ADAPTING FOUNDATION MODELS WITH MECHANISTIC EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models present a promising paradigm for producing estimators with broad domain understanding, serving as strong priors that can be adapted to specialized settings. Current practices often rely on scaling generic architectures, domain-specific fine-tuning, and retrieval-augmented workflows. These paradigms are practically feasible given their deployment simplicity and ability to reuse existing infrastructure, but they often struggle to faithfully embody mechanistic domain knowledge and respect known symmetries or conservation laws in structured environments. This work introduces Mechanistic Residual Adaptation (MechRA), a framework that orchestrates the utilization of domain-specific knowledge characterized by expert models during downstream adaptation. A foundation model is used as global signal guidance, while observations from expert models (which demonstrate key invariances, symmetries, and other properties) facilitate mechanistic adjustment towards more plausible, task-specific outputs. We evaluate this approach under various foundation model types and on a diverse set of downstream tasks, and observe the method's ability to learn and orchestrate task-relevant knowledge provided by expert models translates to performance advantages over other fine-tuning methods and task-specific models.

1 INTRODUCTION

Foundation models offer reusable predictive representations that can be adapted to specialized tasks with relatively limited supervision. In scientific and engineering settings, this statistical knowledge often coexists with a different resource: mechanistic models that describe particular processes, constraints, or relationships within the target system. A simulator may characterize plausible dynamics, a reduced physical model may estimate a relevant flux, and a geometric expert may identify structure in an observed trajectory. These sources can supply case-specific information that complements a foundation model's prediction. The central question is how to learn when and how to use that information while preserving the utility of the pretrained predictor.

Physics-informed learning provides several routes for incorporating domain knowledge, including governing-equation penalties and architectures designed around physical structure Raissi et al. (2019); Yin et al. (2021). Such approaches are especially natural when the relevant constraints and their relationship to the prediction task can be specified during model design. Adapting an existing foundation model presents an additional opportunity: mechanistic models can remain available as external sources of guidance, queried using the observations and conditions of each new case. This interface accommodates frozen predictors and established domain tools, including simulators whose internal computations are inaccessible or non-differentiable. The learned component then determines how their outputs should influence the downstream prediction.

Using mechanistic guidance in this way requires more than making expert outputs available. Experts may describe different quantities, operate at different resolutions, or capture overlapping aspects of a system. Their approximations can also be useful in some regimes and misleading in others. Consequently, the adapter must learn relationships among the observed context, the base prediction, and the supplied evidence. A useful integration mechanism should preserve source-specific content while allowing complementary observations to inform a coordinated correction. We treat mechanistic structure as evidence whose predictive relevance is learned from the task, with expert information remaining available at inference time.

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.

We introduce *Mechanistic Residual Adaptation* (MechRA), an external adapter that learns this integration while keeping the foundation model frozen. Grounded observation maps connect task inputs to expert conditioning spaces. Expert outputs are encoded into continuous tokens and localized through gated writes to a context-dependent shared workspace. A single fusion update integrates these messages, and an output-query decoder produces a residual correction to the base forecast. Optional pooled or attention-based bypasses give the decoder direct access to encoded evidence alongside the shared representation. This arrangement allows the model to learn both where expert information is relevant and how it should contribute to the task. Mechanistic guidance is supplied dynamically for each test context; the trained adapter interprets it through a fixed forward pass.

Our contributions are as follows:

1. A framework for foundation-model adaptation using heterogeneous, context-conditioned mechanistic evidence. It accommodates deterministic and probabilistic black-box experts through grounded observation interfaces, separating evidence generation from task-supervised neural integration.
2. MechRA, a concrete residual adaptation algorithm that coordinates expert representations through localized writes and shared latent integration, with optional direct retrieval of encoded evidence. Modular integration and routing choices support controlled comparisons of how expert information is used.

We study this approach in geometric trajectory forecasting, coupled thermal response prediction, and stochastic bistable dynamics, using time-series and tabular base predictors. Comparisons with native fine-tuning, expert-feature conditioning, pooled integration, and direct expert forecasts examine the value of both the evidence and its integration. Reconstruction tasks and expert-removal interventions further probe whether the adapter can recover and use informative expert content.

2 Background

Adapting pretrained predictors. Foundation models provide reusable prediction priors that can be specialized through labeled context or task-specific parameter updates. Forecasting and tabular models exemplify these interfaces Ansari et al. (2025); Hollmann et al. (2025); Qu et al. (2025). Parameter-efficient methods reduce the number of trainable parameters: bottleneck adapters insert modules into pretrained networks Houlsby et al. (2019), while LoRA learns low-rank weight updates Hu et al. (2022). MechRA uses an external residual interface, keeping the base predictor frozen and learning from public inputs, base predictions, and expert evidence. This makes the adaptation mechanism applicable when internal activations or backbone modification are inconvenient. The experimental comparisons vary both the adaptation mechanism and the evidence available to it, including native fine-tuning with and without expert features.

Mechanistic structure and hybrid modeling. Scientific machine learning incorporates physical knowledge through model structure, training objectives, and combinations of mechanistic and learned components. Physics-informed neural networks penalize governing-equation residuals Raissi et al. (2019); APHYNITY augments incomplete physical dynamics with a learned correction Yin et al. (2021). These approaches motivate using domain models to guide prediction while accounting for discrepancy. We consider a setting in which an existing predictor is accompanied by several approximate specialists. Their outputs may encode trajectories, fluxes, geometric estimates, or other quantities relevant to the task. MechRA learns how these observations should influence a residual correction. The integration architecture is generic, while the domain-specific structure enters through the supplied expert evidence. Physical consistency consequently depends on both expert fidelity and the learned use of that evidence.

External information at inference time. Retrieval-augmented generation conditions predictions on retrieved passages Lewis et al. (2020), and tool-augmented language models learn to incorporate outputs of external computations Schick et al. (2023). These interfaces illustrate how pretrained knowledge can be complemented by information supplied for the current input. Mechanistic adaptation follows this principle with numerical, context-conditioned expert outputs. A simulator can produce plausible future trajectories, while a reduced model can expose a particular physical response.

