TERM: ENRICHING REPRESENTATIONS IN PDE FOUNDATION MODELS WITH TERM-WISE PHYSICAL RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Partial Differential Equation (PDE) foundation models aim to learn transferable solution priors across heterogeneous physical systems, yet existing methods typically represent governing equations as static symbolic structures or combine equations and trajectories only at a global level. Such representations capture which physical terms constitute a PDE, but rarely how each term acts in the current state. Although physical terms have stable mathematical identities, their strength, spatial distribution, and dynamical roles vary along physical trajectories. Motivated by this observation, we propose the Term-wise Enriched Representation Model (TERM), which organizes PDE information at the level of individual physical terms when the governing equation is available. TERM first normalizes the equation into a structured term graph, then computes state-dependent responses for each physical term and explicitly binds them to their symbolic identities. A structure-aware Transformer jointly models these response-enriched term representations and compact trajectory information, while a continuous query decoder enables prediction across equations, resolutions, and continuous coordinates. During autoregressive rollout, the equation structure remains fixed while term responses are recomputed from the evolving predicted state. Experiments across multi-PDE pretraining, equation-specified zero-shot transfer, compositional generalization, cross-resolution prediction, and long-horizon rollout show consistent improvements over the evaluated PDE foundation-model baselines and representation controls. The results suggest that state-dependent, term-level physical representations provide a useful interface for scalable and compositional PDE foundation modeling.

1 INTRODUCTION

Partial differential equations (PDEs) provide central mathematical models for describing physical processes across science and engineering, including fluid dynamics, heat transfer, material processes, and multiphysics phenomena (Brunton & Kutz, 2024; Huang et al., 2025). High-fidelity numerical solvers remain essential for these problems, but repeated PDE solves can become computationally expensive in applications requiring repeated evaluations (Dai et al., 2026; Nooraiepour et al., 2026). Neural operators address this cost by learning mappings between function spaces and approximating solution operators across families of parameterized PDEs (Lu et al., 2021; Li et al., 2021; Kovachki et al., 2023; Subedi & Tewari, 2026). Their ability to share representations across discretizations, parameters, and problem instances has motivated a broader shift from task-specific surrogates toward PDE foundation models that share knowledge across multiple physical systems. Such models explore equation-aware representations (Ye et al., 2024; 2025), multimodal equation-trajectory learning (Liu et al., 2024a), large-scale multi-PDE pretraining (Hao et al., 2024; Chen et al., 2025; Rautela et al., 2025; Sun et al., 2026a), and compositional transfer to unseen physical systems (Sun et al., 2026b). A central question in this direction is how physical knowledge should be represented so that it can be effectively reused across different PDE systems.

Recent work increasingly seeks unified PDE models through large-scale pretraining and shared representations across heterogeneous physical systems(Hao et al., 2024; Chen et al., 2025; Rautela et al., 2025; Sun et al., 2026a), while research explores incorporating governing-equation informa-

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.

tion and symbolic physical structure into the learning process (Ye et al., 2024; 2025; Liu et al., 2024a; Sun et al., 2026b). These developments suggest that transferable PDE modeling depends not only on data scale and model capacity, but also on how physical structure is represented and connected to the evolving system state.

A remaining challenge concerns the granularity at which these two sources of information are connected. A *governing equation* specifies which physical mechanisms are present and how they are composed, while a *physical trajectory* reveals the state in which these mechanisms currently operate. The symbolic identity of a physical term is persistent, but its actual contribution is state-dependent. For example, convection and diffusion remain fixed components of the viscous Burgers equation, while their magnitudes and spatial patterns change as gradients develop and dissipate. Existing equation or trajectory representations can encode these two aspects, yet the correspondence between an individual symbolic term and the response it induces on the current state is generally not treated as an explicit representation unit. This distinction becomes particularly relevant for cross-PDE learning, where recurring physical terms such as convection, diffusion, reaction, dispersion, and pressure gradients provide natural units for sharing knowledge across different equation systems.

We propose the Term-wise Enriched Representation Model (TERM), which represents each physical term through both its persistent symbolic identity and its current physical response. Given a known governing equation, TERM first organizes it into a structured PDE term graph and evaluates each physical term on recent trajectories to obtain its current term-wise physical response. The symbolic identity and current response of each term are then explicitly bound into a response-enriched physical-term representation. These enriched terms are jointly modeled with compact trajectory representation by a structure-aware Foundation Transformer that retains relations from the original PDE graph. A lightweight continuous-query decoder then predicts physical variables at arbitrary coordinates and spatial resolutions. During autoregressive rollout, the symbolic graph remains fixed while term responses and enriched representations are updated from the evolving predicted states. Therefore, TERM provides a shared representation in which equation structure remains stable and the state-dependent activity of each physical mechanism evolves with the trajectory.

The main contributions are as follows:

- We introduce a representation unit that explicitly associates each symbolic PDE term with the physical response it produces on the current trajectory, providing a finer-grained interface for sharing physical information across PDE systems.
- We develop a unified architecture that combines a canonical PDE term graph, shared response encoding, term-local response enrichment, and a structure-aware Foundation Transformer. A continuous query decoder further allows the same representation to support different output resolutions and coordinate queries.
- We evaluate TERM across unseen equations and physical parameters, including novel compositions of previously observed physical terms, as well as resolution changes and long autoregressive rollouts. Comparison with the information-matched Unbound Response control isolates the effect of explicitly preserving term-response correspondence while keeping symbolic identities and response information available to the model.

2 Related Work

General-purpose neural operators and cross-physics pretraining Recent operator-learning research has expanded from individual PDE families toward reusable models that accommodate broader variation in domains, data, and dynamics (Lu et al., 2021; Li et al., 2021). GINO combines graph-based transforms with a Fourier neural operator to learn on varying three-dimensional geometries and discretizations (Li et al., 2023). Unsupervised operator pretraining has also used physics-inspired reconstruction tasks and in-context examples to improve data efficiency when simulated labels are limited (Chen et al., 2024). At the multi-system level (Hao et al., 2024; Chen et al., 2025; Sun et al., 2026b), MPP trains an autoregressive Transformer across heterogeneous spatiotemporal systems with shared field embeddings and normalization (McCabe et al., 2024), while Poseidon couples a multiscale operator Transformer with semigroup-based training to transfer from fluid-dynamics pretraining to diverse downstream PDEs (Herde et al., 2024). These approaches establish that trajectory data from multiple systems can support transferable solution priors. Their

main representation is learned from physical fields, with geometry handling, normalization, and pretraining objectives providing the interface across tasks. TERM adds a term-level interface to this setting. Recurring mechanisms are represented by persistent symbolic identities together with responses evaluated on the current trajectory, so the shared model can track how the same mechanism changes across systems and rollout steps.

Equation-aware and physics-structured learning Governing equations have been incorporated into learned PDE models as supervision, discovered structure, or conditioning information (Ye et al., 2024; Liu et al., 2024a). PDE-Net 2.0 learns differential operators and a symbolic network from observed dynamics, yielding an interpretable recovered equation (Long et al., 2019). PINO imposes PDE constraints on a neural operator, including constraints evaluated at resolutions finer than the available training data (Li et al., 2024). Symbolic expressions can also be treated as a model modality. PROSE jointly predicts numerical operators and governing expressions through multimodal fusion (Liu et al., 2024b), and Unisolver conditions a Transformer on equation symbols, coefficients, boundary conditions, and spatially varying PDE components (Zhou et al., 2025). These methods show that equation information can guide learning beyond trajectory-only prediction. TERM assumes that the governing equation is supplied and uses it to define the representation itself. Each signed additive term is evaluated on the visible state, and the resulting field is bound to that term's symbolic embedding before shared processing. This construction exposes the instantaneous activity of individual mechanisms, including contributions obscured by cancellation in the aggregate right-hand side, and refreshes that activity as predicted states evolve during autoregressive rollout.

3 METHOD

Figure 1 summarizes the information flow in TERM. The model separates persistent equation structure from state-dependent term activity, binds each response to its corresponding symbolic term, and recomputes the dynamic branch as predicted states enter the history during rollout.

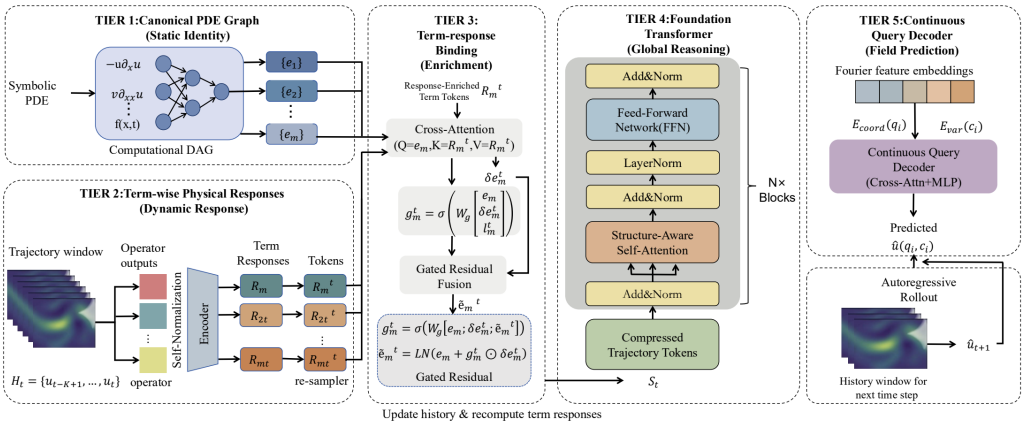

Figure 1: Overview of TERM. The canonical PDE graph provides persistent symbolic term identities, while the observed trajectory is evaluated term by term to obtain state-dependent physical responses. Each response enriches its corresponding term token before the foundation Transformer combines the enriched terms with compressed trajectory tokens. A continuous query decoder predicts the future field, which updates the history and triggers response recomputation during autoregressive rollout.

3.1 PROBLEM SETTING AND CANONICAL TERM GRAPH

We consider evolution PDEs whose governing equations, coefficients, domain, and boundary conditions are available at both training and inference time. A PDE instance is $\mathcal{P} = (\mathcal{E}, \boldsymbol{\lambda}, \mathcal{G}, \mathcal{B})$, where $\mathcal{E}$ is the equation system, $\boldsymbol{\lambda}$ its physical parameters, and $\mathcal{G}$ and $\mathcal{B}$ specify the domain and boundary conditions. After introducing auxiliary variables when needed, let $\mathcal{M}_c$ index the additive physical

terms in the equation for state component c. We write that equation as

$$\partial_t u^{(c)} = \sum_{m \in \mathcal{M}_c} \mathcal{O}_m[u; \boldsymbol{\lambda}]. \tag{1}$$

Each $\mathcal{O}_m$ includes its sign and coefficient. A known forcing field is treated as an additive term. The total number of terms in $\mathcal{P}$ is $M_{\mathcal{P}} = \sum_c |\mathcal{M}_c|$. Given K observed states $\mathcal{H}_t = (u_{t-K+1}, \ldots, u_t)$, TERM predicts $\{\hat{u}^{(c_i)}(q_i)\}_{i=1}^N$ at queries $\mathcal{Q}_{t+1} = \{(q_i, c_i)\}_{i=1}^N$, where $q_i = (x_i, t+1)$. We use the frame index t for notation. Its physical timestamp is supplied to the coordinate encoder when frame intervals vary.

