PRECISE ABOUT THE WRONG LAW: UNCERTAINTY MISSES POOLED SOURCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation inference models estimate a dynamical law and its uncertainty from a set of trajectories in a single forward pass. This interface assumes that every trajectory comes from the same system, yet the input carries no record of where each trajectory came from. We show that when trajectories from different systems are pooled, the models' native uncertainty does not reveal the violation. In the released model for Markov jump processes, a two-source context and a single-law context matched to its one-step transition statistics receive nearly identical uncertainty. The complete trajectories still tell the two apart: a likelihood-ratio witness that knows the candidate laws, but not the source labels, separates them almost perfectly. Decoders trained on the uncertainty output, by contrast, stay near chance on held-out tasks. We call this failure *context aliasing*: the trajectories expose source heterogeneity that native uncertainty fails to flag. As in classical misspecification, the model fits a pooled pseudo-law while losing fidelity to its sources, with little change in uncertainty. Routing the same trajectories by source cuts generator and first-passage-time error by more than half. A simple transition-count test provides an effective check for source heterogeneity. In controlled simulations, the model for ordinary differential equations shows the same decoupling of error and uncertainty. On recorded motion, the response varies across record groups. Uncertainty from these models is conditional on an experiment; it alone does not establish whether its trajectories should be pooled.

1 INTRODUCTION

A foundation inference model can replace hours of per-system optimization with one forward pass. A user supplies several trajectories, the model returns the differential equation or generator that produced them, and an uncertainty head reports its native uncertainty (Berghaus et al., 2024; Seifner et al., 2025; Mauel et al., 2026). This interface compresses an entire statistical experiment into a dense tensor. The tensor records values, times, and shape. It does not record whether every row came from the same physical system.

That omission changes the meaning of confidence (Figure 1). Suppose half the trajectories come from law θ_a and half from θ_b, while the model must return one law $\widehat{\theta}$. Every row can be noiseless, in-domain, and correctly formatted. Pooling them violates the model's one-law assumption. The model can then return a well-determined compromise law while its estimate remains less faithful to either source. Whether uncertainty reflects that mismatch is the question tested by the comparisons below.

Context aliasing concerns experiments that complete paths distinguish but native uncertainty does not. Source error grows under controlled separation without a commensurate uncertainty response, although shape, observation schedule, cardinality, and aggregate rate scale remain the same.

Figure 2 summarizes the finding in the released foundation inference models FIM-MJP and FIM-ODE. Their generator and vector-field outputs have different uncertainty heads, yet both decouple source error from uncertainty. FIM-MJP assigns similar uncertainty to a heterogeneous context and a single-law control matched in cardinality and the stationary distribution of consecutive state pairs. Source-aware inference recovers distinct scientific answers.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

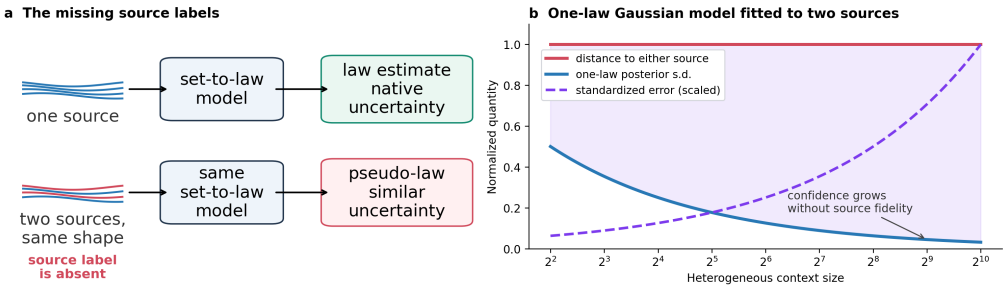

Figure 1: Pooling changes the experiment. (a) Homogeneous and heterogeneous contexts can have the same shape, values in the same domain, and the same number of trajectories. The missing field is source membership. **(b)** In a Gaussian model that assumes one law and is fitted to a balanced two-law context, posterior standard deviation vanishes with context size while distance to either source remains fixed. Standardized scientific error therefore diverges even though the model reports increasing precision. This Gaussian construction illustrates a possible mechanism; Section 4 reports the FIM response.

Our contributions are:

- We formalize the one-law assumption of set-to-law inference and bound the target diameter: without provenance, an uncertainty set that uses only the tensor and is valid for every admissible provenance must span the scientific targets that depend on provenance.
- We isolate context aliasing by one-step Markovization that preserves the adjacent-pair law under stationarity. Complete paths distinguish experiments with the same adjacent-pair law, while native-variance decoders remain near chance on held-out tasks. A matched probe finds heterogeneity signal in the mean output.
- We quantify the value of source identity: oracle routing reduces generator and first-passage-time error, with improvements on every task.
- We connect the neural behavior to one-law Markovization and replicate error–uncertainty decoupling in controlled FIM-ODE simulations. We also benchmark heterogeneity tests and examine recorded motion, where the response varies across record groups. These results motivate source metadata and heterogeneity tests alongside native uncertainty.

2 The One-Law Pooling Trap

This section formalizes the one-law assumption, identifies what missing provenance prevents, and constructs a control that preserves the adjacent-pair law.

2.1 Context Contracts

Let a set-to-law model be a map

$$F : \{(X_k, T_k)\}_{k=1}^{K} \longmapsto (\widehat{\theta}, \widehat{U}), \tag{1}$$

where X_k is a trajectory, T_k its observation schedule, $\widehat{\theta}$ an inferred law, and $\widehat{U}$ a native uncertainty object. The numerical signature fixes dimension and cardinality. The intended statistical experiment additionally assumes

$$\theta \sim p(\theta), \qquad X_k \mid \theta \overset{\text{iid}}{\sim} P_\theta. \tag{2}$$

We call Eq. (2), together with independence units and observation geometry, the *context contract*. A heterogeneous call instead follows

$$Z_k \sim \text{Categorical}(w), \qquad X_k \mid Z_k = s \sim P_{\theta_s}, \tag{3}$$

while the output space still contains one θ. Equations (2) and (3) can generate tensors with identical shapes and individually valid rows. The missing information is relational.

This distinction creates two uncertainties. *Within-law uncertainty* asks how precisely one law is identified if Eq. (2) holds. *Between-source uncertainty* asks whether one law can describe the context at all. Native distributional and heteroscedastic heads are trained for the first question. They need not answer the second.



Figure 2: Source error grows while native uncertainty changes little. **(a)** With tasks, cardinality, and aggregate rate scale fixed, FIM-MJP source error increases while native raw variance does not expand. **(b)** In the repeated aliasing bank at maximum separation, the oracle-law path likelihood ratio (LR) compares mixtures with the homogeneous control P^* (Section 2.3); scores are centered by each task's homogeneous median. Decoders and the Mean + variance probe use leave-one-task-out predictions. Transition LR uses the separate heterogeneity-test bank pooled over separation levels. **(c)** Source-aware routing restores generator, stationary-distribution, and first-passage-time targets; bars report ratios of task-median errors (source-aware over pooled). **(d)** FIM-ODE reproduces the error–uncertainty decoupling across the continuous dynamics families. Curves show task medians; Appendix C.1 lists sample sizes.