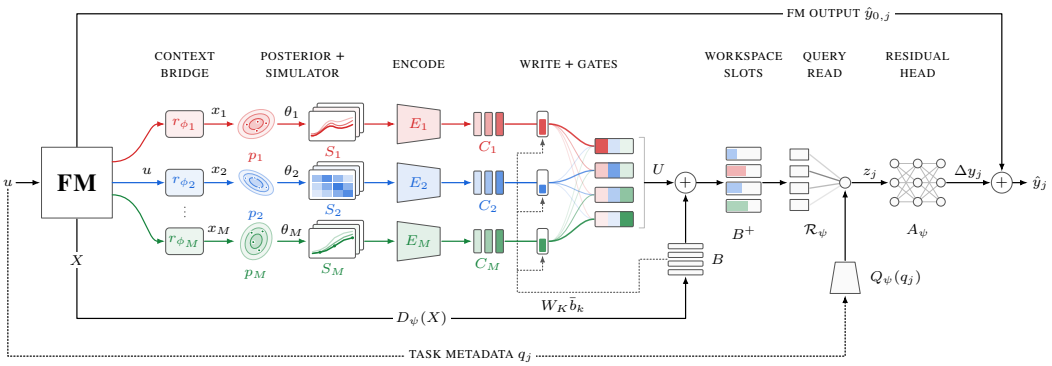


Figure 1: **MechRA: mechanistic adaptation in one pass.** Grounded observations and posterior-conditioned simulation produce tokens C_i. Context features $X = \chi(u, \hat{y}_0)$ initialize the workspace B; addresses and gates localize expert messages before nonlinear fusion yields $B^+ = B + U$. Output queries read B^+ to predict an additive correction. Colors distinguish experts, and routing strengths are schematic. The figure shows the core path; slot interaction and encoded-evidence bypasses extend it as described in the text.

Our setting assumes grounded expert queries and learns the downstream integration of their results. The same process supplies evidence for held-out contexts; model parameters remain fixed during inference. This distinguishes case-dependent guidance from parameter adaptation performed at test time.

Grounded simulation and uncertainty. Simulation-based inference connects observations to parameter uncertainty when likelihood evaluation is difficult Cranmer et al. (2020). Given an observation map and an available posterior, posterior-conditioned simulation yields an ensemble of responses compatible with the observed context under the expert model. Deterministic domain tools provide a simpler instance of the same evidence interface. The observation map establishes how task inputs relate to simulator inputs, and posterior quality determines which parameter uncertainty is represented. Separating evidence generation from neural integration permits black-box experts and independently trained inference procedures. Uncertainty in their outputs remains subject to model discrepancy and the adequacy of the conditioning data.

Representing and coordinating heterogeneous evidence. Deep Sets formalizes invariant aggregation of unordered collections Zaheer et al. (2017), and Set Transformer introduces attention-based interactions and pooling Lee et al. (2019). These provide architectural precedents for encoding variable-sized simulation ensembles with masks and coordinates. Perceiver IO separates input encoding, latent processing, and output-query decoding Jaegle et al. (2022); Slot Attention organizes features through competitive assignments Locatello et al. (2020). MechRA uses a shared latent interface to align heterogeneous expert representations with the prediction context. Its gated, destination-normalized writes allocate evidence to context-dependent slots before one fusion update. Localization supplies a mechanism for selective integration; interpreting individual slots as physical factors requires further validation.

3 Methodology

At a high level, Mechanistic Residual Adaptation (MechRA) adapts a frozen base predictor model to a downstream task using context-conditioned expert evidence at inference time. Expert encoders map sampled outputs into continuous tokens, a gated write mechanism localizes their influence in a common latent workspace, and an output-query decoder predicts a residual correction from the workspace embeddings and task context. At inference time, a single workspace update is performed to dynamically infuse expert model guidance and mechanistically steer the foundation model's output (see Figure 1).

3.1 TASK INTERFACE AND EXPERT MODEL CONDITIONING

Let u denote observed context, y a downstream supervised target, and $\hat{y}_0 = F(u)$ a prediction from a frozen base model. We let $X = \chi(u, \hat{y}_0)$ represent features exposed to the adapter. Our experiments construct X from public inputs and base outputs; internal FM activations could also be included. Expert $i \in \{1, \dots, M\}$ has conditioning space $\mathcal{X}_i$, parameter θ_i, and simulator S_i. A supplied map $x_i = H_i(u)$ or a paired-data-trained bridge $x_i \sim r_{\phi_i}(\cdot \mid u)$ grounds its observation. Given a posterior,

$$\theta_{i\ell} \sim p_i(\theta_i \mid x_i), \qquad v_{i\ell} \sim S_i(\cdot \mid \theta_{i\ell}, \xi_i(u)), \qquad \ell = 1, \dots, L_i, \tag{1}$$

where $\xi_i(u)$ contains known forecast conditions. Deterministic experts produce singleton evidence. We assume that observation bridges and posterior inference are available independently of adapter training. The experimental bridges are fixed domain maps; evidence is generated per case, detached, and cached.

3.2 EXPERT EMBEDDINGS AND LOCALIZATION

An initializer produces $B = D_\psi(X) \in \mathbb{R}^{K \times d}$, a set of context-dependent slots. Encoder E_i maps expert values, coordinates, optional parameters, and a learned identity e_i to tokens $C_i = (c_{it})_{t=1}^{N_i}$. The index t ranges over draws and their positions. Let ν_{it} mark valid tokens, η_i mark expert availability, and $\bar{c}_{it} = \mathrm{LN}(c_{it})$, $\bar{b}_k = \mathrm{LN}(b_k)$. Writes use destination-normalized addresses and scalar gates:

$$a_{itk} = \mathrm{softmax}_k \left((W_{Q,i}\bar{c}_{it})^\top W_K \bar{b}_k / \sqrt{d} \right), \qquad g_{it} = \eta_i \nu_{it} \sigma\big(G_i[\bar{c}_{it}; \bar{b}_{\mathrm{avg}}]\big), \tag{2}$$

$$m_{ik} = D_{ik}^{-1} \sum_t a_{itk} g_{it} W_{V,i} \bar{c}_{it}, \qquad D_{ik} = \max(1, \textstyle\sum_t \nu_{it}), \quad \bar{b}_{\mathrm{avg}} = K^{-1} \textstyle\sum_k \bar{b}_k. \tag{3}$$

Addresses specify *where*, projected values *what*, and gates *how much* to write. Each source token distributes its influence over slots. A nonlinear fusion transforms the summed messages before the residual update:

$$m_k = \sum_i m_{ik}, \qquad U_k = \Phi_\psi([\bar{b}_k; m_k]) - \Phi_\psi([\bar{b}_k; 0]), \qquad B^+ = B + U. \tag{4}$$

The subtraction makes $U = 0$ when all messages vanish. Localization expresses selective influence on learned computational channels.

3.3 RESIDUAL DECODING AND LEARNING

For output index j, q_j concatenates task coordinates with available base-output features. A query embedding reads the updated slots:

$$z_j = \mathcal{R}_\psi(Q_\psi(q_j), B^+), \qquad \Delta y_j = A_\psi(z_j), \qquad \hat{y}_j = \hat{y}_{0,j} + \Delta y_j. \tag{5}$$

The read block $\mathcal{R}_\psi$ includes cross-attention, a query residual, and a residual feed-forward block. Context reaches the head through initialized slots, and base-output features also enter through the query. Neural parameters are trained with task loss, usually mean squared error, while the base predictor and raw evidence remain fixed. A zero-initialized final projection begins training at the base prediction. The adapter can subsequently learn a context-only correction when expert evidence is absent.