A computational graph $G_{\mathcal{P}} = (V_{\mathcal{P}}, E_{\mathcal{P}})$ encodes $\mathcal{E}$, with term nodes corresponding to additive components of the supplied equation. Each term connects to its operator subtree, coefficient, source variables, and target equation; boundary and geometry descriptors provide additional context. A Graph Transformer (Dwivedi & Bresson, 2021) produces a persistent embedding $e_m = T_{\text{graph}}(G_{\mathcal{P}})_m$ for each term. This decomposition sets the granularity of TERM's representation: two mechanisms can depend on the same state variable yet have distinct identities and, subsequently, distinct physical responses. The graph embeddings are reused throughout a rollout. Appendix A.1 defines the term partition and graph construction.

3.2 State-dependent physical responses

For each computable term, we evaluate the same operator on the most recent $K_r \leq K$ observed states:

$$\rho^t_{m,k}(x) = \mathcal{O}_m[u_{t-K_r+k}; \boldsymbol{\lambda}](x), \qquad k = 1, \ldots, K_r. \tag{2}$$

Each $\rho^t_{m,k}$ is a signed spatial field showing where term m acts on an observed state. Keeping these fields separate preserves, for example, large convection and diffusion responses that nearly cancel when summed. We evaluate each term on the observation grid, with that grid's physical spacing and boundary conditions, before forming response patches. Term responses are computed from the observed fields and the supplied governing equation using filtered spectral differentiation on periodic grids and boundary-aware finite differences on nonperiodic structured grids. No hidden simulator states or solver-internal term outputs are required. Differentiating a resampled state would generally give a different field. If a term requires an unobserved variable, such as pressure when only velocity is available, we set its validity mask $m^{\text{resp}}_m = 0$ and treat the response as missing.

Let Ω be the physical spatial domain, excluding padded cells in irregular geometries. The normalization integral below uses its measure $|\Omega|$, and each response patch retains its physical coordinates for the shared resampler. The magnitudes of convection, diffusion, and forcing responses may differ by orders of magnitude across equations and states. We normalize each term over its visible window:

$$a^t_m = \left(\frac{1}{K_r |\Omega|} \sum_{k=1}^{K_r} \int_\Omega |\rho^t_{m,k}(x)|^2 \, \mathrm{d}x \right)^{1/2} + \epsilon, \quad \bar{\rho}^t_{m,k} = \frac{\rho^t_{m,k}}{a^t_m}, \quad \ell^t_m = \log a^t_m. \tag{3}$$

Normalization separates the signed spatial pattern $\bar{\rho}^t_{m,k}$ from its scale ℓ^t_m over the response window. The constant $\epsilon > 0$ keeps both quantities defined for an inactive term, and the scale requires no operator-specific dataset statistics because it comes from the observed window. A shared response encoder maps patches from the K_r normalized fields, together with their physical coordinates, relative times, and ℓ^t_m, to tokens. Learned queries compress each term's tokens into R^t_m of fixed size R, so term-local cross-attention has a fixed token budget per term regardless of grid resolution.

In parallel, a state encoder maps history values, coordinates, timestamps, and variable identities to N_s summary tokens S_t. The response tokens feed term-local enrichment, while S_t provides state context to the shared backbone alongside the enriched terms.

3.3 TERM-LOCAL ENRICHMENT AND SHARED REPRESENTATION

TERM first binds each term embedding to its own response tokens. After projection to a common width, term-local cross-attention and a gated residual update give

$$\begin{aligned} \delta e_m^t &= \mathrm{CrossAttn}(e_m, R_m^t, R_m^t), \\ g_m^t &= \sigma\big(W_g[e_m; \delta e_m^t; \ell_m^t]\big), \\ \tilde{e}_m^t &= \mathrm{LN}\big(e_m + g_m^t \odot \delta e_m^t\big). \end{aligned} \tag{4}$$

For an unavailable response, learned missing-response and missing-scale embeddings replace R_m^t and ℓ_m^t; the mask distinguishes it from a measured zero field. The gate modulates the state-dependent update. Since cross-attention is indexed by m, diffusion updates the diffusion token even when its response cancels convection in the summed right-hand side. Term attribution remains available to the shared backbone.

The foundation Transformer consumes $\{\tilde{e}_m^t\}_{m=1}^{M_\mathcal{P}}$, summaries S_t, and boundary and geometry tokens, producing Z_t. For term pairs, its attention logit is $A_{ij} = Q_i^\top K_j/\sqrt{d} + b_{\mathrm{rel}}(i,j)$. The bias uses typed connections and shortest-path distance in $G_\mathcal{P}$, distinguishing terms coupled through a shared variable from unrelated terms. State tokens attend to all terms to combine operator activity with trajectory context.

The two inputs to enrichment carry different forms of variation. Changing a PDE coefficient or operator changes the persistent graph embedding, whereas moving to another state of the same PDE changes its response field. Their interaction is learned only after each response has been paired with the correct term. With R resampled tokens per term and N_s state summaries, the backbone receives $M_\mathcal{P} + N_s$ main tokens, independent of the number of points on the observation grid. This keeps the shared backbone usable across datasets with different spatial resolutions.

3.4 CONTINUOUS DECODING AND AUTOREGRESSIVE TRAINING

For variable c_i requested at $q_i = (x_i, t+1)$, the shared decoder forms $e_i^q = E_{\mathrm{coord}}(q_i) + E_{\mathrm{var}}(c_i)$. It reads Z_t through cross-attention, $h_i = D_{\mathrm{cross}}(e_i^q, Z_t)$, and applies a linear head, $\hat{u}^{(c_i)}(q_i) = W_{\mathrm{out}} h_i + b_{\mathrm{out}}$. Fourier features locate the query and E_{var} identifies its field, so the same Z_t can be decoded at grid or off-grid coordinates with shared parameters. During rollout, the query set covers the evaluator grid and all variables required by the response operators. The decoded field replaces the oldest frame in $\mathcal{H}_t$. Responses, enriched terms, and Z_t are then recomputed, while the symbolic graph remains fixed.

The objective supervises the entire closed-loop trajectory. At each training step the next predicted field enters the history used by the following step. The rollout and response-preservation losses combine as

$$\begin{aligned} \mathcal{L}_{\mathrm{roll}} &= \frac{1}{T}\sum_{j=1}^{T} \frac{\|\hat{u}_{t+j} - u_{t+j}\|_2}{\|u_{t+j}\|_2 + \epsilon}, \\ \mathcal{L}_{\mathrm{resp}} &= \frac{1}{\max(1, |\mathcal{M}_{\mathrm{resp}}^t|)} \sum_{m \in \mathcal{M}_{\mathrm{resp}}^t} \|D_{\mathrm{resp}}(\tilde{e}_m^t) - y_m^t\|_2^2, \\ \mathcal{L} &= \mathcal{L}_{\mathrm{roll}} + \lambda_{\mathrm{resp}} \mathcal{L}_{\mathrm{resp}}, \end{aligned} \tag{5}$$

where $\mathcal{M}_{\mathrm{resp}}^t = \{m : m_m^{\mathrm{resp}} = 1\}$ contains the terms with computable responses. The response head predicts $y_m^t = \mathrm{Pool}(\{\bar{\rho}_{m,k}^t\}_{k=1}^{K_r})$, a pooled representation of each valid term's normalized response window.

4 THEORETICAL ANALYSIS

This section relates term responses to instantaneous PDE dynamics and derives conditional bounds on numerical response error and autoregressive feedback under the stated regularity and stability assumptions. Appendix A.2 gives the assumptions and proofs.

Response identity and attribution For each component c, the exact responses $\rho^t_{m,K_r} = \mathcal{O}_m[u_t; \boldsymbol{\lambda}]$ satisfy $\partial_t u_t^{(c)} = \sum_{m \in \mathcal{M}_c} \rho^t_{m,K_r}$ by the declared PDE, including when some terms cannot be evaluated from partial observations. These fields decompose the instantaneous right-hand side into signed contributions. On the product space of response fields, the aggregation map $A_c((r_m)_{m \in \mathcal{M}_c}) = \sum_m r_m$ is non-injective when $|\mathcal{M}_c| \geq 2$. Adding any nonzero field to one entry and subtracting it from another leaves A_c unchanged. Thus a summed right-hand side cannot in general identify separate term actions; the effect on prediction is an empirical question.

Numerical response and rollout Let Π_h sample a field on a grid with norm $\|\cdot\|_h$. If a discrete term evaluator has consistency error $\varepsilon_{m,\mathrm{disc}}(h)$ and perturbation factor $\kappa_{m,h}$, then an observation error of at most η gives

$$\left\|\hat{\rho}^t_{m,h} - \Pi_h \rho^t_{m,K_r}\right\|_h \leq \varepsilon_{m,\mathrm{disc}}(h) + \kappa_{m,h}\eta. \tag{6}$$

For autoregressive prediction, let $e_n = \|\hat{u}_{t+n} - u_{t+n}\|$ measure state error on the evaluator grid. Suppose the complete history-to-next-state predictor has one-step error at most $\varepsilon_{\mathrm{step}}$ on exact histories and is locally L_{TERM}-Lipschitz in the maximum history norm. With an exact initial history,

$$e_{n+1} \leq \varepsilon_{\mathrm{step}} + L_{\mathrm{TERM}} \max_{0 \leq j \leq n} e_j, \qquad e_n \leq \varepsilon_{\mathrm{step}} \sum_{j=0}^{n-1} L^j_{\mathrm{TERM}}. \tag{7}$$

The bound is uniform in the rollout horizon when the additional contraction assumption $L_{\mathrm{TERM}} < 1$ holds; the architecture itself does not establish this condition.