Operational definition. Consider a family of paired clean and pooled calls indexed by separation. Let $E_{\mathrm{src}}(\delta)$ denote error to the source laws, and let $U(\delta)$ denote the native uncertainty scale. The burden is the ratio of squared error to variance. Calibration inverts when E_{src} grows while U fails to expand commensurately, causing the burden to rise or source coverage to fall. Absolute confidence contraction, $U(\delta_2) < U(\delta_1)$ for $\delta_2 > \delta_1$, is a stronger special case. This definition applies to architectures with incomparable uncertainty units because each is evaluated through its own error–uncertainty calibration and source coverage.

2.2 An interface bound

If the same tensor permits different scientific targets, any answer based only on that tensor must account for their separation.

The missing field limits what a procedure using only the tensor can certify, even before a model is fitted. For one numerical tensor x, let $\mathcal{Z}(x)$ be its legal provenance or partition annotations and let $T(x, z)$ be the scientific target under z. Define the *target diameter* $\Delta_{\mathcal{C}}(x) = \sup_{z,z' \in \mathcal{Z}(x)} \|T(x, z) - T(x, z')\|$.

Proposition 1 (Target diameter)**.** *For every tensor-only point estimate $\widehat{\theta}(x)$ and any norm,*

$$\sup_{z \in \mathcal{Z}(x)} \|\widehat{\theta}(x) - T(x, z)\| \geq \tfrac{1}{2}\Delta_{\mathcal{C}}(x). \tag{4}$$

Any tensor-only set $C(x)$ guaranteed to contain $T(x, z)$ for every legal z must satisfy $\operatorname{diam} C(x) \geq \Delta_{\mathcal{C}}(x)$. *No deterministic tensor map can identify which external provenance annotation is true.*

Proof. F receives the same x for every annotation. Apply the triangle inequality to any pair approaching the supremum; a set containing both targets has at least their diameter. □

The proposition concerns exact recovery of omitted external provenance. It does not prohibit statistical heterogeneity tests that use path regularities under a generative reference. Those tests provide evidence against one-law pooling; they do not uniquely reconstruct the true real-world annotation for an identical tensor.

2.3 Stationary one-step Markovization

Procedures that assume one law under misspecification target a pooled compromise (White, 1982; Kleijn & van der Vaart, 2006). The control preserves the probability of each adjacent pair under stationarity while discarding the source identity shared along each trajectory.

For grid-observed jump processes, let source s have $P^s = \exp(\mathbf{Q}^s \Delta t)$, stationary occupancy π^s, and weight w_s. Define

$$\rho_i = \sum_s w_s \pi_i^s, \qquad P^\star_{ij} = \frac{\sum_s w_s \pi_i^s P^s_{ij}}{\rho_i}. \tag{5}$$

Proposition 2 (Markovization preserves adjacent pairs, not paths)**.** *If a source is drawn once per stationary trajectory, then the source mixture and the homogeneous Markov chain $(\rho, P^\star)$ have the same distribution of X_t and (X_t, X_{t+1}). Their complete path laws generally differ, with*

$$D_{\mathrm{KL}}(P_{\mathrm{mix}}(X_{0:T}) \| \rho(X_0) \prod_{t=0}^{T-1} P^\star(X_{t+1} \mid X_t)) = \sum_{t=1}^{T-1} I(X_{t+1}; X_{0:t-1} \mid X_t). \tag{6}$$

Equation (6) turns the apparent weakness of a one-step control into a test of context aliasing: if complete paths distinguish the mixture from its Markovization while native uncertainty does not, then the information is present in the context but absent from the uncertainty representation. We use the principal embeddable generator $Q^\star = \log(P^\star)/\Delta t$ only to generate the homogeneous control and evaluate $\exp(\widehat{Q}\Delta t)$ directly against $P^\star$.

3 Controlled evaluation

The evaluation separates error about sources from uncertainty about the pooled law using the models, comparisons, and controls described below. Model parameters, grids, seeds, and thresholds were fixed before the corresponding evaluation. No threshold is selected from an evaluation curve. Banks are selected by experimental role; Appendix C.1 lists configurations and computational budgets.

3.1 Backbones

All checkpoints are used without fine-tuning. Our primary backbone is the released FIM-MJP checkpoint (Berghaus et al., 2024). It maps discretely observed trajectories to native mean and heteroscedastic variance outputs for off-diagonal Markov-generator entries. We use the physical-unit variance, including exponentiation of the log-variance head; we do not assume it is a calibrated Bayesian posterior. The FIM-ODE base checkpoint (Mauel et al., 2026) maps trajectories to a vector field and a native heteroscedastic uncertainty scalar at queried states. Both infer one law from a set of trajectories, with different state spaces and uncertainty heads. Checkpoint revisions are in Table 9.

3.2 MJP separation bank

The MJP separation bank pairs source generators derived from the same base task across separation strengths. Global-clock, preferred-cycle, source-row, and alternating-edge perturbations vary source separation while pairwise rescaling preserves aggregate rate scale. Pooled and homogeneous calls have the same cardinality and observation geometry. Trajectories are independent within each source, and pooled and clean simulations are independent conditional on the generators. Table 9 specifies dimensions and sampling grids.

Generator error is off-diagonal NRMSE averaged against the source laws. We report physical-unit raw variance, rate-normalized variance, squared edge-error divided by variance, and coverage of diagnostic intervals. The intervals describe the scale of the released variance output. We also compare

Table 1: Cross-backbone error–uncertainty decoupling. Entries are absolute pooled medians except rows explicitly labeled as ratios. Sample sizes are listed in Table 9.

Backbone	Endpoint	1.2	1.5	1.9	2.4	Strongest result
FIM-MJP	source NRMSE	0.222	0.394	0.517	**0.634**	3.04× clean
	native raw variance	0.153	0.153	0.140	**0.135**	27/32 decrease
	burden ratio	1.962	4.767	7.465	**8.558**	8.56× clean
	source coverage (%)	100.0	100.0	91.7	**83.3**	lower on 30/32
FIM-ODE	pooled drift NRMSE	0.126	0.292	0.490	**0.699**	11.47× clean
	native uncertainty	0.073	0.072	0.072	**0.079**	remains 0.073–0.079

$\exp(\widehat{\mathbf{Q}}\Delta t)$ with source transition matrices and the pseudo-transition law in Eq. (5). Appendix C.1 specifies intervals, bootstrap units, and the exact paired sign test.

3.3 Context aliasing and source-aware recovery

The pseudo-law/oracle bank compares a source mixture with a one-law experiment having the same adjacent-pair law under stationarity. For each task and strength we construct P^* from Eq. (5), obtain the principal embeddable generator $Q^* = \log(P^*)/\Delta t$, and independently simulate its homogeneous control. All constructions are valid. The repeated aliasing bank independently samples contexts at maximum separation to test whether native uncertainty distinguishes the experiments.

The oracle-law path LR compares the path law of the unlabeled mixture, with its source counts fixed, against the homogeneous P^* path law. It is the exact likelihood ratio for these two path laws. It sums over all source assignments without labeling any path. Native-variance decoders use the logarithms of all off-diagonal variance entries without hyperparameter search. The matched L2 probes likewise use log coordinates, comparing variance, rate-normalized variance, mean, and concatenated mean and variance under identical leave-one-task-out splits. Appendix A gives the likelihood calculation, configurations, and construction checks. The oracle-law path LR knows the candidate laws: it is an information witness, not a deployment algorithm.

In the pseudo-law/oracle bank, an oracle router sends each source's trajectories through the same model separately. Generator NRMSE, stationary-distribution total-variation error, and all-pairs mean first-passage-time NRMSE measure the value of known source labels.