Algorithm 1 summarizes the canonical forward pass. All experts address the same initial workspace, followed by a single fusion update. Appendix C gives the evidence preparation and training procedure.

3.4 MODULAR EXTENSIONS

Normalization and fusion. Slot-mass normalization replaces the denominator in Eq. (3) by $D_{ik} = \max\{\sum_t \eta_i \nu_{it} a_{itk}, \epsilon \max(1, \sum_t \nu_{it})/K\}$, with $\epsilon = 10^{-4}$. Excluding learned gates from this denominator preserves their suppression effect. Under uniform routing, an active expert contributes mean gated values at each slot, avoiding the core normalization's $1/K$ attenuation. Expert-aware fusion instead replaces m_k in Eq. (4) with $[m_{1k}; \dots; m_{Mk}]$, preserving source-specific blocks until the fusion MLP. An optional Transformer stack gives $B^+ = T_\psi(B + U)$, adding interactions among updated slots. These choices are independently configurable.

Algorithm 1 MechRA: canonical forward pass

Require: Base prediction $\hat{y}_0$, context features X, output queries (q_j);
detached expert evidence $(\mathcal{V}_i)$ and availability (η_i)
Ensure: Corrected prediction $\hat{y}$

$B \leftarrow D_\psi(X)$; $\bar{B} \leftarrow \mathrm{LN}(B)$; $\bar{b}_{\mathrm{avg}} \leftarrow K^{-1} \sum_k \bar{b}_k$
$m_k \leftarrow 0$ for each slot $k = 1, \ldots, K$
for each available expert i **do**
  $(C_i, \nu_i) \leftarrow E_i(\mathcal{V}_i; e_i)$ ▷ Encode values, coordinates, and parameters
  $\bar{c}_{it} \leftarrow \mathrm{LN}(c_{it})$ for each token t
  $a_{itk} \leftarrow \mathrm{softmax}_k((W_{Q,i}\bar{c}_{it})^\top W_K \bar{b}_k / \sqrt{d})$
  $g_{it} \leftarrow \eta_i \nu_{it} \sigma(G_i[\bar{c}_{it}; \bar{b}_{\mathrm{avg}}])$; $D_i \leftarrow \max(1, \sum_t \nu_{it})$
  $m_k \leftarrow m_k + D_i^{-1} \sum_t a_{itk} g_{it} W_{V,i} \bar{c}_{it}$ for all k
$U_k \leftarrow \Phi_\psi([\bar{b}_k; m_k]) - \Phi_\psi([\bar{b}_k; 0])$ for all k
$B^+ \leftarrow B + U$ ▷ One shared update
for each output index j **do**
  $z_j \leftarrow \mathcal{R}_\psi(Q_\psi(q_j), B^+)$
  $\hat{y}_j \leftarrow \hat{y}_{0,j} + A_\psi(z_j)$
return $\hat{y}$

Encoded-evidence bypasses. Let $c_0 = c_\psi(X)$ be a masked mean of tokenwise projected context features. A pooled bypass concatenates masked expert means; an attention bypass retrieves encoded tokens using a context- and workspace-dependent query:

$$p_i = \eta_i \frac{\sum_t \nu_{it} c_{it}}{\max(1, \sum_t \nu_{it})}, \qquad e_j^{\mathrm{pool}} = [p_1; \ldots; p_M], \tag{6}$$

$$r_j = R_\psi([q_j; c_0; z_j]), \qquad e_j^{\mathrm{attn}} = \mathcal{R}_\psi^E(r_j, [C_1; \ldots; C_M]), \tag{7}$$

$$\Delta y_j = A_\psi([q_j; c_0; e_j; z_j]). \tag{8}$$

The attention reader masks invalid or unavailable evidence and returns zero when none is active. Both bypasses share expert encoders and availability masks with the workspace route. We train joint, workspace-only, and bypass-only versions of these architectures. Disabling a route zeros its head-input block; workspace removal also zeros z_j in the attention query. Direct context c_0 and raw queries q_j remain in the extended head. Thus its workspace-only variant retains an additional context interface compared with Eq. (5). Same-checkpoint route interventions separately measure reliance after joint training.

4 Experimental results

4.1 Evaluation environments

Three synthetic tasks vary the form and reliability of mechanistic evidence: geometric trajectory estimates, heterogeneous thermal specialists, and stochastic dynamics. Each held-out context has its own expert outputs, generated without future target observations and cached for reproducible comparisons. Appendix B details the generators and expert approximations.

SPHERE: geometric forecasting model. Given noisy three-dimensional positions, predict a clean future trajectory $p_k = r_k d_k$, where $d_{k+1} = R(a_{k+1/2}, \omega \Delta t) d_k$ and R is a Rodrigues rotation. Regimes use a fixed axis, a precessing axis, or a sinusoidally varying radius. A deterministic radius expert emits an estimated radius and its discrepancy from the base forecast; a rotation expert estimates angular velocity and rolls forward a radially projected observation. Both produce lead-indexed tokens. Linear extrapolation and Chronos-2 Ansari et al. (2025) serve as base predictors. The rotation forecast is a direct reference with particularly well-matched dynamics in the fixed-axis regime.

THERMAL: conditional response heat-dynamics. We predict air temperatures in three coupled zones given recent measurements, nominal building properties, and future weather and control

schedules. Following lumped building heat-dynamics models Bacher & Madsen (2011), each zone has air and storage temperatures (A_z, M_z) and lagged HVAC delivery $Q_{h,z}$:

$$C_{a,z}\dot{A}_z = Q_{a,z} + 0.4Q_{s,z} + Q_{\mathrm{int},z} + Q_{h,z}, \qquad C_{m,z}\dot{M}_z = Q_{m,z} + 0.6Q_{s,z}. \tag{9}$$

Here Q_a, Q_m collect conductive and ventilation exchanges and Q_s is solar gain. HVAC follows $t_h\dot{Q}_{h,z} = d_h P_z u_z - Q_{h,z}$ for command u_z. Regimes vary shading, ventilation, and equipment performance. Deterministic experts emit solar irradiance/gain from `pvlib` Holmgren et al. (2018), reduced-envelope temperature/heat flow, and HVAC delivery/electrical demand. Each supplies two-channel tokens indexed by lead and zone. Ridge and TabICL regression use scalar tabular queries and a disjoint labeled support bank Qu et al. (2025); Soda team, Inria (2026).