5 Experiments

5.1 Experimental Setup

Benchmarks Following the multi-physics suite of Origo (Sun et al., 2026b), we jointly pretrain on ten tasks from four public sources. FNO-ν contributes Navier–Stokes (NS) flows at $\nu \in \{10^{-5}, 10^{-4}, 10^{-3}\}$ (Li et al., 2021); PDEBench contributes Burgers (Bur.), diffusion–reaction (DR), shallow-water equations (SWE), and compressible Navier–Stokes (CNS) (Takamoto et al., 2022); PDEArena contributes NS and conditioned NS (NS-c) (Gupta & Brandstetter, 2022); and CFDBench contributes nonperiodic computational-fluid-dynamics data (CFD) (Luo et al., 2023). For zero-shot evaluation, we use the held-out 1D KdV and 2D Allen–Cahn target equations from the unseen-equation evaluation of Origo (Sun et al., 2026b). TERM is evaluated in an equation-specified setting, where the target governing equation is supplied at inference as part of the model input. Further details are shown in Appendix A.3.1 and A.3.2.

Baselines Task-specific experts comprise FNO (Li et al., 2021), U-Net (Ronneberger et al., 2015), factorized FNO (FFNO) (Tran et al., 2023), Galerkin Transformer (GK-T) (Cao, 2021), and GNOT (Hao et al., 2023). Unified pretraining baselines comprise DPOT (Hao et al., 2024), MoE-POT (Wang et al., 2025), OmniArch (Chen et al., 2025), and Origo (Sun et al., 2026b). The former are trained separately on each task; the latter are evaluated after joint pretraining and after 200-epoch task adaptation. We select unified baselines that admit evaluation under the heterogeneous multi-PDE benchmark considered here. Equation-aware models designed for different PDE domains or training protocols are discussed in Section 2 but are not directly compared when an equivalent implementation under the same multi-PDE setting is unavailable.

Evaluation Metrics Prediction error is measured by relative L_2 error (rel-L_2), rel-$L_2(\hat{u}, u) = \|\hat{u} - u\|_2 / \|u\|_2$, on the designated test split. For autoregressive experiments, the same error is reported at each rollout horizon.

Evaluation Scope The evaluation combines joint pretraining and task adaptation across the ten benchmarks, representation ablations, zero-shot transfer to held-out equations and unseen term compositions, and long-horizon rollouts. Appendix A.4 extends these experiments to dataset and parameter shifts, different dynamical regimes, changes in output resolution and query locations, and response-estimation robustness. Appendix A.3.3 and A.3.4 give the common training and evaluation protocol.

5.2 Main Results

Table 1 shows that under joint pretraining, TERM-M obtains the lowest error among unified models on eight of ten tasks and improves over Origo-M on every task. Averaging the per-task relative reductions gives 10.5%. TERM-M also outperforms the best task-specific expert on nine tasks, despite using one shared model across all PDEs. Its 105 M parameters place it below DPOT-M, MoE-POT-M, and OmniArch-M in model size.

After 200 epochs of task adaptation, TERM-200 achieves the lowest error on all ten benchmarks in Table 1. The gains remain substantial on dynamics with different numerical structure: NS decreases from 0.0798 to 0.0218, while CFD decreases from 0.0109 to 0.0039. These results show that the jointly learned representation remains adaptable to individual data distributions.

Table 1: Rel-L_2 across ten PDE benchmarks. **Bold** denotes the lowest error within each regime.

Category	Method	Params.	FNO-ν			PDEBench				PDEArena		CFD
			10^{-5}	10^{-4}	10^{-3}	Bur.	DR	SWE	CNS	NS	NS-c	
Task-specific experts	FNO	0.5 M	0.1827	0.0976	0.0151	0.0205	0.1416	**0.0052**	0.1154	0.1063	**0.3740**	0.0088
	UNet	25 M	0.2326	0.1405	0.0290	0.3743	0.1143	0.0612	0.3928	0.1202	0.3942	0.2454
	FFNO	1.3 M	**0.1421**	0.0589	0.0116	0.0141	0.0669	0.0136	**0.0248**	**0.0985**	0.7045	**0.0084**
	GK-T	1.6 M	0.1581	0.0931	**0.0115**	0.0147	0.0421	0.0081	0.0400	0.1115	0.4947	0.0123
	GNOT	1.8 M	0.1844	**0.0518**	0.0146	**0.0123**	**0.0365**	0.0079	0.0379	0.2015	0.3792	0.0103
Unified pretraining	DPOT-M	122 M	0.0482	0.0334	0.0056	0.0123	0.0342	0.0034	0.0136	0.0950	0.3232	0.0088
	MoE-POT-M	288 M	0.0420	0.0410	0.0067	0.0111	0.0350	0.0035	0.0107	0.1152	0.4102	**0.0060**
	OmniArch-M	144 M	0.2415	0.1691	0.0530	0.0104	0.0186	**0.0019**	0.0230	0.2027	0.6430	0.6434
	Origo-M	97 M	0.0342	0.0257	0.0048	0.0099	0.0025	0.0033	0.0089	0.0871	0.3124	0.0125
	TERM-M	105 M	**0.0306**	**0.0228**	**0.0045**	**0.0086**	**0.0022**	0.0029	**0.0080**	**0.0798**	**0.2862**	0.0109
Fine-tuned for 200 epochs	DPOT-200	122 M	0.0300	0.0168	0.0050	0.0045	0.0166	0.0026	0.0144	0.0384	0.2257	0.0053
	MoE-POT-200	288 M	0.0410	0.0188	0.0046	0.0041	0.0163	0.0023	0.0087	0.0721	0.2591	0.0046
	OmniArch-200	144 M	0.0963	0.0597	0.0176	0.0083	0.0120	0.0018	0.0179	0.1120	0.4445	0.0047
	Origo-200	97 M	0.0268	0.0155	0.0044	0.0037	0.0142	0.0015	0.0085	0.0247	0.2155	0.0044
	TERM-200	105 M	**0.0239**	**0.0138**	**0.0040**	**0.0033**	**0.0020**	**0.0014**	**0.0076**	**0.0218**	**0.1915**	**0.0039**

5.3 Ablation Studies

We compare the complete *TERM* model with the component controls *Trajectory-only*, *Static Graph*, *Response-only*, and *Global Response*, and with the information-matched controls *RHS-Matched*, *Identity-Free*, *Shuffled Binding*, and *Unbound Response*. Appendix A.3.5 gives their exact constructions. As shown in Table 2, the strongest information-matched control is Unbound Response, yet explicit pairing reduces KdV error from 0.0114 to 0.0087 and AC error from 0.0923 to 0.0836. Shuffled Binding performs worse than Unbound Response on all four tasks, showing that correspondence quality matters even when the available inputs are unchanged. The component ablations give the same ordering: equation structure improves over Trajectory-only, current responses improve over Static Graph, and TERM improves over Global Response on every task. The final gain is largest on KdV, where preserving the term–response association reduces error by 26.9% relative to Global Response.

Table 2: Representation ablations measured by rel-L_2. Panel (a) matches the available information and token budget; panel (b) removes major components. **Bold** denotes the lowest error in each panel.

(a) Binding and information controls

Variant	Bur.	NS	KdV	AC
Static Graph	0.0101	0.0876	0.0142	0.1115
RHS-Matched	0.0095	0.0847	0.0127	0.1010
Identity-Free	0.0093	0.0838	0.0122	0.0975
Shuffled Binding	0.0094	0.0842	0.0124	0.0990
Unbound Response	0.0091	0.0821	0.0114	0.0923
TERM	**0.0086**	**0.0798**	**0.0087**	**0.0836**

(b) Component ablations

Variant	Bur.	NS	CNS	KdV	AC
Trajectory-only	0.0118	0.0947	0.0104	0.0169	0.1285
Static Graph	0.0101	0.0876	0.0093	0.0142	0.1115
Response-only	0.0097	0.0859	0.0090	0.0135	0.1058
Global Response	0.0093	0.0834	0.0087	0.0119	0.0964
TERM	**0.0086**	**0.0798**	**0.0080**	**0.0087**	**0.0836**

5.4 ZERO-SHOT GENERALIZATION

We evaluate zero-shot generalization to held-out equations and to novel compositions of previously observed physical terms.

Unseen equation generalization Under the unseen-equation protocol introduced in Section 5.1, all parameters remain frozen during evaluation. The target equation is supplied at inference to construct TERM's graph and responses, making this an equation-specified zero-shot setting.

Table 3: Zero-shot rel-L_2 on unseen KdV and AC equations. **Bold** denotes the lowest error on each unseen equation.

Method	Params.	KdV	AC
DPOT-M	122M	0.0385	0.2713
MoE-POT-M	288M	0.0237	0.1521
OmniArch-M	144M	0.0342	0.2496
Origo-M	97M	0.0103	0.0952
TERM-M	105M	**0.0087**	**0.0836**

Table 3 shows that TERM-M reaches 0.0087 on KdV and 0.0836 on AC, reducing error relative to Origo-M by 15.5% and 12.2%, respectively. On AC, the result measures reuse of known term families in an unseen equation. On KdV, the complete third-order dispersion term u_{xxx} is absent from pretraining. TERM constructs its operator tree compositionally from derivative primitives represented by the canonical graph vocabulary, while the response evaluator computes its state-dependent action from the supplied equation and observed state. The experiment therefore evaluates equation-specified transfer to an unseen higher-order operator composition; it does not require TERM to infer the governing equation from trajectories.

Strict physical-term compositional generalization Appendix A.3.6 gives the complete benchmark construction and signature audit. Let A, D, R, and S denote advection, diffusion, reaction, and dispersion. Training contains all six pairwise combinations, whereas testing contains the four unseen 1D triple combinations and a 2D $A + D + R$ equation trained only from its three pairs. Every test term and pairwise subcomposition has therefore been observed, while the complete target composition has not. Coefficient ranges and the initial-condition family are shared across the split.

Table 4: Zero-shot rel-L_2 under strictly held-out physical-term compositions. **Bold** denotes the lowest error for each held-out composition and the 1D average.

Variant	1D held-out compositions					2D held-out
	$A+D+R$	$A+D+S$	$A+R+S$	$D+R+S$	Average	$A+D+R$
Trajectory-only	0.0418	0.0364	0.0449	0.0396	0.0407	0.0630
Static Graph	0.0356	0.0315	0.0387	0.0342	0.0350	0.0550
RHS-Matched	0.0319	0.0286	0.0344	0.0308	0.0314	0.0490
Identity-Free	0.0298	0.0273	0.0328	0.0294	0.0298	0.0470
Shuffled Binding	0.0324	0.0291	0.0351	0.0312	0.0320	0.0500
Unbound Response	0.0271	0.0249	0.0296	0.0267	0.0271	0.0430
TERM	**0.0219**	**0.0197**	**0.0238**	**0.0215**	**0.0217**	**0.0360**

Table 4 shows that TERM achieves the lowest error for every held-out composition. Its 1D average of 0.0217 improves on the information-matched Unbound Response value of 0.0271 by 19.9%; the corresponding reduction on 2D $A + D + R$ is 16.3%. Shuffled Binding and RHS-Matched reach 0.0320 and 0.0314 on the 1D average, respectively. The comparison with Unbound Response isolates the effect of explicitly preserving each response's association with its symbolic term during recombination.