3.4 Equal-cardinality ODE bank

The ODE bank covers linear restoration, cubic restoration, sinusoidal restoration, and double-well dynamics. Source B scales the active coefficient of source A by the MJP separation strengths. Pooled and homogeneous calls have exactly matched cardinality, dimension, time grid, query grid, and initial-condition slots. Drift NRMSE is evaluated against both source fields, and native uncertainty is the mean released $\exp(u)$ head on the common query grid. Table 9 gives the grid sizes.

3.5 Controls and reproducibility

Separate controls test homogeneous evidence, exact row duplication in FIM-ODE, overlapping windows, and source composition. The heterogeneity-test bank evaluates trajectory features and Dirichlet-smoothed transition counts (Section 6). These scores require no learned parameters. Appendices B and C report their complete results and the machine-readable configurations.

4 One-law uncertainty tracks the wrong experiment

We compare errors and uncertainty, test what their outputs reveal, and measure the effect of restoring source identity.

4.1 NATIVE UNCERTAINTY DOES NOT TRACK SOURCE ERROR

Figure 2 and Table 1 show that source error in FIM-MJP rises with separation without a corresponding expansion of native uncertainty. The pooled law can remain well determined even as it becomes less faithful to either source. Physical-unit raw variance decreases slightly but consistently under task pairing, whereas rate-normalized variance increases modestly. The former measures absolute dispersion and supports mild contraction; the latter measures dispersion relative to the inferred rates and does not contract. Neither tracks source error. Table 10 reports the paired sign test; Table 3 summarizes response within tasks.

At maximum separation, the clean diagnostic intervals already cover every source target. Pooling raises squared error relative to variance and reduces coverage from that conservative baseline (Table 16). The error–uncertainty mismatch holds across the perturbation mechanisms (Table 11).

4.2 THE PATHS REVEAL THE VIOLATION; NATIVE UNCERTAINTY DOES NOT

The repeated aliasing bank exposes a representation gap. Matching adjacent pairs leaves differences in complete paths that native uncertainty may fail to reflect. The path LR attains **0.999 AUROC**, demonstrating that the complete contexts are distinguishable even though their adjacent-pair law is exactly matched. The best native-variance decoder reaches only **0.555 AUROC** on held-out tasks. Scalar variance is likewise nearly unchanged between mixtures and homogeneous controls. Table 7 reports the task-level results, intervals, and decoder comparisons.

The mean output retains evidence of heterogeneity. Native variance therefore loses information that remains available in other outputs. Under the same grouped splits and L2 probe, mean plus variance reaches **0.809 AUROC**. Mean-only and normalized variance probes also contain signal, while native variance alone does not reliably identify source heterogeneity across tasks. Table 4 summarizes the comparison; Section 6 evaluates a transition LR that does not know the generating laws.

FIM-MJP still fits the pooled law: its generator error to Q^* remains small for both mixtures and homogeneous contexts (Table 4). The pooled target describes adjacent transitions, so fitting it need not recover either source. Native uncertainty describes the pooled estimate while providing little evidence about the validity of pooling.

4.3 RESTORING SOURCE IDENTITY RESTORES THE SCIENTIFIC ANSWER

At maximum separation, oracle routing reduces generator error by **60.5%** and first-passage-time error by **55.5%**, using the median of per-task relative reductions. Routing by source removes the pooling that forces trajectories from different laws into one estimate. Both improve on every task. Stationary-law error also improves on most tasks. Source identity therefore changes long-horizon scientific quantities as well as parameter-space scores. Table 12 reports the absolute errors and task counts.

4.4 THE DIRECTION REPLICATES IN CONTINUOUS-LAW INFERENCE

FIM-ODE reproduces error–uncertainty decoupling in the equal-cardinality study. Its head summarizes one inferred law while the source partition remains absent. Drift error rises with separation as native uncertainty changes little (Table 1). At maximum separation, the median pooled/clean uncertainty ratio is **0.818**: the pooled call reports lower uncertainty despite its higher source error (Table 13).

Exact row duplication separately leaves uncertainty and predictions nearly unchanged (Table 13), showing stability to this duplication. Across these controlled studies, generator and vector-field outputs show the same decoupling: neither uncertainty head represents the omitted source partition.

4.5 RECORDED MOTION SHOWS A NONUNIFORM RESPONSE

CMU MoCap extends the comparison to recorded motion using a shared representation fitted only to training sequences. Mixed contexts increase held-out forecast error and reduce native uncertainty in the aggregate (Table 2). The response varies by record-group pair: uncertainty also rises in one pair, and routing improves most comparisons, rather than all. Same-group predictions outperform persistence on every test sequence. These observations concern forecasts and the native head at matched states; the head supplies no calibrated forecast intervals. Record groups are proxies for sources, and dimensionality reduction discards motion information, so this stress test complements the controlled experiments without identifying true source laws.

Table 2: CMU MoCap: mixing record groups. H uses same-group sequences and P pools groups, both with four trajectories; R routes P's target-group half. E is forecast RMSE; U is native uncertainty at common held-out states. Entries are medians of cell ratios after averaging context draws. Cells share sequences across pairs; Appendix D gives the protocol, absolute errors, and all context sizes.

Record groups	Cells	E_P/E_H	U_P/U_H	E_R/E_P
09–35	6	1.229	1.167	0.905
09–39	4	1.835	0.779	0.575
35–39	6	1.161	0.752	0.948
All pairs	16	1.308	0.789	0.838

5 MECHANISM: A STABLE PSEUDO-LAW

Three controls test whether the decoupling reflects pooling under one law or a broader response to context changes.

Homogeneous evidence works. Additional independent trajectories from one generator satisfy the context contract and make FIM-MJP more accurate and less uncertain (Appendix B.1). On the raw scale, median elasticity is positive for added homogeneous evidence and negative for increased source separation (Table 3).

The estimate approaches the pseudo-generator. The pseudo-law/oracle bank separates fidelity to the sources from fit to their stationary adjacent-pair law. Markovization preserves adjacent pairs while discarding the dependence associated with persistent source identity. As separation increases, FIM-MJP stays close to $P^\star$ while losing source fidelity. A rowwise Dirichlet-smoothed estimator on the same paths shows the same geometry, with nearly constant posterior transition variance (Table 14). The pooled target can thus remain well determined while omitting source-specific dynamics.

Overlapping windows increase uncertainty. Overlapping windows from one physical realization increase median FIM-MJP error and rate-normalized variance (Table 14). This control changes dependence within one realization, whereas source pooling combines different laws. The response depends on both the context change and the uncertainty scale; these comparisons support no universal direction for context violations.

Together these controls give a mechanistic chain:

source mixture $\Rightarrow$ one-law compression $\Rightarrow$ fit to a pseudo-law $\Rightarrow$ source error without wider uncertainty.

These controls are consistent with fitting the pooled law while losing fidelity to its sources.

6 REPAIRING THE INTERFACE

The strongest repair is structural: carry source, device, batch, intervention, or physical-replicate identity beside every trajectory and route by that field. The oracle experiment demonstrates its scientific value. When identity is unavailable, native uncertainty needs an external heterogeneity test.

Table 3: Response within tasks. Median and interquartile range of $\epsilon_i = \Delta \log S_i / \Delta \log E_i$. MJP uses $S_i = \sqrt{V_i}$ on the stated variance scale; ODE uses its native $\exp(u)$ separately. These descriptive summaries are not calibration tests. Endpoints and computation are specified in Appendix C.6.