BISTABLE: stochastic forecasting. Observed two-dimensional states follow the discrete Langevin model

$$x_{n+1} = x_n - \nabla V_\theta(x_n)\Delta t + \sigma\sqrt{\Delta t}\,\varepsilon_n, \quad \varepsilon_n \sim \mathcal{N}(0, I_2), \tag{10}$$

$$V_\theta(x) = \frac{a}{4}\sum_{r=1}^{2}(x_r^2 - 1)^2 + \frac{\kappa}{2}(x_1 - x_2)^2 - \tau x_1, \quad \theta = (a, \kappa, \tau, \sigma). \tag{11}$$

A uniform finite-grid prior and Gaussian transition likelihood define an available parameter posterior. A stochastic expert samples this posterior and simulates uncoupled forecast dynamics ($\kappa = 0$); a local expert supplies linearized-rollout means and standard deviations; a geometry expert supplies posterior-averaged drift and potential on a spatial grid. Expert and target innovations are independent. Base choices include linear extrapolation and Chronos-2. The stochastic ensemble mean provides a direct reference; well-crossing probabilities furnish an auxiliary event task.

4.2 Comparisons and evaluation protocol

We compare raw FM predictions, native conditioning on fixed expert summaries, Chronos-2 LoRA adaptation with and without those summaries, and external residual adapters. The external no-expert control trains the workspace architecture on context alone. Pooled controls vary access to context and base-output features; workspace variants vary normalization, fusion, slot interaction, and bypass routes. Encoded-evidence adapters retain individual draws. These controls separate access to information from its integration strategy.

The table labels use the same route definitions across environments. "FM raw" is the frozen base forecast (Chronos-2 for SPHERE and BISTABLE, and the selected tabular baseline for THERMAL); "FM (experts)" exposes the expert summaries through the base model's native input interface. "FM+LoRA (no experts)" and "FM+LoRA (experts)" fine-tune the Chronos-2 backbone with the corresponding context, where that comparison is available. "FM+MechRA (no experts)" is the residual adapter trained without expert branches. The remaining MechRA rows retain the expert channel: "experts" is the canonical workspace route, "slot norm" changes the write normalization, "workspace" isolates the workspace path, and "full" enables that path together with the direct encoded-evidence bypass. Thus the rows compare both information access and the way evidence is routed to the residual head; they are not additional base-model families.

Mixed-mode training fits one model per case and seed across regimes. Independent-subtask mode fits separate models, support banks, and validation selections for each regime, then macro-averages subtask metrics. We report trajectory or temperature errors by regime and overall, with geometric and event metrics where applicable. Repeated runs summarize seed-level variation. Independently trained reconstruction tasks assess recovery of expert evidence; same-checkpoint removal and mismatch interventions assess reliance. Appendix D specifies budgets, metrics, and aggregation.

4.3 Overall performance

Across the reported tables, the full FM+MechRA configuration is generally the strongest or among the strongest methods after averaging over regime-specific metrics. Its advantage is most visible when the base forecast is imperfect and experts provide complementary structure: the workspace route can share evidence across experts, while the encoded-evidence bypass preserves task-relevant detail. The no-expert MechRA control often improves the frozen forecast, separating gains from

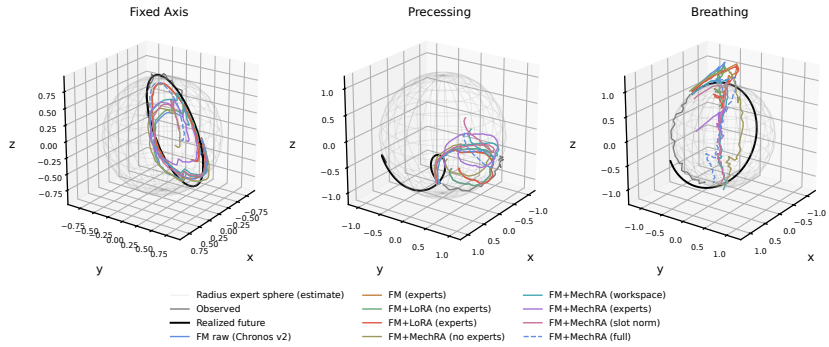


Figure 2: SPHERE trajectory forecasts across fixed-axis, precessing, and breathing regimes. Observed history and clean future positions are shown alongside base, adapted, and direct-expert predictions. The spherical reference surface visualizes the estimated radius. Axis precession and radial variation expose departures from the rotation expert's fixed-axis, fixed-radius approximation, illustrating how the usefulness of geometric guidance varies across regimes.

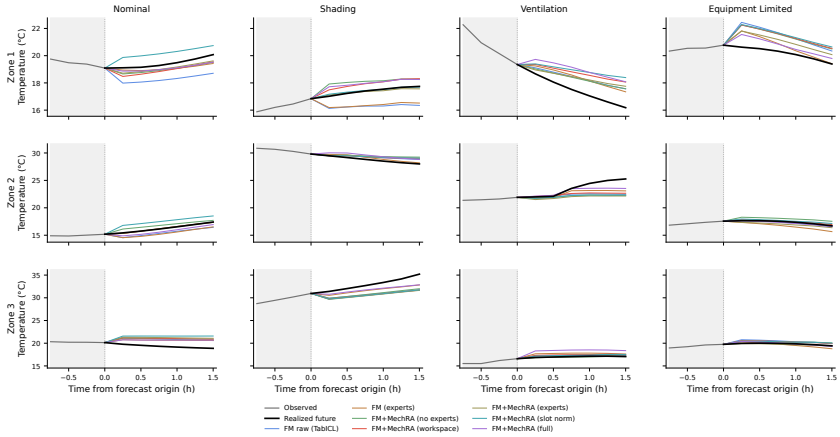


Figure 3: THERMAL response forecasts for a three-zone building with known future weather and control schedules. Predictions are compared with realized air temperatures under the displayed operating regimes. Solar, envelope, and HVAC experts provide complementary estimates of heat gain, temperature response, and delivered power; their approximations differ in sensitivity to shading, ventilation, and equipment limitations. The examples contextualize adaptation using heterogeneous physical quantities.

residual adaptation alone from gains due to mechanistic evidence. Native expert conditioning and LoRA help in selected regimes, but their performance is less consistent across heterogeneous expert outputs.

These results also expose meaningful limits. A direct expert can win when its inductive bias is unusually well matched to a regime, and stochastic references can remain difficult to exploit under point-error losses. Slot normalization and workspace-only variants are competitive in several settings, indicating that the shared route is useful even when the bypass is removed. The table-level trends should therefore be read as evidence for flexible evidence integration rather than as a claim that one route dominates every task; they aggregate small, controlled environments with deliberately varied expert quality.

The qualitative forecasts in Figures 2, 3, and 4 illustrate the regime-specific behavior summarized quantitatively in Tables 1, 2, and 3. Together they show both the shape of the forecasting problems and the corresponding route-level comparisons.