5.5 Numerical Robustness of Physical Responses

Because response evaluation differentiates observed states, we test input noise and coarsened response grids on Burgers, NS, and zero-shot KdV; Appendix A.4.5 gives the full results. At 5% noise, TERM yields a macro-average rel-L_2 of 0.0419, below Static Graph (0.0487) and unregularized TERM-Direct (0.0597), and remains best at response resolutions 32, 64, and 128. This pattern is consistent with the response-error analysis in Section 4 and motivates controlling noise amplification in the discrete evaluator.

5.6 Long-Horizon Rollout Prediction

Starting from ten observed frames, each model performs a free autoregressive rollout for 10, 25, 50 or 100 steps. Cumulative rel-L_2 is the mean error over the predicted horizon. Static Graph removes responses, Frozen Response improves over Static Graph but remains above TERM at longer horizons, while TERM recomputes term responses from the evolving predicted state at every rollout step. TERM-Direct is slightly better at 10 steps, whereas TERM achieves lower cumulative error from 25 to 100 steps on both equations. These results are consistent with the benefit of updating term responses during rollout and with the stabilized evaluator becoming increasingly useful as prediction errors accumulate.

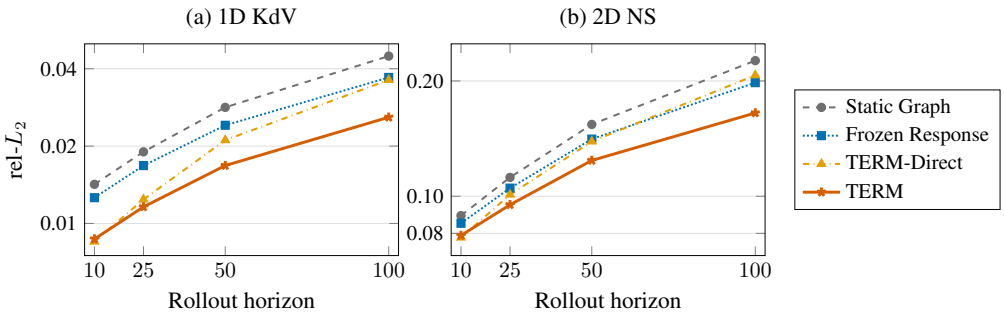

Figure 2: Cumulative rel-L_2 across long-horizon rollout horizons. Both y-axes are logarithmic.

Figure 2 shows that, at 100 steps, TERM reduces cumulative error from 0.226 to 0.165 on NS and from 0.0448 to 0.0259 on zero-shot KdV compared with Static Graph. Frozen Response improves on Static Graph but remains above TERM at long horizons, indicating that initial response information becomes stale as the state evolves. TERM-Direct is slightly better at 10 steps, then accumulates error more quickly. TERM is best from 25 through 100 steps on both equations. Updating term responses is therefore useful over long rollouts when the discrete evaluator controls feedback from prediction error.

6 Conclusion

TERM connects persistent PDE structure with evolving dynamics at the level of individual physical terms. A canonical graph encodes symbolic term identities and relations, while term-wise evaluation provides their state-dependent responses. Binding these views before shared Transformer processing preserves response attribution and allows the representation to evolve during autoregressive prediction. Across the evaluated settings, TERM consistently improves over static, globally aggregated, and incorrectly bound representations, while information-matched ablations show that explicit term-response correspondence contributes beyond simply providing more equation or trajectory information. Our analysis further characterizes the information lost through response aggregation and the propagation of numerical response errors during enrichment and rollout. TERM assumes that the governing equation is supplied and evaluable, and its accuracy depends on response estimation and observation quality. Future work may extend the framework to uncertain equations, partial observations, noise-robust response evaluation, and higher-dimensional geometries.

AI USE STATEMENT

Generative AI tools assisted with drafting and revising portions of the manuscript, providing feedback on research methodology, checking notation and LaTeX, and preparing plotting and table code. They also assisted in summarizing author-provided experimental results and critiquing explanations of the method and theoretical analysis. The authors reviewed all AI-assisted material, checked the cited sources, mathematical statements, implementation details, figures, and reported values, and took responsibility for the final content. Generative AI was not used as a substitute for numerical simulation or model evaluation.

ETHICS STATEMENT

This study uses numerical simulations and public PDE benchmarks and does not involve human participants or personal or sensitive data. TERM is a surrogate modeling method whose predictions can be unreliable under equation misspecification, partial observation, or distribution shift; outputs should therefore be validated against appropriate numerical solvers before use in safety-critical scientific or engineering workflows. The authors have reviewed and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Section 3 specifies the representation, architecture, response construction, training objective, and autoregressive procedure. Section 4 and Appendix A.2 state the assumptions and provide the complete analysis. Appendix A.3 documents dataset sources and splits, preprocessing, the TERM-M architecture, response evaluators, optimization, ablation constructions, and the strict-composition benchmark. Appendix A.4 gives the supplementary evaluation protocols and results. Together with the cited public dataset implementations, these sections document the reported protocol and comparison settings.

REFERENCES

Steven L Brunton and J Nathan Kutz. Promising directions of machine learning for partial differential equations. *Nature Computational Science*, 4(7):483–494, 2024.

Shuhao Cao. Choose a transformer: Fourier or galerkin. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24924–24940, 2021. URL https://papers.nips.cc/paper/2021/hash/d0921d442ee91b896ad95059d13df618-Abstract.html.

Tianyu Chen, Haoyi Zhou, Ying Li, Hao Wang, Chonghan Gao, Rongye Shi, Shanghang Zhang, and Jianxin Li. OmniArch: Building foundation model for scientific computing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9860–9887. PMLR, 2025.

Wuyang Chen, Jialin Song, Pu Ren, Shashank Subramanian, Dmitriy Morozov, and Michael W. Mahoney. Data-efficient operator learning via unsupervised pretraining and in-context learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0201. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0bac492172db3311c7e116098cfcf521-Abstract-Conference.html.

Yilong Dai, Shengyu Chen, Ziyi Wang, Xiaowei Jia, Yiqun Xie, Vipin Kumar, and Runlong Yu. Learning pde solvers with physics and data: A unifying view of physics-informed neural networks and neural operators. *arXiv preprint arXiv:2601.14517*, 2026.

Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.

Jayesh K. Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized PDE modeling. *arXiv preprint arXiv:2209.15616*, 2022. URL https://arxiv.org/abs/2209.15616.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: A general neural operator transformer for operator learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12556–12569. PMLR, 2023. URL https://proceedings.mlr.press/v202/hao23c.html.

Zhongkai Hao, Chang Su, Songming Liu, Julius Berner, Chengyang Ying, Hang Su, Anima Anandkumar, Jian Song, and Jun Zhu. DPOT: Auto-Regressive Denoising Operator Transformer for Large-Scale PDE Pre-Training. In *International Conference on Machine Learning*, pp. 17616–17635. PMLR, 2024.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for PDEs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2311. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/84e1b1ec17bb11c57234e96433022a9a-Abstract-Conference.html.

Shudong Huang, Wentao Feng, Chenwei Tang, Zhenan He, Caiyang Yu, and Jiancheng Lv. Partial differential equations meet deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):13649–13669, 2025.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Zongyi Li, Nikola Kovachki, Chris Choy, Boyi Li, Jean Kossaifi, Shourya Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, and Anima Anandkumar. Geometry-informed neural operator for large-scale 3d PDEs. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1556. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/70518ea42831f02afc3a2828993935ad-Abstract-Conference.html.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/IMS Journal of Data Science*, 1(3):1–27, 2024. doi: 10.1145/3648506. URL https://doi.org/10.1145/3648506.

Yuxuan Liu, Jingmin Sun, Xinjie He, Griffin Pinney, Zecheng Zhang, and Hayden Schaeffer. PROSE-FD: A multimodal PDE foundation model for learning multiple operators for forecasting fluid dynamics. In *NeurIPS 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*, 2024a.

Yuxuan Liu, Zecheng Zhang, and Hayden Schaeffer. PROSE: Predicting multiple operators and symbolic expressions using multimodal transformers. *Neural Networks*, 180:106707, 2024b. doi: 10.1016/j.neunet.2024.106707. URL https://doi.org/10.1016/j.neunet.2024.106707.

Zichao Long, Yiping Lu, and Bin Dong. PDE-Net 2.0: Learning PDEs from data with a numeric-symbolic hybrid deep network. *Journal of Computational Physics*, 399:108925, 2019. doi: 10.1016/j.jcp.2019.108925. URL https://doi.org/10.1016/j.jcp.2019.108925.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3):218–229, 2021.

Yining Luo, Yingfa Chen, and Zhen Zhang. CFDBench: A large-scale benchmark for machine learning methods in fluid dynamics. *arXiv preprint arXiv:2310.05963*, 2023. URL https://arxiv.org/abs/2310.05963.

Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Géraud Krawezik, François Lanusse, Mariel Pettee, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. Multiple physics pretraining for spatiotemporal surrogate models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3791. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d7cb9db5ade2db7814fbd01ee59f4c7b-Abstract-Conference.html.

Mohammad Nooraiepour, Jakub Wiktor Both, Teeratorn Kadeethum, and Saeid Sadeghnejad. Partial differential equations in the age of machine learning: a critical synthesis of classical, machine learning, and hybrid methods. *Artificial Intelligence Review*, 2026. ISSN 1573-7462. doi: 10.1007/s10462-026-11677-5.

Mahindra Singh Rautela, Alexander Most, Siddharth Mansingh, Bradley C Love, Ayan Biswas, Diane Oyen, and Earl Lawrence. Morph: Shape-agnostic pde foundation models. *arXiv preprint arXiv:2509.21670*, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Unique Subedi and Ambuj Tewari. Operator learning: A statistical perspective. *Annual Review of Statistics and Its Application*, 13:123–148, 2026. doi: 10.1146/annurev-statistics-042424-070908.

Dengdi Sun, Xiaoya Zhou, Xiao Wang, Hao Si, Wanli Lyu, Jin Tang, and Bin Luo. Nestor: A nested moe-based neural operator for large-scale pde pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6147–6156, June 2026a.

Li Sun, Hongbo Lv, Zhikai Jiang, Zhongtian Sun, Lanxu Yang, and Philip S. Yu. Origo: Interpretable multi-physics PDE foundation model through neural operator splitting. In *Forty-third International Conference on Machine Learning*, 2026b.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBENCH: An Extensive Benchmark for Scientific Machine Learning. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks*, 2022.