Comparison	Tasks	Median	Interquartile range
Homogeneous evidence (raw)	32	0.137	$[0.012, 0.236]$
Homogeneous evidence (normalized)	32	0.333	$[-0.161, 0.683]$
Overlapping windows (raw)	16	0.018	$[-0.143, 0.107]$
Overlapping windows (normalized)	16	0.227	$[0.151, 0.511]$
Source separation (raw)	32	-0.032	$[-0.099, -0.014]$
Source separation (normalized)	32	0.104	$[0.014, 0.175]$
FIM-ODE (native scale)	16	0.170	$[0.004, 0.276]$

Table 4: Context aliasing and source recovery at maximum separation. AUROCs and the variance ratio use the repeated aliasing bank; error medians use the pseudo-law/oracle bank. Sample sizes are in Table 9. A dash denotes a target not defined for that call.

Endpoint	Heterogeneous	Homogeneous P^*	Source oracle
Oracle-law path LR AUROC		**0.999**	–
Best native-variance decoder AUROC		0.555	–
Mean + variance probe AUROC		0.809	–
Taskwise raw-variance ratio (95% CI)		0.992 (0.934–1.014)	–
Generator NRMSE to Q^*	0.203	0.166	–
Source generator NRMSE	0.620	–	**0.257**
Stationary-distribution TV	0.129	–	**0.068**
First-passage-time NRMSE	0.597	–	**0.229**

Native-variance decoder labels in Figure 2: L2 native variance, RBF native variance, ExtraTrees native variance, and kNN native variance. The Mean + variance probe uses the same L2 decoder.

The heterogeneity-test bank compares training-free scores against independent homogeneous P^* controls. Each context is partitioned by deterministic two-means on trajectory features; the same complete procedure applies to mixtures and controls. The scores are feature silhouette, between/within feature dispersion, a Dirichlet-Wald score, a smoothed transition LR, or CONTRACTUQ, which standardizes subgroup FIM-MJP generator disagreement by native variance. The bank includes balanced, imbalanced, and multi-source compositions (Table 5).

Transition LR reaches **0.816 AUROC** for the balanced composition, pooling separation levels. It tests differences in transitions directly, while CONTRACTUQ measures disagreement between inferred generators relative to native variance. Transition LR is stronger overall and cheaper than an additional model query. CONTRACTUQ provides a complementary score for strong conflicts when raw transitions are unavailable. Table 5 and Appendix B report all composition and separation results.

Table 5: Source-heterogeneity detection AUROC under source composition. Values pool all four separation levels; positives are source mixtures and negatives are homogeneous P^* contexts.

Test	16/16	24/8	29/3	11/11/10
Feature silhouette	0.701	0.634	0.547	0.711
Feature between/within	0.704	0.669	0.537	0.715
Transition Wald	0.797	0.811	**0.751**	**0.743**
Transition LR	**0.816**	**0.833**	0.742	0.724
CONTRACTUQ	0.740	0.768	0.713	0.718

A deployed model should report uncertainty conditional on one law alongside evidence that pooling is valid. Low native uncertainty is actionable only when disagreement is low; high disagreement calls for source-aware inference regardless of native precision. Thresholds should be calibrated on homogeneous reference tasks from the intended deployment regime.

7 Relation to prior work

Foundation inference models and uncertainty. FIM-MJP, FIM-SDE, and FIM-ODE establish set-to-law inference in the zero-shot setting (Berghaus et al., 2024; Seifner et al., 2025; Mauel et al., 2026). We study their deployment boundary: what statistical experiment does a shape-valid set represent? Inductive-bias probes test whether foundation models learn the underlying world structure (Vafa et al., 2025); we test whether native uncertainty detects that trajectories come from different systems. Predictive uncertainty can degrade under shift (Ovadia et al., 2019), and amortized inference can be brittle to selected observations (Gloeckler et al., 2023); ensembles improve within-model dispersion (Lakshminarayanan et al., 2017; Yadav et al., 2026). Here every trajectory remains admissible. Only the relational source partition changes, so an ensemble conditioned on one law can remain precise around the same pseudo-law.

Misspecification, mixtures, and latent environments. Pseudo-true parameters and posterior contraction under misspecification are classical (White, 1982; Kleijn & van der Vaart, 2006). Sandwich posteriors and bagged posteriors target covariance mismatch or instability around such parameters (Hoff & Wakefield, 2013; Huggins & Miller, 2023). Mixture models and environment inference recover latent regimes when heterogeneity is explicitly modeled (Kausik et al., 2023; Liu et al., 2025); classical transition-count tests detect unequal Markov laws (Anderson & Goodman, 1957). Posterior and holdout predictive checks assess a model against observed or held-out data (Gelman et al., 1996; Moran et al., 2024).

Misspecification in simulation-based inference. Robust neural posterior estimation models simulation–data discrepancies (Ward et al., 2022); learned diagnostics detect misspecification (Schmitt et al., 2024), and learned statistics reduce its impact (Huang et al., 2023). Our controlled comparison isolates source pooling with stationary adjacent-pair laws matched: complete paths retain evidence of heterogeneity that native variance fails to reveal across held-out tasks.

8 Deployment consequences

Uncertainty is conditional on an experiment, not a certificate for it. A narrow output answers how concentrated inference is *assuming* one law; it does not answer whether the rows should have been pooled.

Scientific inference APIs should expose source, batch, device, intervention, and physical replicate as first-class metadata. Without these fields, an observable-space heterogeneity test should accompany native uncertainty; strong disagreement triggers abstention, partitioning, or a mixture-valued output.

Source metadata and independence units determine which trajectories can support a shared law. At fixed context size, increasing separation between source laws raises source error while native uncertainty changes little.

9 Scope and conclusion

Controlled simulations cover one jump-process and one vector-field model across four mechanism families. They do not test the effect of context size in Figure 1b. FIM-MJP provides the strongest result on the uncertainty scale, with directional replication in FIM-ODE. Recorded motion provides an observational stress test with a nonuniform response. Heterogeneity tests lose power for mild or highly imbalanced conflicts, where metadata, abstention, or a richer mixture model may be required.

The central conclusion is operational: uncertainty is meaningful only inside the statistical experiment it conditions on. Complete paths distinguish a source mixture from its stationary Markovization, whereas native variance provides little signal across held-out tasks. Other model outputs and transition counts retain evidence of heterogeneity, and source-aware routing recovers scientific accuracy. Foundation inference models therefore need a context contract alongside every uncertainty estimate.

Reproducibility. The supplement contains configurations, hashes, task rows, and construction checks. Appendix C.1 records the evaluation and computational budgets.

AI USE DISCLOSURE

The authors used generative AI assistants (OpenAI Codex and Claude) in a supporting role, for literature search, assistance with code implementation, checking derivations, figure formatting, language editing, and translation. The research questions, experimental design, proofs, analysis, and conclusions are the authors' own. All numerical results were produced by the authors' experiment code using the released model checkpoints. The authors reviewed and verified all AI-assisted content, including code and citations, and take full responsibility for the paper.

REFERENCES

Theodore W. Anderson and Leo A. Goodman. Statistical inference about Markov chains. *The Annals of Mathematical Statistics*, 28(1):89–110, 1957. doi: 10.1214/aoms/1177707039. URL https://doi.org/10.1214/aoms/1177707039.

David Berghaus, Kostadin Cvejoski, Patrick Seifner, César Ojeda, and Ramsés J. Sánchez. Foundation inference models for Markov jump processes. In *Advances in Neural Information Processing Systems*, volume 37, pp. 129407–129442. Curran Associates, Inc., 2024. doi: 10.52202/079017-4112. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e9df36b21ff4ee211a8b71ee8b7e9f57-Abstract-Conference.html.

Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–807, 1996. URL https://www3.stat.sinica.edu.tw/statistica/j6n4/j6n41/j6n41.htm. With discussion.

Manuel Gloeckler, Michael Deistler, and Jakob H. Macke. Adversarial robustness of amortized Bayesian inference. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11493–11524. PMLR, 2023. URL https://proceedings.mlr.press/v202/gloeckler23a.html.

Pashupati Hegde, Çağatay Yıldız, Harri Lähdesmäki, Samuel Kaski, and Markus Heinonen. Variational multiple shooting for bayesian ODEs with gaussian processes. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 790–799, 2022. URL https://proceedings.mlr.press/v180/hegde22a.html.

Peter Hoff and Jon Wakefield. Bayesian sandwich posteriors for pseudo-true parameters: A discussion of "Bayesian inference with misspecified models" by Stephen Walker. *Journal of Statistical Planning and Inference*, 143(10):1638–1642, 2013. doi: 10.1016/j.jspi.2013.05.014. URL https://doi.org/10.1016/j.jspi.2013.05.014.

Daolang Huang, Ayush Bharti, Amauri Souza, Luigi Acerbi, and Samuel Kaski. Learning robust statistics for simulation-based inference under model misspecification. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/16c5b4102a6b6eb061e502ce6736ad8a-Abstract-Conference.html.

Jonathan H. Huggins and Jeffrey W. Miller. Reproducible model selection using bagged posteriors. *Bayesian Analysis*, 18(1):79–104, 2023. doi: 10.1214/21-BA1301. URL https://doi.org/10.1214/21-BA1301.

Chinmaya Kausik, Kevin Tan, and Ambuj Tewari. Learning mixtures of Markov chains and MDPs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15970–16017. PMLR, 2023. URL https://proceedings.mlr.press/v202/kausik23a.html.

Bas J. K. Kleijn and Aad W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, 34(2):837–877, 2006. doi: 10.1214/009053606000000029. URL https://doi.org/10.1214/009053606000000029.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Shixuan Liu, Yue He, Haotian Wang, Wenjing Yang, Yunfei Wang, Peng Cui, and Zhong Liu. Environment inference for learning generalizable dynamical system. In *Advances in Neural Information Processing Systems*, volume 38, pp. 164076–164105. Curran Associates, Inc., 2025. doi: 10.52202/085713-5475. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f00649987484cba88e5a3699249df65c-Abstract-Conference.html.

Maximilian Mauel, Johannes R. Hübers, David Berghaus, Patrick Seifner, and Ramsés J. Sánchez. Foundation inference models for ordinary differential equations. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.08733.

Gemma E. Moran, David M. Blei, and Rajesh Ranganath. Holdout predictive checks for Bayesian model criticism. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1): 194–214, 2024. doi: 10.1093/jrsssb/qkad105. URL https://doi.org/10.1093/jrsssb/qkad105.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Marvin Schmitt, Paul-Christian Bürkner, Ullrich Köthe, and Stefan T. Radev. Detecting model misspecification in amortized Bayesian inference with neural networks. In *Pattern Recognition: DAGM GCPR 2023*, volume 14264 of *Lecture Notes in Computer Science*, pp. 541–557. Springer, 2024. doi: 10.1007/978-3-031-54605-1_35. URL https://doi.org/10.1007/978-3-031-54605-1_35.

Patrick Seifner, Kostadin Cvejoski, David Berghaus, César Ali Ojeda Marin, and Ramsés J. Sánchez. In-context learning of stochastic differential equations with foundation inference models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 152636–152676. Curran Associates, Inc., 2025. doi: 10.52202/085713-5104. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/dff573cecb50646083b5186b565e23a3-Abstract-Conference.html.

Keyon Vafa, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. What has a foundation model found? Using inductive bias to probe for world models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60727–60747. PMLR, 2025. URL https://proceedings.mlr.press/v267/vafa25a.html.

Daniel Ward, Patrick Cannon, Mark Beaumont, Matteo Fasiolo, and Sebastian Schmon. Robust neural posterior estimation and statistical model criticism. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/db0eac6747e3631eb91095cd76065611-Abstract-Conference.html.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. doi: 10.2307/1912526. URL https://doi.org/10.2307/1912526.

Akash Yadav, Taiwo A. Adebiyi, and Ruda Zhang. Calibrating scientific foundation models with inference-time stochastic attention. arXiv preprint arXiv:2604.19530, 2026. URL https://arxiv.org/abs/2604.19530.

A MARKOVIZATION AND CONTEXT ALIASING

A.1 WHY ADJACENT-PAIR MATCHING IS EXACT

Let a source $S \sim w$ be drawn once for a stationary trajectory. Then $\Pr(X_t = i) = \rho_i = \sum_s w_s \pi_i^s$ and

$$\Pr(X_t = i, X_{t+1} = j) = \sum_s w_s \pi_i^s P_{ij}^s = \rho_i P_{ij}^\star. \tag{7}$$

Thus $(\rho, P^\star)$ matches every stationary adjacent pair exactly. For the complete path, apply the chain rule to the mixture law and subtract the Markovized log density:

$$D_{\mathrm{KL}}(P_{\mathrm{mix}} \| P_{\mathrm{Markov}}) = \sum_{t=1}^{T-1} \mathbb{E} \log \frac{P(X_{t+1} \mid X_{0:t})}{P(X_{t+1} \mid X_t)} \tag{8}$$

$$= \sum_{t=1}^{T-1} I(X_{t+1}; X_{0:t-1} \mid X_t). \tag{9}$$

Persistent source identity makes the sum positive unless the source-specific chains induce identical conditional futures. This proves Proposition 2 and precisely separates one-step matching from path equivalence. For the irreducible jump processes used here, $\exp(Q\Delta t)$ has positive entries, so every finite state sequence has common support under both source laws and P^*. A path likelihood can accumulate strong statistical evidence, but the numerical tensor alone does not deterministically reveal its external provenance annotation. This is the concrete MJP witness for Proposition 1.

A.2 EMBEDDABILITY AND NUMERICAL BRANCH CHECKS

We use SciPy 1.13.1's principal matrix logarithm and explicitly reconstruct the transition matrix. Table 6 reports worst cases over the 32-task repeated aliasing bank; the four-strength pseudo-law/oracle bank contains 128/128 valid constructions as well. The spectrum remains far from zero and the negative real-axis branch cut.

Table 6: P^* construction checks at separation 2.4.

Check	Worst case
$\max \lvert\Im\{\log(P^*)/\Delta t\}\rvert$	0
$\max \lvert\exp(Q^*\Delta t) - P^*\rvert$	1.33×10^{-15}
$\max \lvert\pi(Q^*) - \rho\rvert$	1.33×10^{-15}
Minimum $\Re\{\lambda(P^*)\}$	0.785
Maximum $\lvert\arg \lambda(P^*)\rvert$	0.0146
Valid generators	32/32

A.3 ORACLE-LAW PATH LR

For each task we generate 50 contexts from each experiment. A mixture context has exactly 16 trajectories from source A and 16 from source B, with the row order permuted. A homogeneous context has 32 trajectories from Q^*. All trajectories have 96 observations at $\Delta t = 0.08$ and begin in the corresponding stationary distribution.