Table 1: SPHERE forecasting with Chronos-2 at the medium resource tier. Entries are mean errors across seeds, by regime and overall. Lower is better; bold and underlining mark the best and second distinct scores among displayed methods, ranked before rounding.

Method	Breathing					Fixed Axis				
	RMSE	MAE	ADE	FDE	Final MAE	RMSE	MAE	ADE	FDE	Final MAE
FM raw (Chronos v2)	0.486	0.318	0.646	0.964	0.465	0.522	0.321	0.643	1.074	0.529
FM+MechRA (no experts)	0.356	0.262	0.526	0.812	0.415	0.391	0.275	0.550	0.921	0.457
FM+LoRA (no experts)	0.463	0.306	0.620	0.930	0.450	0.498	0.308	0.617	1.029	0.509
FM+LoRA (experts)	0.432	0.277	0.567	0.885	0.429	0.471	0.282	0.569	0.965	0.469
FM (experts)	0.446	0.284	0.582	0.910	0.442	0.483	0.288	0.582	0.991	0.481
FM+MechRA (experts)	0.382	0.289	0.586	0.860	0.428	0.370	0.273	0.544	0.902	0.456
FM+MechRA (slot norm)	0.349	0.255	0.520	0.875	0.454	0.337	0.242	0.484	0.877	0.436
FM+MechRA (workspace)	0.338	0.254	0.512	0.809	0.413	0.336	0.239	0.479	0.869	0.434
FM+MechRA (full)	**0.333**	**0.240**	**0.491**	**0.763**	**0.388**	**0.320**	**0.222**	**0.446**	**0.816**	**0.404**

Method	Precessing					Overall				
	RMSE	MAE	ADE	FDE	Final MAE	RMSE	MAE	ADE	FDE	Final MAE
FM raw (Chronos v2)	0.729	0.508	1.023	1.564	0.781	0.590	0.379	0.762	1.200	0.591
FM+MechRA (no experts)	0.580	0.428	0.856	1.239	0.623	0.455	0.322	0.644	0.998	0.499
FM+LoRA (no experts)	0.715	0.498	1.003	1.538	0.767	0.570	0.367	0.738	1.163	0.574
FM+LoRA (experts)	0.718	0.497	1.001	1.563	0.781	0.554	0.348	0.703	1.131	0.555
FM (experts)	0.724	0.502	1.009	1.576	0.788	0.564	0.354	0.715	1.153	0.566
FM+MechRA (experts)	0.588	0.441	0.887	**1.229**	**0.619**	0.452	0.329	0.659	0.992	0.499
FM+MechRA (slot norm)	0.585	0.420	0.851	1.251	0.623	0.434	0.301	0.607	0.990	0.496
FM+MechRA (workspace)	**0.570**	0.416	0.834	1.246	0.630	0.425	0.297	0.596	0.975	0.491
FM+MechRA (full)	0.578	**0.411**	**0.833**	1.256	0.632	**0.421**	**0.285**	**0.576**	**0.942**	**0.472**

Table 2: THERMAL response prediction with TabICL at the small resource tier. Entries are mean temperature errors across seeds, by operating regime and overall. Lower is better; bold and underlining mark the best and second distinct scores among displayed methods, ranked before rounding.

Method	Equipment Limited			Nominal			Shading		
	RMSE	MAE	Final MAE	RMSE	MAE	Final MAE	RMSE	MAE	Final MAE
FM raw (TabICL)	1.521	1.163	1.165	1.313	1.042	1.231	1.407	1.064	1.257
FM+MechRA (no experts)	1.347	1.074	1.088	1.196	0.937	1.092	1.224	0.917	1.126
FM (experts)	1.457	1.168	1.355	0.963	0.752	0.903	0.946	0.753	0.941
FM+MechRA (experts)	1.409	1.093	1.101	1.254	0.986	1.159	1.304	0.971	1.204
FM+MechRA (slot norm)	1.412	1.118	1.111	1.239	0.966	1.129	1.221	0.910	1.108
FM+MechRA (workspace)	1.277	1.008	**0.997**	1.227	0.975	1.131	1.202	0.897	1.093
FM+MechRA (full)	**1.116**	**0.862**	1.034	**0.870**	**0.681**	**0.801**	**0.863**	**0.678**	**0.781**

Method	Ventilation			Overall		
	RMSE	MAE	Final MAE	RMSE	MAE	Final MAE
FM raw (TabICL)	1.365	1.088	1.269	1.406	1.089	1.237
FM+MechRA (no experts)	1.167	0.948	1.111	1.235	0.967	1.109
FM (experts)	1.039	0.810	1.010	1.127	0.862	1.047
FM+MechRA (experts)	1.216	0.956	1.142	1.297	0.999	1.155
FM+MechRA (slot norm)	1.264	1.028	1.217	1.295	1.007	1.151
FM+MechRA (workspace)	1.185	0.965	1.123	1.229	0.962	1.095
FM+MechRA (full)	**0.949**	**0.750**	**0.911**	**0.956**	**0.740**	**0.881**

5 CONCLUSION

We presented MechRA, a residual adaptation framework that equips a frozen foundation model with dynamically supplied mechanistic evidence. The method encodes heterogeneous expert outputs, localizes them into a shared workspace, and combines the resulting representation with a base forecast through a residual head. A direct encoded-evidence route complements the workspace when the task

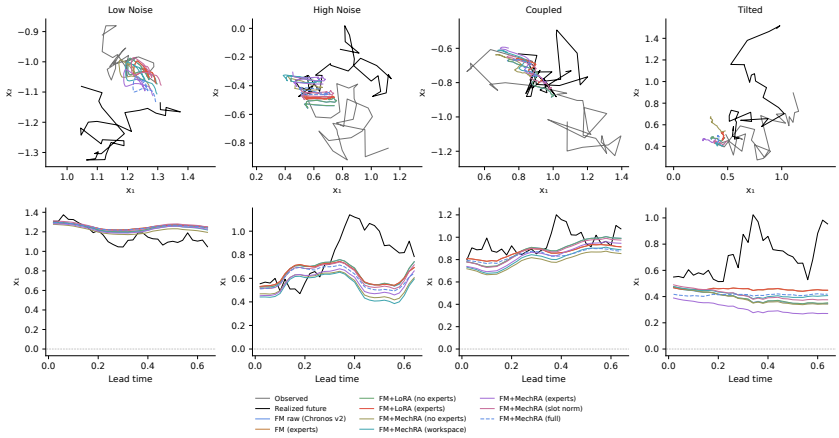


Figure 4: BISTABLE forecasts across noise, coupling, and tilt regimes. Phase-space trajectories (top) and the first state coordinate versus forecast lead (bottom) compare the realized future with base, adapted, and direct-expert predictions. The stochastic direct forecast is the expert ensemble mean. Independent process noise separates the realized path from its conditional mean, so smooth point forecasts can remain appropriate even when individual trajectories fluctuate substantially.