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized fourier neural operators. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tmIiMPl4IPa.

Hong Wang, Haiyang Xin, Jie Wang, Xuanze Yang, Fei Zha, Huanshuo Dong, and Yan Jiang. Mixture-of-experts operator transformer for large-scale PDE pre-training. *arXiv preprint arXiv:2510.25803*, 2025. URL https://arxiv.org/abs/2510.25803.

Zhanhong Ye, Xiang Huang, Leheng Chen, Hongsheng Liu, Zidong Wang, and Bin Dong. Pdeformer: Towards a foundation model for one-dimensional partial differential equations. In *ICLR 2024 Workshop on AI4DifferentialEquations In Science*, 2024.

Zhanhong Ye, Zining Liu, Bingyang Wu, Hongjie Jiang, Leheng Chen, Minyan Zhang, Xiang Huang, Qinghe Meng, Jingyuan Zou, Hongsheng Liu, and Bin Dong. Pdeformer-2: A versatile foundation model for two-dimensional partial differential equations. *arXiv preprint arXiv:2507.15409*, 2025.

Hang Zhou, Yuezhou Ma, Haixu Wu, Haowen Wang, and Mingsheng Long. Unisolver: PDE-conditional transformers towards universal neural PDE solvers. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79061–79088. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhou25r.html.

A APPENDIX

A.1 CANONICAL TERM GRAPH CONSTRUCTION

Term partition. Let $F^{(c)}[u; \boldsymbol{\lambda}]$ be the declared right-hand side of the evolution equation for component c, after any auxiliary state variables have been introduced. Define $\mathrm{Split}_+(F^{(c)}) = (s_m)_{m \in \mathcal{M}_c}$ by flattening additions and collecting the signed summands at the outermost additive level. Products, nonlinear functions, and derivatives remain inside their summands. Then

$$\partial_t u^{(c)} = F^{(c)}[u; \boldsymbol{\lambda}] = \sum_{m \in \mathcal{M}_c} s_m[u; \boldsymbol{\lambda}], \qquad \mathcal{O}_m := s_m. \tag{8}$$

The sign belongs to s_m. A prescribed external forcing is a summand even when it has no state-variable input. In a coupled system, c remains part of the term identity, so identical operator expressions in different target equations are distinct terms.

Term records and graph edges. For each $m \in \mathcal{M}_c$, let χ_m be its signed outer coefficient and T_m its remaining operator tree. Set $\chi_m = 1$ when there is no separable outer coefficient; a variable coefficient inside a derivative stays in T_m. The term record and its evaluation are

$$v_m = \big(c, \mathsf{T}_m, \chi_m, \mathrm{src}(\mathsf{T}_m)\big), \qquad \mathcal{O}_m[u; \boldsymbol{\lambda}] = \chi_m \, \mathrm{Eval}(\mathsf{T}_m; u, \boldsymbol{\lambda}). \tag{9}$$

Here $\mathrm{src}(\mathsf{T}_m)$ is the set of state variables used by the tree. The graph contains typed nodes for variables, coefficients, constants, differential and algebraic operators, term roots, target equations, boundary conditions, and geometry. Directed, typed edges record operator inputs, the term-to-target-equation link, and the declared associations of boundary and geometry descriptors. Thus the graph retains both the internal structure of a term and the variables shared by different terms.

Canonicalization. Before building the graph, a syntactic normalizer standardizes derivative notation, the position of an outer scalar coefficient, the ordering of commutative operands, and variable names while preserving their roles. Higher derivatives can be represented by compositions of first-derivative primitives, for example $\partial_{xx} u \mapsto D_x(D_x(u))$. The normalizer does not distribute products over additions or move coefficients across derivatives. In particular, conservative and nonconservative expressions retain different trees unless an explicit equivalence rule is supplied. This convention fixes the term boundary relative to the declared equation form; it does not identify every mathematically equivalent PDE presentation.

For example, the declared form

$$\partial_t u = -u \, \partial_x u + \nu \, \partial_{xx} u + f(x, t) \tag{10}$$

produces separate convection, diffusion, and forcing terms. The alternative convection expression $-\partial_x(u^2/2)$ retains its conservative operator tree rather than being silently replaced by $-u \, \partial_x u$. For each PDE instance, the resulting graph and its term embeddings remain fixed along a trajectory; the evaluated term responses change with the state.

A.2 PROOFS AND SUPPLEMENTARY ANALYSIS

All statements below condition on a supplied PDE instance and a time interval on which its exact solution and term responses exist. Numerical and representation bounds concern terms marked computable by the response mask. Comparisons between histories use the same PDE coefficients, geometry, boundary conditions, and query grid. Constants are local to the stated state region.

A.2.1 RESPONSE IDENTITY AND AGGREGATION

Fix a PDE instance and a state u_t at which every term in equation c is defined. The declared decomposition gives

$$\partial_t u_t^{(c)} = \sum_{m \in \mathcal{M}_c} \mathcal{O}_m[u_t; \boldsymbol{\lambda}] = \sum_{m \in \mathcal{M}_c} \rho^t_{m, K_r}. \tag{11}$$

This is an identity conditional on the supplied governing equation. For a nontrivial linear space X of response fields, define $A_c : X^{|\mathcal{M}_c|} \to X$ by $A_c((r_m)_m) = \sum_m r_m$. If $|\mathcal{M}_c| \geq 2$, choose distinct

indices i, j and any $0 \neq \delta \in X$. Replacing (r_i, r_j) by $(r_i + \delta, r_j - \delta)$ changes the response tuple but preserves its sum. Thus A_c is not injective on $X^{|\mathcal{M}_c|}$. The set of tuples realizable by one fixed equation and a restricted state family may be smaller; the result makes no claim of non-injectivity on that set.

A.2.2 Response estimation and representation stability

Discrete response error. Let Π_h sample sufficiently regular states on a spatial grid. Distinguish the state norm $\|\cdot\|_{h,u}$ from the response norm $\|\cdot\|_{h,\rho}$; the main text abbreviates the latter as $\|\cdot\|_h$. At an exact state u_t, assume the term evaluator satisfies

$$\big\|\widehat{\mathcal{R}}_{m,h}(\Pi_h u_t) - \Pi_h \mathcal{O}_m[u_t; \boldsymbol{\lambda}]\big\|_{h,\rho} \leq \varepsilon_{m,\mathrm{disc}}(h). \tag{12}$$

Assume also that, on the grid states under consideration,

$$\|\widehat{\mathcal{R}}_{m,h}(v_h + \xi_h) - \widehat{\mathcal{R}}_{m,h}(v_h)\|_{h,\rho} \leq \kappa_{m,h}\|\xi_h\|_{h,u}. \tag{13}$$

With $\hat{\rho}^t_{m,h} = \widehat{\mathcal{R}}_{m,h}(\Pi_h u_t + \xi_h)$ and $\|\xi_h\|_{h,u} \leq \eta$, the triangle inequality applied at $v_h = \Pi_h u_t$ yields

$$\big\|\hat{\rho}^t_{m,h} - \Pi_h \rho^t_{m,K_r}\big\|_{h,\rho} \leq \kappa_{m,h}\eta + \varepsilon_{m,\mathrm{disc}}(h). \tag{14}$$

For a predicted grid state $\hat{v}_{t,h}$ in the same stability region, the same argument makes the feedback dependence explicit:

$$\big\|\widehat{\mathcal{R}}_{m,h}(\hat{v}_{t,h}) - \Pi_h \rho^t_{m,K_r}\big\|_{h,\rho} \leq \varepsilon_{m,\mathrm{disc}}(h) + \kappa_{m,h}\|\hat{v}_{t,h} - \Pi_h u_t\|_{h,u}. \tag{15}$$

No grid-independent value of $\kappa_{m,h}$ is assumed.

From a response window to an enriched term. For one term and each $k = 1, \ldots, K_r$, set $v_k = u_{t-K_r+k}$, $r_k = \Pi_h \mathcal{O}_m[v_k; \boldsymbol{\lambda}]$, and $\hat{r}_k = \widehat{\mathcal{R}}_{m,h}(\Pi_h v_k + \xi_{k,h})$. Write $r = (r_1, \ldots, r_{K_r})$ and $\hat{r} = (\hat{r}_1, \ldots, \hat{r}_{K_r})$. The response window has norm

$$\|r\|_W = \left(\frac{1}{K_r}\sum_{k=1}^{K_r}\|r_k\|^2_{h,\rho}\right)^{1/2}, \qquad a(r) = \|r\|_W + \epsilon, \qquad \Phi(r) = \big(r/a(r), \log a(r)\big), \tag{16}$$

where $\epsilon > 0$. This norm agrees with the RMS normalization in the method when $\|\cdot\|_{h,\rho}$ uses the corresponding spatial average. If the consistency and stability bounds hold at each v_k, with constants $\varepsilon_{m,k,\mathrm{disc}}(h)$ and $\kappa_{m,k,h}$ and noise bound $\|\xi_{k,h}\|_{h,u} \leq \eta_k$, then

$$d := \|\hat{r} - r\|_W \leq \left(\frac{1}{K_r}\sum_{k=1}^{K_r}[\varepsilon_{m,k,\mathrm{disc}}(h) + \kappa_{m,k,h}\eta_k]^2\right)^{1/2}. \tag{17}$$

The reverse triangle inequality gives $|a(\hat{r}) - a(r)| \leq d$. Since $a(r), a(\hat{r}) \geq \epsilon$ and $\|r\|_W \leq a(r)$, the decomposition $\hat{r}/a(\hat{r}) - r/a(r) = [\hat{r} - r]/a(\hat{r}) + r[1/a(\hat{r}) - 1/a(r)]$ gives the first bound below. The second follows from the mean value theorem for $\log$ on $[\epsilon, \infty)$.

$$\left\|\frac{\hat{r}}{a(\hat{r})} - \frac{r}{a(r)}\right\|_W \leq \frac{2d}{\epsilon}, \qquad |\log a(\hat{r}) - \log a(r)| \leq \frac{d}{\epsilon}. \tag{18}$$

Consequently, the product norm $\|(v, b)\|_{W\times\mathbb{R}} = (\|v\|^2_W + |b|^2)^{1/2}$ obeys $\|\Phi(\hat{r}) - \Phi(r)\|_{W\times\mathbb{R}} \leq \sqrt{5}\, d/\epsilon$.