For path x_k, define $\ell_s(k) = \log \pi_s(x_{k0}) + \sum_t \log P^s_{x_{kt}, x_{k,t+1}}$. The unlabeled fixed-count likelihood is

$$\log L_H = \log \sum_{A:|A|=16} \exp\left[\sum_{k \in A} \ell_a(k) + \sum_{k \notin A} \ell_b(k)\right] - \log \binom{32}{16}. \tag{10}$$

We compute the combinatorial sum with a 17-state log-space dynamic program. The score $\log L_H - \sum_k \ell_*(k)$ never receives a per-path source label. Table 7 reports the outcomes.

Raw variance averages the six released variance coordinates in physical units; its decrease describes contraction on that scale. Rate-normalized variance divides each coordinate by the squared predicted

Table 7: Repeated aliasing bank. The raw-variance interval resamples tasks as clusters.

Endpoint	Result
Oracle-law path LR AUROC	0.999
Median / minimum task AUROC	1.000 / 0.999
Tasks with AUROC ≥ 0.90	32/32
Median path-LR power	100%
L2 native variance (95% cluster CI)	0.555 [0.527,0.590]
RBF native variance AUROC	0.543
ExtraTrees native variance AUROC	0.545
kNN native variance AUROC	0.522
Normalized variance: grouped / task median	0.637 / 0.758
Mean + variance (95% cluster CI)	0.809 [0.752,0.864]
Median taskwise raw-variance ratio	0.992
Task-bootstrap 95% interval	[0.934, 1.014]
Task ratios in $[0.75, 1.25]$	32/32
Median taskwise normalized-variance ratio	1.101

rate; its increase describes uncertainty relative to the rate scale. Both change less than the 2.86-fold rise in source error, while burden and coverage measure the resulting task-level mismatch.

Normalized variance is detectable when a separate classifier is fit inside each task (median AUROC 0.758), but its leave-one-task-out AUROC is 0.637 and leave-one-family-out AUROC is 0.598. Native variance therefore does not reliably identify source heterogeneity on held-out tasks. We do not claim that every task-specific nonlinear function of the output is uninformative.

B Operating controls and heterogeneity tests

B.1 Homogeneous evidence

The homogeneous bank asks whether the native output responds normally when its context contract holds. With 4, 8, 16, 32, and 64 independent trajectories from one generator, median NRMSE is $0.479, 0.357, 0.293, 0.260, 0.220$ and raw variance is $0.223, 0.205, 0.178, 0.181, 0.176$. Accuracy improves on 31/32 tasks. These operating curves describe how error and the released variance output change with additional independent trajectories from one generator. The output responds to valid evidence within a one-law experiment but does not verify that a pooled context belongs to that experiment.

B.2 Heterogeneity tests

Table 8 reports all five scores. Intervals are task-bootstrap 95% intervals.

Table 8: Balanced 16/16 heterogeneity tests.

Test	All strengths AUROC	Strength 2.4 AUROC
Feature silhouette	0.701 [0.605,0.793]	0.875 [0.723,0.992]
Feature between/within	0.704 [0.610,0.795]	0.816 [0.629,0.984]
Transition Wald	0.797 [0.719,0.874]	0.957 [0.879,1.000]
Transition LR	**0.816** [0.732,0.888]	**0.961** [0.891,1.000]
ContractUQ	0.740 [0.652,0.824]	0.930 [0.816,1.000]

The exact path LR in Appendix A is an information witness that uses the known source candidate laws. The transition LR is a heterogeneity score for deployment: it estimates two subgroups from observable transition counts and does not receive the generating laws or true source labels.

C CONFIGURATIONS AND FULL NUMERICAL RESULTS

C.1 EXPERIMENTAL CONFIGURATIONS

The banks have distinct roles and are not pooled across studies. The main uncertainty curves use the MJP separation bank; source recovery uses the pseudo-law/oracle bank, and decoder comparisons use the repeated aliasing bank.

Table 9: Experimental configurations and computational budgets.

Component	Configuration
Checkpoints	FIM-MJP revision `cee2bb53`; FIM-ODE base revision `00875c25`.
MJP generators	Two three-state sources; six off-diagonal rates per generator; pairwise arithmetic mean over 12 source rates held equal across strengths.
Separation grid	$1.2, 1.5, 1.9, 2.4$; global-clock, preferred-cycle, source-row, alternating-edge; eight tasks per family, 32 tasks and 128 task–strength evaluations.
MJP observations	32 independent trajectories per call, length 96, spacing 0.08; pooled source allocation 16/16; homogeneous calls use 32 paths from one law. Pooled and clean simulations are conditionally independent.
MJP separation bank	128 valid evaluations from 384 model calls.
Pseudo-law/oracle bank	128 valid evaluations from 512 calls; all 128 constructions valid. Oracle routes the same 16 observed paths per source separately.
Repeated aliasing bank	Separation 2.4; 50 mixture and 50 homogeneous contexts per task, 32 tasks, 3,200 calls; six native variance coordinates; four decoders. Fixed-count path likelihood sums $\binom{32}{16}$ assignments with a 17-state log-space program.
ODE bank	Linear, cubic, sinusoidal restoration and double-well dynamics; same four strengths; 16 tasks per strength, 64 evaluations/192 calls. Four trajectories per call on 128 time points; pooled allocation two per source, same four initial-condition slots; 161 query states.
Heterogeneity-test bank	256 pairs/1,024 calls; 32 trajectories per context; source compositions 16/16, 24/8, 29/3, and 11/11/10. Deterministic two-means on the same trajectory features.
Uncertainty intervals	Diagnostic edges $\widehat{Q}_e \pm 1.96\sqrt{\widehat{V}_e}$; these describe the scale of the released variance output. Task-paired bootstrap 95% intervals; primary medians use 5,000 draws, test AUROCs use 2,000 draws. Strongest-to-weakest variance uses an exact paired sign test.

C.2 MJP RESPONSE AND MECHANISM BREAKDOWN

Table 10: Complete MJP endpoints for the MJP separation bank.

Endpoint	Reported result
Pooled/clean generator error	$1.42\times$ at separation 1.2; $3.04\times$ at 2.4.
Absolute source NRMSE	0.222 (95% bootstrap interval 0.211–0.245) at 1.2; 0.634 (0.597–0.647) at 2.4. Intermediate medians at 1.5/1.9: 0.394/0.517.
Native raw variance	Medians at strengths 1.2/1.5/1.9/2.4: 0.153/0.153/0.140/0.135. Under task pairing, 27/32 decrease; one-sided exact sign-test $p = 5.65 \times 10^{-5}$; median strongest/weakest ratio 0.932.
Normalized variance	Strongest/weakest ratio $1.24\times$. Neither normalization tracks the 2.86-fold absolute source-error increase (nearly threefold).

Endpoint	Reported result
Pooled/clean burden	Rounded endpoints $1.96\times$ and $8.56\times$; full displayed sequence 1.962, 4.767, 7.465, 8.558.
Diagnostic edge coverage	At separation 2.4, pooled 95% coverage is lower on 30/32 tasks; median 83.3%, versus clean 100%. Pooled medians across strengths: 100.0, 100.0, 91.7, 83.3%.

Table 11: Paired mechanism breakdown from separation 1.2 to 2.4.