Table 3: BISTABLE forecasting at the medium resource tier with independent subtask training. Entries are mean errors across seeds; Overall averages subtask metrics equally. Lower is better; bold and underlining mark the best and second distinct scores among displayed methods, ranked before rounding.

Method	**Coupled**					**High Noise**					**Low Noise**				
	RMSE	MAE	ADE	FDE	Final MAE	RMSE	MAE	ADE	FDE	Final MAE	RMSE	MAE	ADE	FDE	Final MAE
FM raw (Chronos v2)	0.274	0.190	0.300	0.426	0.268	0.330	0.252	0.395	0.536	0.339	0.123	0.095	0.150	0.199	0.127
FM+MechRA (no experts)	0.244	0.172	0.269	**0.359**	**0.228**	0.321	0.247	0.389	0.512	**0.325**	0.121	0.094	0.148	0.193	0.123
FM+LoRA (no experts)	0.274	0.190	0.300	0.425	0.267	0.329	0.251	0.394	0.535	0.339	0.123	0.095	0.149	0.198	0.126
FM+LoRA (experts)	0.266	0.184	0.292	0.410	0.257	0.324	0.247	0.388	0.524	0.333	0.121	0.094	0.147	0.194	0.124
FM (experts)	0.266	0.184	0.292	0.410	0.257	0.325	0.247	0.389	0.525	0.333	0.121	0.094	0.147	0.194	0.124
FM+MechRA (experts)	0.254	0.179	0.282	0.380	0.240	0.323	0.247	0.390	0.520	0.331	0.122	0.095	0.149	0.195	0.124
FM+MechRA (slot norm)	0.244	0.171	0.269	0.364	0.230	0.324	0.250	0.393	0.520	0.333	0.117	0.092	0.144	0.186	0.119
FM+MechRA (workspace)	0.247	0.173	0.272	0.368	0.233	0.320	0.247	0.388	0.514	0.329	**0.116**	**0.091**	**0.143**	**0.183**	**0.117**
FM+MechRA (full)	**0.243**	**0.171**	**0.268**	0.362	0.228	**0.319**	**0.246**	**0.387**	**0.512**	0.328	0.117	0.091	0.144	0.189	0.119

Method	**Tilted**					**Overall**				
	RMSE	MAE	ADE	FDE	Final MAE	RMSE	MAE	ADE	FDE	Final MAE
FM raw (Chronos v2)	0.256	0.179	0.281	0.384	0.244	0.246	0.179	0.281	0.386	0.244
FM+MechRA (no experts)	0.242	0.172	0.270	0.356	0.227	0.232	0.171	0.269	0.355	0.225
FM+LoRA (no experts)	0.256	0.179	0.280	0.383	0.244	0.246	0.179	0.281	0.385	0.244
FM+LoRA (experts)	0.249	0.174	0.273	0.369	0.234	0.240	0.175	0.275	0.374	0.237
FM (experts)	0.249	0.174	0.273	0.369	0.234	0.240	0.175	0.275	0.375	0.237
FM+MechRA (experts)	0.247	0.173	0.273	0.367	0.233	0.236	0.174	0.274	0.365	0.232
FM+MechRA (slot norm)	0.244	0.173	0.272	0.354	0.226	0.232	0.171	0.270	0.356	0.227
FM+MechRA (workspace)	0.243	0.174	0.273	0.356	0.228	0.232	0.171	0.269	0.355	0.227
FM+MechRA (full)	**0.239**	**0.169**	**0.266**	**0.348**	**0.222**	**0.229**	**0.169**	**0.266**	**0.353**	**0.224**

depends on details that are costly to compress. This design keeps the foundation model's broad prior intact while learning how to use domain evidence at inference time.

Across geometric, thermal, and stochastic environments, the full configuration usually improves on the corresponding frozen forecast and remains competitive with alternatives that expose expert summaries directly or fine-tune the base model. Controlled ablations show that the gains depend on both the evidence channel and its integration: workspace sharing helps coordinate specialists, while the bypass is useful for strong or highly localized signals. Direct experts can still be best when their inductive bias matches a regime closely, and stochastic evidence remains sensitive to posterior quality and the training objective. These observations support mechanistic adaptation as a flexible interface rather than a universal replacement for specialized simulators.

AI USE STATEMENT

We have not used generative AI tools for any of the required disclosures, namely generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refine hypotheses, designing or providing feedback on research methodology or experiments, most of which are not applicable to this work.

We did use generative AI tools for several of the recommended disclosure tasks, namely assisting with software creation, modifying scientific figures, suggesting paper structure, helping to draft section and identify relevant literature, and proposing keywords. We have reviewed all AI-assisted work with manual literature review and LLM-generated code was verified and extensively tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Abdul Fatir Ansari et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. URL https://arxiv.org/abs/2510.15821.

Peder Bacher and Henrik Madsen. Identifying suitable models for the heat dynamics of buildings. *Energy and Buildings*, 43(7):1511–1522, 2011. doi: 10.1016/j.enbuild.2011.02.005.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. doi: 10.1073/pnas.1912789117. URL https://arxiv.org/abs/1911.01429.

Noah Hollmann et al. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6.

William F. Holmgren, Clifford W. Hansen, and Mark A. Mikofski. pvlib python: a python package for modeling solar energy systems. *Journal of Open Source Software*, 3(29):884, 2018. doi: 10.21105/joss.00884. URL https://joss.theoj.org/papers/10.21105/joss.00884.

Neil Houlsby et al. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Andrew Jaegle et al. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2107.14795.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Francesco Locatello et al. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.15055.

Jingang Qu et al. TabICL: A tabular foundation model for in-context learning on large data. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2502.05564.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

Patrick Seifner, Kostadin Cvejoski, David Berghaus, Cesar Ojeda, and Ramses J. Sanchez. In-context learning of stochastic differential equations with foundation inference models. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2502.19049. Updated arXiv version 3, June 2026.

Soda team, Inria. TabICLv2: An open tabular foundation model. Official software repository, 2026. URL https://github.com/soda-inria/tabicl. Regression implementation and checkpoint documentation; accessed September 25, 2026.

Yuan Yin et al. Augmenting physical models with deep networks for complex dynamics forecasting. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.04456.

Manzil Zaheer et al. Deep sets. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1703.06114.

Appendix

A Additional positioning **13**

B Environment generators and experts **13**

C Adapter implementation and assumptions **14**

D Evaluation and implementation protocol **15**

A ADDITIONAL POSITIONING

Scope of mechanistic evidence. Expert simulations combine observed context with modeling assumptions. For stochastic dynamics, their independent future draws describe possible outcomes, and point-error objectives favor conditional means. FIM-SDE estimates drift and diffusion from observed trajectories Seifner et al. (2025), providing a related modeling approach to our use of experts with a declared grid posterior. The thermal pvlib branch contributes established solar-modeling software Holmgren et al. (2018) within our synthetic building generator. These experiments probe information integration under controlled mismatch; field calibration and real-system validation remain separate evaluation questions.