Let $\mathsf{E}_{m,h}$ denote the complete patch encoder and resampler acting on $\Phi(r)$ and returning both the response tokens and the log-scale feature, with coordinates held fixed. Let $\Gamma_m(e_m, \cdot)$ denote term-local enrichment through LayerNorm. If these maps are locally L_E- and L_Γ-Lipschitz, respectively, along the two inputs under comparison, then

$$\begin{aligned}\|\tilde{e}^t_m - \tilde{e}^{t,\star}_m\| &\leq L_\Gamma\|\mathsf{E}_{m,h}(\Phi(\hat{r})) - \mathsf{E}_{m,h}(\Phi(r))\| \\ &\leq L_\Gamma L_E\|\Phi(\hat{r}) - \Phi(r)\|_{W\times\mathbb{R}} \leq L_\Gamma L_E\frac{\sqrt{5}}{\epsilon}d.\end{aligned} \tag{19}$$

Combining this chain with the response-window estimate gives

$$\|\tilde{e}_m^t - \tilde{e}_m^{t,\star}\| \le L_\Gamma L_E \frac{\sqrt{5}}{\epsilon} \left(\frac{1}{K_r} \sum_{k=1}^{K_r} [\varepsilon_{m,k,\mathrm{disc}}(h) + \kappa_{m,k,h}\eta_k]^2 \right)^{1/2}. \tag{20}$$

The starred representation uses exact sampled response fields with the same coordinates, term identity, and validity mask. A change in response validity is outside this bound. The estimate compares two representations normalized on the same grid. If a_{cont} denotes the continuous-window scale in the method and a quadrature estimate gives $|a(r) - a_{\mathrm{cont}}| \le q_m(h)$, replacing the grid scale by a_{cont} adds at most $\sqrt{2}\, q_m(h)/\epsilon$ to the bound on the pair $\Phi(r)$, before multiplication by $L_E L_\Gamma$. This follows by applying the same reciprocal-scale and logarithm estimates with the response fields fixed.

A.2.3 Gated update and composition

Gated residual. Before LayerNorm the enriched term is $p_m^t = e_m + g_m^t \odot \delta e_m^t$. For the Euclidean norm, the sigmoid gate gives $0 \le g_{m,j}^t \le 1$, hence

$$\|p_m^t - e_m\|_2 = \|g_m^t \odot \delta e_m^t\|_2 \le \|\delta e_m^t\|_2. \tag{21}$$

This inequality concerns the pre-normalization vector. It does not imply $\|\tilde{e}_m^t - e_m\|_2 \le \|\delta e_m^t\|_2$, because $\tilde{e}_m^t = \mathrm{LN}(p_m^t)$.

Term composition. Let $\mathbb{V}$ be a vocabulary of admissible term operator trees. Provided the graph encoder accepts finite graphs within its token capacity and the shared response encoder and enrichment map accept every term in $\mathbb{V}$, the same parameterized representation map is defined for any finite admissible combination of these terms. Each new combination supplies a finite term graph, the shared modules process its term nodes, and the backbone receives the resulting finite token sequence. This is architectural closure under a new term combination; prediction quality on that combination is an empirical property.

A.2.4 Finite history and spatial reduction

An input approximation bound needs separate regularity assumptions. Suppose a full-history next-state map F_∞ has a K-frame truncation F_K such that $\|F_\infty(H_\infty) - F_K(H_K)\| \le C_t \varrho^K$ for $0 < \varrho < 1$. Suppose every retained frame admits a grid reconstruction $I_h \Pi_h u$ with error at most $C_x h^\beta$, and F_K is L_K-Lipschitz in the maximum history norm. The triangle inequality then gives

$$\begin{aligned} \|F_\infty(H_\infty) - F_K(I_h\Pi_h H_K)\| &\le \|F_\infty(H_\infty) - F_K(H_K)\| + \|F_K(H_K) - F_K(I_h\Pi_h H_K)\| \\ &\le C_t \varrho^K + L_K C_x h^\beta, \end{aligned} \tag{22}$$

Here $I_h\Pi_h H_K$ applies the same reconstruction to each retained frame. This bound concerns truncation and grid reconstruction before learned token compression; it does not assign a reconstruction rate to the resampler.

A.2.5 Autoregressive rollout

Let $F_{\mathcal{P}}$ be the exact history-to-next-state map on the evaluator grid for the specified PDE and F_θ the complete TERM map, including response evaluation on the input history. Timestamps and PDE metadata are held fixed in each comparison; the following constants are uniform over the stated rollout horizon. The output state norm is $\|\cdot\|_{h,u}$, and a K-frame history has norm $\|H - H'\|_{\max} = \max_{1\le k\le K} \|H_k - H'_k\|_{h,u}$. Assume that exact and predicted histories stay in a region where

$$\begin{aligned} &\|F_\theta(H) - F_\theta(H')\| \le L_{\mathrm{TERM}} \|H - H'\|_{\max}, \\ &\|F_\theta(H) - F_{\mathcal{P}}(H)\| \le \varepsilon_{\mathrm{step}} \quad \text{for exact histories } H. \end{aligned} \tag{23}$$

The Lipschitz assumption on F_θ can be traced to its modules. With the PDE graph, validity masks, and query grid fixed, write the one-step map schematically as

$$F_\theta(H) = D_\theta\big(B_\theta\big(\Gamma_\theta(e, \mathsf{R}_\theta(H)), \mathsf{S}_\theta(H)\big)\big). \tag{24}$$

Here R_θ stacks the response-token and log-scale outputs of all terms, S_θ is the state-summary branch, Γ_θ stacks the term-local updates, B_θ is the shared backbone, and D_θ queries the evaluation grid.

Suppose these maps are locally Lipschitz with constants $L_R, L_S, L_\Gamma, L_B, L_D$, respectively, in compatible product norms. The composition bound is

$$L_{\mathrm{TERM}} \le L_D L_B \big[(L_\Gamma L_R)^2 + L_S^2\big]^{1/2}. \tag{25}$$

For a fixed validity mask, let $L_{E,m}$ bound the token encoder of term m, and let $\kappa_{m,h}$ bound its discrete evaluator on each response frame. The window-normalization estimate above then permits

$$L_R \le \frac{\sqrt{5}}{\epsilon} \left(\sum_{m \in \mathcal{M}^t_{\mathrm{resp}}} (L_{E,m}\kappa_{m,h})^2 \right)^{1/2}. \tag{26}$$

These are sufficient local bounds. They explain how derivative sensitivity enters the full map, but do not imply $L_{\mathrm{TERM}} < 1$. The one-step approximation term $\varepsilon_{\mathrm{step}}$ is defined for the complete implemented predictor on exact histories; it already includes its discretization and response-estimation errors. The Lipschitz factor controls the additional change when predicted histories replace exact ones. With $e_n = \|\hat{u}_{t+n} - u_{t+n}\|_{h,u}$ on the evaluator grid and an exact initial K-frame history, set $e_j = 0$ for $j \le 0$. At step n, let H_{t+n} and $\hat{H}_{t+n}$ be the exact and predicted history windows. Adding and subtracting $F_\theta(H_{t+n})$ gives

$$\hat{u}_{t+n+1} - u_{t+n+1} = F_\theta(\hat{H}_{t+n}) - F_\theta(H_{t+n}) + F_\theta(H_{t+n}) - F_{\mathcal{P}}(H_{t+n}). \tag{27}$$

The two terms are controlled by the Lipschitz and one-step assumptions, respectively, yielding

$$e_{n+1} \le \varepsilon_{\mathrm{step}} + L_{\mathrm{TERM}} \max_{n-K+1 \le j \le n} e_j \le \varepsilon_{\mathrm{step}} + L_{\mathrm{TERM}} \max_{0 \le j \le n} e_j. \tag{28}$$

Set $E_n = \max_{0 \le j \le n} e_j$ and $B_n = \varepsilon_{\mathrm{step}} \sum_{j=0}^{n-1} L^j_{\mathrm{TERM}}$, with $B_0 = 0$. The sequence B_n is nondecreasing. If $E_n \le B_n$, the recurrence gives $e_{n+1} \le \varepsilon_{\mathrm{step}} + L_{\mathrm{TERM}} B_n = B_{n+1}$; therefore $E_{n+1} \le B_{n+1}$, completing the induction. For $L_{\mathrm{TERM}} < 1$ this yields $e_n \le \varepsilon_{\mathrm{step}}/(1 - L_{\mathrm{TERM}})$; for $L_{\mathrm{TERM}} = 1$ it gives $e_n \le n\varepsilon_{\mathrm{step}}$. The constant L_{TERM} is an assumption on the full predictor, rather than a property established by its term-local architecture.

For an inexact initial history, set $E_0 = \max_{-K+1 \le j \le 0} e_j$ and $E_n = \max_{-K+1 \le j \le n} e_j$. The same recurrence implies $E_{n+1} \le \max\{E_n, \varepsilon_{\mathrm{step}} + L_{\mathrm{TERM}} E_n\}$. If $L_{\mathrm{TERM}} < 1$, the interval $[0, \max\{E_0, \varepsilon_{\mathrm{step}}/(1 - L_{\mathrm{TERM}})\}]$ is invariant under this update, yielding a uniform bound. If $L_{\mathrm{TERM}} = 1$, then $E_n \le E_0 + n\varepsilon_{\mathrm{step}}$. If $L_{\mathrm{TERM}} > 1$, iteration gives $E_n \le L^n_{\mathrm{TERM}} E_0 + \varepsilon_{\mathrm{step}} \sum_{j=0}^{n-1} L^j_{\mathrm{TERM}}$.

A.3 Implementation and Experimental Details

A.3.1 Datasets and preprocessing

Table 5 gives the trajectory-level splits used for joint pretraining and evaluation. The ten tasks are drawn from FNO-ν (Li et al., 2021), PDEBench (Takamoto et al., 2022), PDEArena (Gupta & Brandstetter, 2022), and CFDBench (Luo et al., 2023), following the suite assembled by Origo (Sun et al., 2026b). Each training update first selects a dataset and then samples a trajectory and a valid temporal window from that dataset, balancing the contribution of sources with different sizes.

Table 5: Datasets used for joint pretraining. The three FNO-ν rows are separate viscosity settings.

Source	PDE or setting	Train	Test
FNO-ν	NS, $\nu = 10^{-5}$	1,000	200
	NS, $\nu = 10^{-4}$	1,000	200
	NS, $\nu = 10^{-3}$	1,000	200
PDEBench	Burgers (Bur.)	900	60
	Diffusion–reaction (DR)	900	60
	Shallow-water equations (SWE)	900	60
	Compressible NS, CNS-$(1, 0.1)$	9,000	200
PDEArena	Navier–Stokes (NS)	2,000	200
	Conditioned NS (NS-c)	1,000	200
CFDBench	CFD benchmark	9,000	1,000

One-dimensional fields are mapped to 128 points and two-dimensional fields to 128×128 grids. Coarser fields are interpolated and finer fields are downsampled for the common model input. The response branch evaluates operators on the native-resolution fields before remapping, whereas the state branch consumes the common-grid representation; native-resolution fields are also retained for cross-resolution tests. Variable channels are padded to the maximum channel count and accompanied by a validity mask. A geometry mask marks physical cells for nonrectangular domains and obstacles. Each variable has its own type embedding, so channel position does not define variable identity.