Mechanism	Raw-variance ratio	Decreases	Source NRMSE
Global clock	0.875	7/8	$0.223 \to 1.086$
Preferred cycle	0.982	5/8	$0.233 \to 0.610$
Source row	0.804	8/8	$0.192 \to 0.452$
Alternating edge	0.955	7/8	$0.249 \to 0.638$

C.3 Readout and source recovery

The repeated aliasing bank's path LR has task-centered pooled AUROC 0.999 (each task's homogeneous median is subtracted before pooling scores); all 32 task AUROCs exceed 0.998. Median power is 100% at each task's empirical 95th-percentile homogeneous threshold. Table 7 gives the decoder results. The matched L2 mean-only probe has AUROC 0.603, versus 0.637 for normalized variance and 0.809 for mean plus variance. The best native-variance decoder is 0.555. All 32 taskwise raw-variance ratios lie in $[0.75, 1.25]$; their median is 0.992 with task-bootstrap 95% interval 0.934–1.014. In the four-strength pseudo-law/oracle bank, generator NRMSE to Q^* is 0.203 for mixtures and 0.166 for homogeneous controls.

Table 12: Source recovery at maximum separation (32 tasks). Absolute errors are task medians. Relative reductions are $\mathrm{median}_i(1 - E_{i,\mathrm{oracle}}/E_{i,\mathrm{pooled}})$, shown as rounded percentages.

Endpoint	Pooled	Oracle	Improved	Reduction
Generator NRMSE	0.620	0.257	32/32	60.5%
Stationary-law TV	0.129	0.068	23/32	–
First-passage NRMSE	0.597	0.229	32/32	55.5%

The first-passage reduction is 55.5% and the generator reduction is 60.5%; both exceed one half. Panel (c) reports ratios of task-median errors: 0.41, 0.52, and 0.38 for generator, stationary, and first-passage errors. The text reports medians of per-task relative reductions.

C.4 ODE response and controls

Table 13: ODE results and controls.

Endpoint	Reported result
Pooled drift NRMSE	At strengths 1.2/1.5/1.9/2.4: 0.126/0.292/0.490/0.699; 5.54-fold growth (more than fivefold).
Native uncertainty	0.073/0.072/0.072/0.079, with endpoint values 0.073 and 0.079. At maximum separation, median pooled/clean ratio 0.818.
Strongest clean comparison	Clean drift NRMSE 0.044; median pooled/clean error ratio $11.47\times$.
Exact row duplication	Uncertainty ratio $0.999989\times$; prediction-error ratio $0.999806\times$; median prediction displacement 0.000151 NRMSE. The head does not collapse simply because four rows are present.

C.5 MECHANISM AND HETEROGENEITY-TEST DETAILS

Table 14: Pseudo-law and operating controls.

Control	Reported result
Homogeneous evidence	Context sizes 4, 8, 16, 32, 64; NRMSE 0.479/0.357/0.293/0.260/0.220; raw variance 0.223/0.205/0.178/0.181/0.176; improvement on 31/32 tasks.
FIM-MJP transition error	At separation 2.4: 0.606 to sources, 0.198 to $P^\star$; at 1.2: 0.208 and 0.157.
Dirichlet transition estimator	Source errors 0.192/0.360/0.495/0.594; pseudo-law errors 0.105/0.134/0.107/0.126; posterior transition variance approximately 5.2×10^{-5}.
Overlapping-window control	Error increases $1.647\times$ and rate-normalized variance increases $1.171\times$; these are medians of within-task ratios.
Transition LR / ContractUQ	All-strength balanced AUROC 0.816/0.740; 24/8 AUROC 0.833/0.768. At maximum separation, balanced AUROC 0.961/0.930 and 29/3 AUROC 0.918/0.879.

Table 15: Complete source-composition comparison. Values pool all four separation levels; positives are source mixtures and negatives are homogeneous P^* contexts.

Test	16/16	24/8	29/3	11/11/10
Feature silhouette	0.701	0.634	0.547	0.711
Feature between/within	0.704	0.669	0.537	0.715
Transition Wald	0.797	0.811	**0.751**	**0.743**
Transition LR	**0.816**	**0.833**	0.742	0.724
CONTRACTUQ	0.740	0.768	0.713	0.718

C.6 PAIRED RESPONSE AND ABSOLUTE DIAGNOSTICS

Response elasticity is calculated within each task before aggregation. Here E_i is the task's NRMSE: generator error in the homogeneous and overlap controls, and error to the source laws in the separation comparisons. For endpoints a, b, we use $\epsilon_i = \log(S_{i,b}/S_{i,a})/\log(E_{i,b}/E_{i,a})$ and report the median and empirical interquartile range across tasks. For MJP, S_i is the square root of the reported raw or rate-normalized variance; for ODE it is the native $\exp(u)$ scale. The latter is not identified as a calibrated standard deviation or variance. Every error and uncertainty endpoint is positive, and every log-error difference is nonzero, so all tasks are retained.

The homogeneous comparison uses 4 and 64 trajectories from the same generator; the overlap comparison uses independent trajectories and overlapping windows within each task. The latter's reported variance ratio of 1.171 refers to rate-normalized variance. Source-separation rows pair the same base task at strengths 1.2 and 2.4. These comparisons use different banks, changes to the context, and uncertainty scales. Their endpoint elasticities describe measured responses; a value of one is not imposed as a common calibration target. Task-level signs vary, and raw and normalized source-separation rows have different median signs.

Table 16 reports absolute burden alongside coverage, including means. At maximum separation, both mean and median clean coverage are 100%. Each task averages two clean source calls with six off-diagonal entries each: all 384 source–edge comparisons are covered. The clean burden median is 0.273, compared with 2.377 for pooled calls. These measurements indicate conservative diagnostic intervals relative to the observed clean errors in this bank. Edge errors share a task and an inferred generator, so independent Bernoulli trials cannot be assumed. The intervals diagnose the released variance scale; they do not establish nominal frequentist coverage.

Table 16: Absolute burden and diagnostic coverage. Burden columns are task medians of squared edge error divided by variance; coverage columns report means and medians over 32 tasks at each separation.

	Burden median		Mean coverage (%)		Median coverage (%)	
Separation	Clean	Pooled	Clean	Pooled	Clean	Pooled
1.2	0.110	0.226	100.00	100.00	100.0	100.0
1.5	0.132	0.691	100.00	98.44	100.0	100.0
1.9	0.187	1.335	100.00	91.41	100.0	91.7
2.4	0.273	2.377	100.00	83.07	100.0	83.3

Table 17: Transition LR and ContractUQ across source compositions. All-strength and maximum-separation AUROCs use their respective evaluation sets; the two aggregation scopes are kept separate.

	All strengths		Maximum separation	
Composition	Transition LR	CONTRACTUQ	Transition LR	CONTRACTUQ
16/16	0.816	0.740	0.961	0.930
24/8	0.833	0.768	0.926	0.902
29/3	0.742	0.713	0.918	0.879
11/11/10	0.724	0.718	0.930	0.852

D CMU MoCap record-group comparison

D.1 Data, representation, and comparisons

The CMU Graphics Lab Motion Capture Database provides recorded human motion. We use the public 50-channel benchmark sequences distributed by Hegde et al. (2022), which also underpin the MoCap evaluation of FIM-ODE (Mauel et al., 2026). The data originate from https://mocap.cs.cmu.edu/; collection was funded by NSF EIA-0196217. We include all three benchmark groups, 09, 35, and 39, and every pair among them. CMU notes that one person can appear under multiple subject numbers and that each number has a separately calibrated skeleton. We therefore call these *record groups*; they can differ in motion, skeleton calibration, and recording conditions, rather than only in a dynamical law.