Shared representations and direct retrieval. Summing messages compresses source distinctions before nonlinear fusion. Concatenation fusion preserves those distinctions through the joint MLP, while slot interaction mixes information across latent locations. Bypasses expose summaries or token detail directly to the residual head. These choices offer complementary representations, whose utility depends on the task, expert quality, and optimization. Capacity-controlled comparisons and route-specific training help distinguish representational effects from the consequences of adding parameters or context inputs.

Model families and releases. TabPFN and FIM-SDE provide literature context and are excluded from the reported experimental comparisons. Experimental configurations specify the selected checkpoints and software versions; TabICL's regression interface and Chronos-2's series-forecasting model are documented by their release sources Soda team, Inria (2026); Ansari et al. (2025). Our comparisons concern simulator conditioning, expert encoding, shared integration, and direct retrieval under stated information and resource budgets.

B ENVIRONMENT GENERATORS AND EXPERTS

Sphere specifics. The default physical settings are $\Delta t = 0.12$, initial radius uniform on $[0.8, 1.5]$, angular speed uniform on $[0.35, 1.1]$, and independent Gaussian observation noise with standard deviation 0.025. Fixed, precessing, and breathing regimes have probabilities 0.5, 0.3, and 0.2; precession and breathing are mutually exclusive. The radius estimate is the median norm over at most 16 recent steps. Rotation uses regularized least squares for $v \simeq \omega \times x$ over at most 24 observations, with estimated ω supplied as parameter metadata. The offline predictor fits a line over the last eight observations.

Thermal equations and mismatch. The exchange terms in Eq. (9) are

$$Q_{a,z} = (G_{o,z} + G_{v,z})(T_o - A_z) + G_{m,z}(M_z - A_z) + \sum_j G_{zj}(A_j - A_z), \tag{12}$$

$$Q_{m,z} = G_{m,z}(A_z - M_z) + G_{mo,z}(T_o - M_z). \tag{13}$$

Capacitances are in J/K, conductances in W/K, heat rates in W, temperatures in °C, and time in seconds. Symmetric inter-zone conductances conserve exchanged heat. Ventilation uses $G_v = \rho c_p V \,\mathrm{ACH}(1 + 0.08w)/3600$ for wind speed w. Solar gain is $Q_{s,z} = A_{w,z}\tau_{w,z}(1 - b_z)I_z$, with incident irradiance I_z, window area A_w, transmittance τ_w, and blind attenuation b_z. Future forcing schedules are supplied to predictors and experts. Nominal coefficients are public, but the first five thermal coefficients per zone are independently perturbed by factors in $[0.88, 1.12]$ for generation. The four regimes are sampled uniformly. Equipment limitation uses $d_h \in [0.45, 0.70]$ and $t_h = 1200$ seconds, versus nominal values 1 and 300 seconds. Observations are spaced 900 seconds apart, with 60-second explicit temperature updates, exponential actuator updates, and 24 warmup intervals. Measurement noise is 0.08 degrees Celsius. Numerical convergence has not been established by this setup. Reference solar transposition uses Hay–Davies; the pvlib expert uses isotropic transposition. Shared pvlib geometry and different transposition models provide controlled solar mismatch. The envelope uses $C_a + 0.15C_m$, omits zone coupling, and uses crude solar and instantaneous HVAC terms. HVAC starts at zero delivery and uses nominal capacity and lag. Targets are trained as $(T - 20)/10$ and reported in physical temperature units. Tabular inputs summarize observed history; storage temperatures and coefficient perturbations remain hidden.

Algorithm 2 MechRA: grounded evidence preparation and adapter training

Require: Training pairs $\mathcal{D} = \{(u, y)\}$; frozen F; grounded bridges and experts; draw counts (L_i), dropout probability ρ, task loss $\mathcal{L}$

Initialize adapter parameters ψ; zero the final output projection
for each $(u, y) \in \mathcal{D}$ **do**
  $\hat{y}_0 \leftarrow F(u)$; $X \leftarrow \chi(u, \hat{y}_0)$; construct (q_j) from task coordinates and $\hat{y}_0$
  for each available expert i **do**
    Obtain $x_i = H_i(u)$ or $x_i \sim r_{\phi_i}(\cdot \mid u)$ using the fixed grounded bridge
    if expert i is stochastic **then**
      Sample $\theta_{i\ell} \sim p_i(\cdot \mid x_i)$ and $v_{i\ell} \sim S_i(\cdot \mid \theta_{i\ell}, \xi_i(u))$, $\ell = 1, \ldots, L_i$
    else
      Evaluate the deterministic expert once from x_i and known conditions
    Package outputs, coordinates, optional parameters, and validity as $\mathcal{V}_i$
  Cache detached $(X, \hat{y}_0, (q_j), (\mathcal{V}_i))$ with target y
for each training minibatch **do**
  Sample $\eta_i \sim \text{Bernoulli}(1 - \rho)$ per available expert and case
  Compute $\hat{y}$ with Algorithm 1
  Update ψ using $\nabla_\psi \mathcal{L}(\hat{y}, y)$
return adapter checkpoint selected on a separately prepared validation split

Bistable posterior and experts. The 16-point grid is $\{0.7, 1.3\} \times \{0, 0.5\} \times \{0, 0.25\} \times \{0.25, 0.65\}$ with $\Delta t = 0.02$. The likelihood multiplies the transition densities in Eq. (10); the posterior is exact only for this declared discrete model, prior, and observation scheme. The local expert retains sampled coupling but linearizes drift around the last observation. The geometry grid has 9×9 locations over $[-2, 2]^2$. Regime labels prioritize nonzero tilt, then coupling, then high noise, leaving low noise otherwise; their frequencies follow the prior mass of the corresponding parameter subsets. The event is a sign change of x_1 relative to its last observed value at any forecast step. The direct expert event probability is its empirical crossing frequency; learned adapters use an auxiliary sigmoid head with loss weight 0.1. Plots show empirical 10–90% expert sample intervals.

Transient bistable setting. Default initial coordinates are centered at ± 1 with Gaussian standard deviation 0.15. An optional transient setting moves these centers to ± 1.7, uses 16 context observations and 96 forecast steps (1.92 time units), and draws 64 expert trajectories. This retains initial relaxation while extending the forecast beyond its locally near-linear portion. Around an uncoupled well, the deterministic relaxation timescale is approximately $1/(2a)$. More expert draws reduce Monte Carlo variation in the ensemble mean; forecast realizations retain process noise. The preset is an alternative task distribution and should be identified separately from the default tiered benchmarks.