The governing equation supplied with each sample is canonicalized before training. Additive terms, coefficients, source and target variables, boundary conditions, and geometry descriptors form the graph described in Appendix A.1. Graphs are fixed within a trajectory and cached for rollout. The canonical operator-tree signature is also used to audit held-out compositions: a test composition is excluded as a complete graph while its prescribed constituent terms, and for triple tests its pairwise subcompositions, remain represented in training.

A.3.2 UNSEEN-EQUATION TERM COVERAGE

To distinguish the two unseen-equation settings, we audit the physical-term vocabulary present during pretraining. Table 6 lists the principal operator families in each training PDE. No pretraining equation contains the third-order dispersion operator u_{xxx}, so KdV tests transfer to a novel operator, while AC recombines diffusion and reaction families observed during pretraining.

Table 6: Physical-term vocabulary coverage in the pretraining PDEs. ✓ denotes presence, × denotes absence, and "Task" denotes task-dependent inclusion.

Source	PDE or setting	Adv.	Diff.	React.	Press.	Grav.	Comp.	Disp. u_{xxx}
FNO-ν	NS, $\nu = 10^{-5}$	✓	✓	×	✓	×	×	×
	NS, $\nu = 10^{-4}$	✓	✓	×	✓	×	×	×
	NS, $\nu = 10^{-3}$	✓	✓	×	✓	×	×	×
PDEBench	Burgers (Bur.)	✓	✓	×	×	×	×	×
	Diffusion–reaction (DR)	×	✓	✓	×	×	×	×
	Shallow-water equations (SWE)	✓	×	×	✓	✓	×	×
	Compressible NS, CNS-$(1, 0.1)$	✓	✓	×	✓	×	✓	×
PDEArena	Navier–Stokes (NS)	✓	✓	×	✓	×	×	×
	Conditioned NS (NS-c)	✓	✓	×	✓	×	×	×
CFDBench	CFD benchmark	Task	Task	×	✓	×	×	×

A.3.3 TERM-M ARCHITECTURE

All main, zero-shot, robustness, and rollout experiments use TERM-M unless a scale is named explicitly. Table 7 consolidates the settings from the model specification and the implementation record. The graph encoder and foundation Transformer use pre-LayerNorm and GELU; the reported dropout is fixed at 0.05.

Table 7: TERM-M implementation used in the reported experiments.

Component	Configuration
History / response / training rollout	$K = 10$, $K_r = 4$, $T = 10$
Graph encoder	4 layers, width 384, 6 heads, FFN 1536
Response encoder	2 layers, width 256
Response resampler	4 tokens per physical term
Trajectory encoder	4 layers, width 384, 6 heads
Trajectory resampler	32 summary tokens
Foundation Transformer	12 layers, width 768, 12 heads, FFN 3072
Continuous query decoder	2 cross-attention layers, width 384, 6 heads, FFN 1536
Coordinate encoding	32 Fourier frequencies
Two-dimensional state patch size	8×8
Dropout / trainable parameters	0.05 / 105.4M

Responses are evaluated on the original observation grid before spatial remapping. Periodic structured grids use Fourier differentiation with the $2/3$ de-aliasing rule. Nonperiodic structured grids use fourth-order centered differences in the interior and fourth-order one-sided stencils at boundaries. A simulator-provided discrete operator is used when available; irregular grids use local polynomial or moving-least-squares differentiation. Each term is normalized by the RMS of its visible four-frame response window as in Equation 3. An unavailable response is represented by a learned missing-response token and its validity mask.

Before the learned resampler, response fields use 32 spatial positions in one dimension, an 8×8 layout in two dimensions, and a $4 \times 4 \times 4$ layout in three dimensions. Every physical term is then compressed to four response tokens. The state branch uses values, coordinates, relative time, physical time step, and variable identity, and produces 32 summary tokens. The query decoder combines 32 Fourier coordinate frequencies with a variable embedding and reads from the backbone through two cross-attention layers.

A.3.4 Optimization and comparison protocol

TERM-M is pretrained on four NVIDIA RTX 5090 GPUs with AdamW, an initial learning rate of 2×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.05, and global gradient clipping at 1.0. Training runs for 1,000 epochs with 50 warm-up epochs followed by cosine decay, a total batch size of 24, and BF16 mixed precision. The closed-loop objective in Equation 5 uses a ten-step rollout and $\lambda_{\text{resp}} = 0.1$. Every predicted frame is fed into the next history window. Whenever an ablation retains an identifiable response representation, we preserve the response-prediction objective with $\lambda_{\text{resp}} = 0.1$, and the prediction target follows the response actually retained by that variant. No auxiliary target is introduced when doing so would restore term-level information removed by the ablation.

Task-specific experts are trained independently on each target dataset. Unified models share one pretraining run across the ten tasks. Their adapted variants use 200 additional epochs on the target dataset. TERM and any reimplemented baselines use the dataset splits described in Table 5. For methods whose benchmark results are taken from prior work, we report the published values under the corresponding benchmark setting and identify their provenance in the relevant table captions. In particular, Origo results are taken from Sun et al (Sun et al., 2026b) rather than from a separate reimplementation.

A.3.5 Ablation variant construction

Let $E_{\mathcal{P}} = (e_1, \ldots, e_{M_{\mathcal{P}}})$ denote the symbolic term embeddings, $R^t = (R^t_1, \ldots, R^t_{M_{\mathcal{P}}})$ the resampled response-token sets, and S_t the trajectory summaries. The complete model forms one conditioning token per physical term through the local enrichment map Γ,

$$C^t_{\text{TERM}} = \big(\Gamma(e_m, R^t_m, \ell^t_m)\big)_{m=1}^{M_{\mathcal{P}}}, \tag{29}$$

and supplies C^t_{TERM} and S_t to the foundation Transformer. All ablations retain the same trajectory encoder, foundation Transformer, query decoder, data split, optimizer, and training schedule.

Response-based variants use the same evaluator, window normalization, response encoder, and resampler unless the removed component makes one of these modules inapplicable. The information-matched controls retain the response feature width and resampling budget of TERM.

Trajectory-only Only S_t is supplied to the foundation Transformer. The equation graph, symbolic term tokens, graph-relation attention bias, response fields, and response-enrichment loss are removed. Coordinates, timestamps, variable identities, and the query decoder remain unchanged.

Static Graph The model receives S_t together with the persistent term embeddings $E_{\mathcal{P}}$ and the graph-relation bias. The response evaluator and response encoder are disabled, and e_m enters the backbone directly in place of the enriched token $\tilde{e}_m^t$. This construction preserves equation identity while preventing the term representation from changing with the observed state.

Response-only Each physical term is evaluated and encoded exactly as in TERM, but the symbolic term embeddings and graph-relation bias are removed before the backbone. The response-derived tokens therefore expose the current operator outputs without explicit operator type, coefficient, or graph-neighborhood features. The governing equation is used by the external evaluator to construct the response fields but is not presented to the learned predictor as a symbolic graph.

Global Response The symbolic graph and S_t are retained. Instead of enriching term m from R_m^t, all response tokens are concatenated and compressed by shared learned queries into a single global response bank,

$$R_{\mathrm{global}}^t = \mathrm{Resample}_{\mathrm{global}}\big(\mathrm{Concat}_m\, R_m^t\big). \tag{30}$$

This bank is supplied alongside the graph tokens, with no index indicating which response generated an individual term update. It retains current physical activity at the set level while removing term-wise attribution.

RHS-Matched For every response frame and state component, the signed term fields are summed before response encoding,

$$\rho_{\mathrm{RHS},c,k}^t = \sum_{m \in \mathcal{M}_c} \rho_{m,k}^t. \tag{31}$$

The resulting right-hand-side fields use the same response encoder and the same total number and width of learned resampling queries as TERM. The model also retains $E_{\mathcal{P}}$ and S_t, but the response stream contains no recoverable source label for an individual summand. Thus this control preserves the aggregate instantaneous dynamics and the response-token budget while discarding term-level decomposition.

Identity-Free The separate response sets R_m^t and their scale features are retained, so no response field is summed or discarded. Symbolic term embeddings, coefficients encoded in the graph, typed graph edges, and graph-relation biases are withheld. Each response set passes through the same shared encoder and an anonymous term slot. This preserves response granularity and token count while removing explicit information about which operator produced each response.

Shuffled Binding For each training sample and temporal window, we independently draw a fixed-point-free permutation π over the valid response terms within each equation. The same permutation is applied to all frames in that response window, while a new permutation is sampled for another sample. At evaluation, a fresh fixed-point-free permutation is likewise sampled for each sample/window.

$$\Gamma(e_m, R_m, \ell_m) \longrightarrow \Gamma(e_m, R_{\pi(m)}, \ell_{\pi(m)}). \tag{32}$$

Unbound Response Both $E_{\mathcal{P}}$ and every R_m^t are available, but the term-local attention restriction is removed. Each symbolic term attends to the pooled union of response tokens,

$$\delta e_m^t = \mathrm{CrossAttn}\left(e_m, \mathrm{Concat}_j\, R_j^t, \mathrm{Concat}_j\, R_j^t\right), \tag{33}$$

followed by the same gated residual update used by TERM. The backbone therefore receives the same symbolic identities and response information, but the architecture does not supply the association $e_m \leftrightarrow R_m^t$.

TERM The complete construction in Equation 29 restricts each local enrichment update to the response generated by the same symbolic term. It retains the graph relation bias and passes the resulting response-enriched term tokens, together with S_t, through the shared backbone. Comparisons with the information-matched controls therefore test the organization of the available physical information, whereas the component ablations test which information streams contribute to prediction.

A.3.6 STRICT-COMPOSITION BENCHMARK CONSTRUCTION

The controlled vocabulary consists of advection $A[u] = -u\partial_x u$, diffusion $D[u] = \nu\partial_{xx}u$, reaction $R[u] = \kappa(u - u^3)$, and dispersion $S[u] = -\beta\partial_{xxx}u$. The one-dimensional training set contains the six pairwise equations $A + D$, $A + R$, $A + S$, $D + R$, $D + S$, and $R + S$. Testing uses $A + D + R$, $A + D + S$, $A + R + S$, and $D + R + S$. Each training composition has 1,000 training and 100 validation trajectories; each held-out composition has 200 test trajectories. All are solved on the periodic domain $[0, 2\pi]$ with 256 grid points.