The published training/test sequence counts are 6/2, 16/4, and 6/2, respectively; validation sequences are unused. We retain the supplied split. The arrays do not identify original recording trials, so held-out benchmark sequences are the evaluation units, without a claim of independent people or recording trials. No evaluation window is an exact duplicate of a training window. We use the dataset revision `c08472981 7e09cb3910b45ec268eb4688e0a44f8`; the accompanying manifest records file hashes and download URLs.

Following the benchmark loader, channels 24, 25, 31, and 32 (zero-based) are set to 10^{-6}. One PCA is fitted to the first 100 frames of every training sequence, jointly across groups, and retains three components, the dimensionality supported by the checkpoint. Its coordinates are standardized using the mean and standard deviation of those same training frames, with 10^{-5} added to each standard deviation. There is no group-specific PCA or centering. The retained training variance is 77.94%; discarded components limit the scope of this comparison. The time step is 0.01 in the published benchmark convention, with no claim about the raw capture frequency.

Contexts contain $n \in \{2, 4, 6\}$ sequences with 100 frames each; $n = 4$ is the primary comparison. For each of five context draws, a within-group permutation is generated by NumPy's default random generator with seed $20260925 + 100r + g$, where $r = 0, \ldots, 4$ and g is the numerical group label. The same-group context H uses the first n training sequences from the target group. The mixed context P uses the first $n/2$ from each member of a pair. The routed context R uses exactly the target group's $n/2$ sequences from P, so routing introduces no additional observations. H and P have equal cardinality; R isolates routing of the same observations with fewer trajectories per call. All selections are without replacement within a context. Context draws overlap and are not independent replications.

D.2 Prediction, uncertainty, and aggregation

The released base FIM-ODE checkpoint and source revisions are unchanged from Table 9; no parameters are fitted. Each held-out sequence supplies its initial state, followed by a 50-step forecast using RK4 at the observation step size. Future held-out states are targets only, and never enter the context or rollout. The primary error E is RMSE over these future frames and the three standardized PCA coordinates. Persistence keeps the initial state fixed.

We separately evaluate $U = \exp(u)$ on all 51 observed held-out states, including the initial state, using the same query locations for H, P, and R. This retrospective query comparison separates changes in the native head from changes in rollout locations. U remains in the checkpoint's normalized units and is not propagated into a trajectory variance. Table 18 also reports U_0, queried only at the initial state available when forecasting.

A cell is a pair and a target test sequence. Within each cell, we average RMSE and uncertainty over the five context draws, form paired ratios, and then take medians across cells. Thus a median error ratio need not equal the ratio of median errors. There are 16 cells, based on eight distinct test sequences appearing in two pairs each. Tables summarize these overlapping cells descriptively; they do not attach population confidence intervals or significance tests.

All 135 specified contexts produced finite forecasts. Halving the RK4 step changed RMSE by at most 0.000466 in the routed check and 0.000339 for the first draw of all three primary pooled contexts (less

than 0.1% relatively). Protocol, data hashes, context indices, preprocessing parameters, predictions, and per-cell results accompany the experiment.

D.3 ALL PAIR AND CONTEXT-SIZE RESULTS

The primary comparison has higher pooled error in 12/16 cells and higher pooled uncertainty in 5/16; 7/16 have higher error without higher uncertainty. Routing improves 12/16 cells. The aggregate error and uncertainty ratios are 1.308 and 0.789, while the routed/pooled error ratio is 0.838. These are medians of paired ratios, not ratios of the absolute medians in Table 19.

The response is nonuniform. For 09–35, native uncertainty increases at every context size. At $n = 2$, aggregate uncertainty also increases slightly. At $n = 6$, routing in 35–39 has a median error ratio just above one and improves only half of its cells. Table 18 retains these outcomes. The real-data comparison has neither known vector-field truth nor an exact adjacent-pair-matched control; it tests the operational effect of mixing recorded groups rather than isolating the mechanism established by the simulations. Its scope is limited to this representation and forecast horizon.

Table 18: All context sizes and record-group pairs. Ratios are defined above. The final columns count higher pooled error and lower routed error, respectively; denominators are cells.

n	Groups	E_P/E_H	U_P/U_H	$U_{0,P}/U_{0,H}$	E_R/E_P	$E_P > E_H$	$E_R < E_P$
2	09–35	1.035	1.177	0.828	1.312	3/6	3/6
2	09–39	2.017	0.956	0.883	0.647	4/4	4/4
2	35–39	1.845	0.937	0.586	0.799	6/6	4/6
2	All	1.470	1.010	0.803	0.853	13/16	11/16
4	09–35	1.229	1.167	0.806	0.905	4/6	4/6
4	09–39	1.835	0.779	0.822	0.575	4/4	4/4
4	35–39	1.161	0.752	0.648	0.948	4/6	4/6
4	All	1.308	0.789	0.759	0.838	12/16	12/16
6	09–35	1.388	1.318	1.039	0.762	4/6	4/6
6	09–39	2.055	0.814	0.810	0.475	4/4	4/4
6	35–39	1.194	0.812	0.725	1.002	5/6	3/6
6	All	1.512	0.846	0.789	0.719	13/16	11/16

Table 19: Absolute medians for the primary comparison. Each method's RMSE and native uncertainty are first averaged over context draws within a cell. Native U has the checkpoint's normalized units. Persistence fixes the initial state; a native uncertainty value is not defined for it.

Groups	Quantity	Same-group H	Mixed P	Routed R	Persistence
09–35	Latent RMSE	0.481	0.462	0.493	1.030
09–35	50-channel RMSE	5.365	4.651	5.427	8.143
09–35	Native U	0.826	0.690	0.938	—
09–39	Latent RMSE	0.539	1.013	0.535	1.592
09–39	50-channel RMSE	6.119	9.355	6.112	13.452
09–39	Native U	0.747	0.678	0.869	—
35–39	Latent RMSE	0.506	0.695	0.541	1.016
35–39	50-channel RMSE	5.385	6.131	5.532	8.143
35–39	Native U	0.731	0.498	0.939	—
All	Latent RMSE	0.506	0.576	0.535	1.030
All	50-channel RMSE	5.886	6.651	5.964	9.289
All	Native U	0.747	0.560	0.938	—

At the primary context size, same-group forecasts beat persistence on all eight distinct test sequences; mixed and routed forecasts each beat it in 14/16 pair–sequence cells in latent coordinates. We also inverse-project forecasts into the 50 benchmark channels and compute RMSE there after the same

zero-channel treatment. The pooled/same-group median error ratio is 1.142 and the routed/pooled ratio is 0.898 in this representation. The median RMSE floor from projecting the observed future states into the shared PCA subspace and back is 4.277. This floor uses future states only to quantify representation loss; it is not a forecasting baseline. The corresponding floors for 09–35, 09–39, and 35–39 are 3.656, 5.106, and 3.656. Absolute RMSE in these channels reflects both discarded motion information and forecast error, and does not measure error against known physical parameters.

Table 20: Primary comparison by target group within each pair. These rows expose both directions of each pooling comparison; U uses the common held-out query states.

Pair → target	Cells	E_P/E_H	U_P/U_H	E_R/E_P	$E_P > E_H$	$E_R < E_P$
09–35 → 09	2	1.269	1.499	0.761	2/2	2/2
09–35 → 35	4	1.027	0.813	1.120	2/4	2/4
09–39 → 09	2	1.781	0.949	0.576	2/2	2/2
09–39 → 39	2	2.056	0.729	0.565	2/2	2/2
35–39 → 35	4	1.064	0.669	1.037	2/4	2/4
35–39 → 39	2	1.695	0.752	0.695	2/2	2/2