Physical scope. Sphere experts overlap: the rotation rollout already incorporates the radius estimate. The thermal envelope is the only thermal expert that directly predicts temperature; solar and HVAC branches expose auxiliary fluxes. Their approximations provide different failure modes across regimes. The bistable stochastic expert removes posterior-sampled coupling while its local and geometry branches retain that information. These controlled mismatches support evaluation of selective expert use without requiring a uniformly reliable simulator.

C Adapter implementation and assumptions

Evidence preparation and optimization. Algorithm 2 separates fixed evidence generation from learning the neural adapter. Here ψ collects all trainable adapter parameters, including encoders, identity embeddings, writers, initializer, and decoder. The cached record $\mathcal{V}_i$ contains raw values, coordinates, optional sampled parameters, and validity masks; encoded tokens are recomputed during training. Held-out contexts follow the same preparation procedure using only their available observations and forecast conditions. Training dropout is disabled at inference, where η_i reflects expert availability or an explicit intervention.

Architectural realization. The initializer projects layer-normalized context features X and cross-attends with K learned slot queries. Expert tokenizers are two-layer GELU MLPs; optional parameter embeddings are broadcast over each draw's tokens. A learned vector identifies each expert. Pre-norm Transformer layers operate separately on draws before flattening their token axes. Draw permutations leave aggregation and readout unchanged unless explicit draw-index tags are introduced. Continuous temporal or spatial coordinates retain within-draw indexing. Fusion and decoder MLPs have hidden width $2d$; attention-block feed-forward networks use $4d$. Optional slot interaction uses zero dropout. Expert dropout samples one availability decision per expert and case, shared by writes and bypasses.

Context and output queries. For trajectory tasks, X contains observed and forecast positions, finite-difference velocities, relative time, and an observation flag. The output query q_j concatenates relative lead with the base forecast at that lead. Thermal queries contain lead, zone, and normalized base temperature; context features derive from the public tabular inputs and base predictions. In the core, X enters through $D_\psi(X)$, and q_j remains in the residual attention readout. Bypass architectures add the direct summary $c_0 = c_\psi(X)$ and concatenate raw queries at the head.

Bridge scope. A learned bridge requires supervision or a justified coupling between task context and simulator observations. With black-box simulation and posterior inference, downstream task loss alone supplies no bridge gradient. A grounded bridge can additionally use $s_i = \mathrm{Attention}(Q_i(e_i), B)$, where Q_i is a learned query projection. Such latent conditioning requires calibration against the simulator observation space and a consistent input representation, for example by freezing the initializer after calibration. Our environments use fixed observation maps. Posterior quality remains an assumption when replacing the declared inference models with learned SBI components Cranmer et al. (2020).

Route controls and pooled comparisons. The pooled expert comparator applies an MLP to concatenated per-expert masked means, a learned context summary, and the output query. Two restricted controls remove the direct context summary, or retain only expert means and task coordinates at the head. All predict an additive residual to the base output. Bypass variants share encoders with the workspace and retain their extended head when either route is disabled. We distinguish training under fixed route switches from zeroing a route after joint training: the latter intervention changes the input distribution seen by the fitted head. The original workspace's decoder interface is given separately in Eq. (5).

Fallback and representational scope. An optional exact fallback subtracts the same decoder applied to the initial workspace, with the same interaction stack and zero bypass evidence. This yields the base prediction when all evidence is absent. Default experiments instead permit a context-only correction. Nonlinear decoding supports curved forecasts; recovering a useful expert trajectory still depends on encoding, retrieval, and optimization. Slots provide shared computational channels, with no enforced physical factorization or calibrated uncertainty. Graph constraints and repeated simulator updates remain possible extensions beyond the present architecture.

D Evaluation and implementation protocol

Resource tiers. All counts below refer to independent cases. Small/medium/large presets use training sizes 192/384/1024, validation sizes 64/128/256, and test sizes 128/256/512. Widths are 32/64/128, slots 4/8/16, encoder depths 1/2/3, and heads 4/4/8. Context/horizon lengths are 32/8, 64/16, 128/32 for sphere; 4/6, 8/12, 16/24 for thermal; and 32/16, 64/32, 128/64 for bistable. Bistable expert draws are 4/16/32. Deterministic experts retain one draw. Changing tiers changes the prediction problem as well as capacity; comparisons of architecture alone must hold data settings fixed.

Support, optimization, and reporting. Only thermal generates a labeled support bank: 16/64/128 cases, each expanding to $3H$ rows. It fits ridge or conditions the tabular FM and is disjoint from adapter training, validation, and test cases. Series models receive each case's observed trajectory and skip the global support bank. External adapters use AdamW (learning rate $3 \cdot 10^{-4}$, weight decay 10^{-4}), gradient clipping at norm 1, and expert dropout 0.15. Maximum epochs are 30/60/100 with patience 8/12/20; validation MSE selects checkpoints, including epoch zero. Chronos LoRA uses

a separate native training loop. Checkpoints, versions, ensemble counts, and conditioning budgets must accompany reported scores.

RMSE, MAE, final MAE, and bias are reported overall and by regime, with leadwise MAE; trajectory tasks additionally report average/final Euclidean distance and sphere reports radial MAE. Bistable reports event Brier scores where available. Repeated evaluations bootstrap whole seeds, with paired differences formed within seed; timestep rows and ensemble draws are not independent replicates. Intervals are pointwise and do not correct repeated model selection on test data. Reconstruction standardizes each target channel using training evidence, predicts aligned draw means or individually tagged draws, and trains fresh no-expert, pooled, and original-workspace controls. Bypass variants are not currently part of that fixed reconstruction suite. No empirical ranking is asserted here.

Mixed and independent-subtask training. Mixed runs draw each split from the environment's regime mixture and fit one model per comparison case and seed. Independent runs condition generation on each regime and train separate adapters, native LoRA models, and thermal support contexts. All configured case counts apply per subtask, increasing total data and compute relative to the same mixed-run configuration. Validation selection is performed within each task. Bistable experts retain their full-grid inference prior even when the data generator is conditioned on a regime; they receive only the observed trajectory. This preserves the expert interface across protocols.

Mixed-run Overall metrics pool held-out cases. Independent-run Overall metrics are arithmetic means of subtask metrics, including RMSE. Repeated summaries first form this mean within each seed, then estimate variability across seeds. Compact tables report means, optionally accompanied by sample standard deviations; fuller reports retain bootstrap confidence intervals. Independently trained slot indices are analyzed within subtask, since their meanings can differ across models. Reconstruction likewise fits task-specific training normalizations before combining standardized metrics. Plot examples use the first held-out case per regime, selected independently of prediction error.