The two-dimensional setting trains on $A + D$, $A + R$, and $D + R$ and tests on $A + D + R$ at 128×128 resolution. Training and test equations use the same coefficient ranges, initial-condition family, and numerical solver protocol. Before training, canonical operator-tree signatures verify that no complete test composition occurs in training and that every constituent term and pairwise subcomposition is covered. All zero-shot evaluations freeze model parameters and provide the target equation as an input specification.

A.4 GENERALIZATION ANALYSIS

A.4.1 CROSS-DATASET GENERALIZATION

This experiment evaluates transfer across dataset distributions, discretizations, and parameter regimes through independent leave-one-dataset-out pretraining runs. All trajectories associated with the target PDE are removed during pretraining. At test time, the governing equation and ten recent frames are provided without gradient updates. The target compositions use A for advection, D for diffusion, G for gravity, P for pressure, and C for compressibility.

Table 8: Leave-one-dataset-out zero-shot rel-L_2. **Bold** denotes the lowest reported error for each target PDE and for the average.

Method	Bur. $A + D$	SWE $A + G$	NS $A + P + D$	CNS $A+P+D+C$	Avg.
DPOT-M	0.0196	0.00931	0.1281	0.0194	0.0441
MoE-POT-M	0.0181	0.00874	0.1194	0.0178	0.0409
OmniArch-M	0.0217	0.00821	0.1422	0.0213	0.0483
Origo-M	–	–	–	–	–
Static Graph	0.0174	0.00862	0.1168	0.0171	0.0395
Global Response	0.0153	0.00791	0.1085	0.0155	0.0367
TERM-M	**0.0126**	**0.00658**	**0.0943**	**0.0132**	**0.0317**

Table 8 shows that TERM-M is best on all four held-out datasets and lowers the macro-average from 0.0367 for Global Response to 0.0317, a 13.6% reduction. Static Graph improves on the unified pretrained baselines on average, and adding global response information lowers its error further. The remaining gain from Global Response to TERM-M is consistent with a contribution from term-level organization under changes in dataset distribution and discretization. Because related full compositions may occur in other sources, this experiment measures cross-dataset transfer; the benchmark in Appendix A.3.6 supplies the stricter composition control.

A.4.2 UNSEEN PHYSICAL PARAMETER GENERALIZATION

For each FNO-NS viscosity, a separate model is pretrained on the other two values and evaluated at the held-out value. The governing coefficient is supplied at test time and no parameter update is performed.

Table 9: Zero-shot rel-L_2 at unseen viscosity values. **Bold** denotes the lowest error for each held-out value and for the average.

Method	10^{-5} Extra.	10^{-4} Interp.	10^{-3} Extra.	Avg.
Trajectory-only	0.0618	0.0386	0.0112	0.0372
Static Graph	0.0547	0.0331	0.0094	0.0324
Unbound Response	0.0489	0.0292	0.0081	0.0287
TERM	**0.0432**	**0.0251**	**0.0070**	**0.0251**

Table 9 shows that TERM gives the lowest error for interpolation and both extrapolation directions. Relative to Unbound Response, its reductions are 11.7%, 14.0%, and 13.6% at 10^{-5}, 10^{-4}, and 10^{-3}, respectively. The low-viscosity extrapolation remains the hardest setting for every variant. The progression from Trajectory-only to Static Graph to Unbound Response to TERM helps distinguish the value of the supplied coefficient, current response information, and explicit term–response association.

A.4.3 DYNAMICAL REGIME ANALYSIS

Burgers test windows are partitioned by gradient intensity into smooth, transition, and sharp regimes. NS windows are partitioned by Reynolds number and flow complexity into low-, mid-, and high-Re regimes. Within each PDE the governing graph is unchanged across regimes, while TERM recomputes its responses from the current history.

Table 10: Rel-L_2 across Burgers and NS dynamical regimes. **Bold** denotes the lowest error in each regime.

Method	Burgers			NS		
	Smooth	Transition	Sharp	Low-Re	Mid-Re	High-Re
Trajectory-only	0.0084	0.0127	0.0219	0.0478	0.0818	0.1377
Static Graph	0.0076	0.0111	0.0186	0.0433	0.0735	0.1213
Global Response	0.0071	0.0096	0.0152	0.0404	0.0661	0.1074
TERM	**0.0067**	**0.0083**	**0.0124**	**0.0378**	**0.0596**	**0.0914**

Table 10 shows that TERM is best in all six regimes, with its margin increasing as the dynamics become more demanding. Against Static Graph, the reduction grows from 11.8% to 33.3% between smooth and sharp Burgers windows, and from 12.7% to 24.6% between low- and high-Re NS, based on the displayed values. Global Response narrows the gap but remains above TERM, especially for sharp gradients and high-Re flows, where the relative activity of individual mechanisms changes most strongly.

A.4.4 CROSS-RESOLUTION AND IRREGULAR-QUERY PREDICTION

All models receive histories sampled on the 128×128 grids and are evaluated against solver references on the target grids or at 16,384 irregular query coordinates. Native-resolution solver fields are retained only to provide evaluation targets at the requested coordinates. TERM-Grid Decoder replaces the continuous coordinate-query decoder with a fixed-grid output head trained only on the 128×128 lattice. Predictions at 64×64, 256×256, and irregular query coordinates are obtained by deterministic bilinear interpolation from the predicted 128×128 field.

Table 11: Macro-average rel-L_2 across target resolutions and irregular queries. **Bold** denotes the lowest error for each query set and for the average.

Method	128→64	128→128	128→256	Irregular	Avg.
FNO	0.0528	0.0551	0.0714	0.0768	0.0640
Static Graph	0.0406	0.0427	0.0529	0.0564	0.0482
TERM-Grid Decoder	0.0358	0.0369	0.0498	0.0537	0.0441
TERM	**0.0324**	**0.0331**	**0.0395**	**0.0418**	**0.0367**

Table 11 shows that TERM obtains the lowest error for every query set. Increasing the target grid from 128×128 to 256×256 changes its error from 0.0331 to 0.0395, using solver-generated high-resolution targets. Compared with TERM-Grid Decoder, the continuous decoder reduces error by 20.7% at 256×256 and 22.2% on irregular coordinates; the macro-average decreases from 0.0441 to 0.0367. This pattern indicates that the query interface is most beneficial when evaluation points depart from the training grid.

A.4.5 Response evaluator robustness

We evaluate sensitivity to observation noise and response resolution on Burgers, NS, and zero-shot KdV, reporting their macro-average rel-L_2. Zero-mean Gaussian noise is added only to the observed history, with standard deviation measured relative to each variable's training-set standard deviation; target fields remain clean. TERM-Direct recomputes term responses at every rollout step, as in TERM, but applies the corresponding discrete differential operators directly to the predicted field, without the filtering, de-aliasing, or numerical stabilization used by the standard response evaluator. The resolution study changes only the spatial representation supplied to the response encoder. All learned modules, response recomputation frequency, and training objectives are otherwise unchanged.

Table 12: Response-estimation robustness. **Bold** denotes the lowest error at each noise level or response resolution.

(a) Input-noise standard deviation

Method	0%	0.5%	1%	2%	5%
Static Graph	0.0373	0.0380	0.0391	0.0415	0.0487
TERM-Direct	**0.0321**	0.0338	0.0366	0.0428	0.0597
TERM	0.0324	**0.0329**	**0.0338**	**0.0356**	**0.0419**

(b) Response resolution

Method	32	64	128
Static Graph	0.0373	0.0373	0.0373
Global Response	0.0364	0.0354	0.0349
TERM	**0.0346**	**0.0333**	**0.0324**

Without noise, TERM-Direct and TERM are similar. As noise increases, unregularized differentiation degrades sharply and becomes worse than Static Graph at 2%, whereas TERM remains below Static Graph through 5%. Finer response grids improve TERM, but even resolution 32 still outperforms both controls. The results suggest that approximate term-wise responses can remain useful when their numerical evaluation limits noise amplification.

A.4.6 Inference Efficiency

We evaluate the inference cost introduced by term-wise response modeling through a controlled comparison of Static Graph, TERM-Direct, and TERM-M. For this comparison, all three variants

Table 13: Inference efficiency of PDE foundation models and TERM response variants on 128×128 inputs.

Category	Method	Params.	Inference Speed ↑	Latency / Frame ↓	Response Eval. ↓
Published reference	DPOT-L	–	~ 135 FPS	~ 7.41 ms	–
Published reference	MoE-POT-M	288M	~ 148 FPS	~ 6.76 ms	–
Published reference	OmniArch-M	144M	~ 155 FPS	~ 6.45 ms	–
Published reference	Origo	97M	~ 1050 FPS	~ 0.95 ms	–
Controlled TERM study	Static Graph	–	~ 128 FPS	7.8 ± 0.2 ms	–
Controlled TERM study	TERM-Direct	105.4M	~ 110 FPS	9.1 ± 0.2 ms	0.36 ± 0.03 ms
Controlled TERM study	**TERM-M**	105.4M	~ 104 FPS	9.6 ± 0.3 ms	0.62 ± 0.04 ms

are measured on the same hardware and under the same inference protocol. We use batch size 1, a $K = 1$ history window, and 128×128 inputs on a single NVIDIA RTX 5090 GPU with the same mixed-precision setting. Each measurement uses 50 warm-up iterations followed by 500 timed test-sample forward passes, with CUDA synchronization before and after each timing interval. For TERM-M and TERM-Direct, numerical term-response evaluation is timed separately from the learned response encoder, resampler, term-local enrichment, and the remaining network. Response Eval. reports only the numerical computation of term-wise physical responses.

For broader context, Table 13 also reports published throughput values from Sun et al. (2026b). These values are reproduced from their Appendix E.2 and are not used for direct efficiency comparisons because the reported baselines use different hardware and implementation environments. The controlled TERM variants therefore provide the relevant comparison for quantifying the computational overhead of dynamic term-response modeling.

The controlled comparison quantifies the additional cost of dynamic term-response modeling. Relative to Static Graph, TERM-M increases single-step latency from 7.8 to 9.6 ms, corresponding to a 23.1% overhead. Numerical response evaluation accounts for 0.62 ms per step, while the remaining increase comes from response encoding, resampling, and term-local enrichment. TERM-Direct reduces response-evaluation time to 0.36 ms and total latency to 9.1 ms, indicating that numerical response construction contributes only part of the overall overhead. These results show that the state-dependent response pathway introduces a moderate additional inference cost relative to the static-graph variant. Published throughput values from other PDE foundation models provide scale context only and are not directly comparable to the controlled TERM measurements because their timing environments are not matched.