STRATATOK: LOCAL-VARIATION-GUIDED DYNAMIC TOKENIZATION FOR PDE FORECASTING FROM SPARSE AND IRREGULAR OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Observations of physical systems are often sparse and irregularly distributed across space and time. Forecasting PDE dynamics from such observations requires compact spatiotemporal tokens, but fixed tokenizers use the same predefined partition for every input, even when observation locations and locally active regions change. Existing dynamic tokenizers address this rigidity, but under sparse and irregular sampling, soft slicing leaves local variation implicit while patch-based refinement estimates it from unevenly populated or empty patches. To address this limitation, we present STRATATOK, a variation-stratified dynamic tokenizer. STRATATOK estimates local variation from each observation's visible spatiotemporal neighborhood and uses the resulting score to allocate its representation mass between global and variation-enriched pathways. Learned soft assignments then form mass-normalized tokens within each pathway. The resulting assignments admit a mass-conserving discrete transport interpretation. The two pathways form complementary token banks for global and localized dynamics. Across nine simulated PDE settings and four RealPDEBench systems, STRATATOK achieves the lowest RMSE in all thirteen comparisons among six tokenizers. The advantage persists across nearly all tested observation densities and sampling patterns, and STRATATOK ranks first at both full and halved token budgets in all four settings. Code is available at STRATATOK.

1 INTRODUCTION

Partial differential equations describe how physical systems evolve across space and time (Karniadakis et al., 2021; Kovachki et al., 2023). Forecasting their dynamics supports weather prediction (Lam et al., 2023; Bi et al., 2023), fluid simulation (Pfaff et al., 2021; Sanchez-Gonzalez et al., 2020), and climate projections (Nguyen et al., 2023; Watson-Parris et al., 2022). High-fidelity numerical solvers are accurate but costly when simulations must be repeated across initial conditions, geometries, or design parameters (Luo et al., 2025). This cost has driven the development of data-driven surrogate models for fast PDE prediction (Li et al., 2021; Lu et al., 2021).

Data-driven PDE surrogates learn mappings from physical inputs to solution fields and now achieve strong results across diverse systems (Lu et al., 2021; Li et al., 2021; Yin et al., 2023; Luo et al., 2025; Alkin et al., 2024). These surrogates predict solution fields directly at inference without a separate numerical solve for each new input (Li et al., 2021). Most methods, however, operate on complete fields sampled over dense grids or meshes (Lippe et al., 2023; Herde et al., 2024; Pfaff et al., 2021). These static discretizations fix sampling locations and neighborhood structure before representation learning begins (Li et al., 2021; Lippe et al., 2023; Herde et al., 2024).

Real-world observations of physical systems are often sparse and irregularly distributed across space and time (Lienen & Günnemann, 2022; Tong & Chen, 2026). Forecasting from these observations therefore calls for representations that operate on sparse locations (Lienen & Günnemann, 2022) while compressing multiscale field information (Wu et al., 2023). Fixed partitions are poorly matched because their regions differ in observation density and variation, while informative regions move as the physical state evolves (Ryoo et al., 2021; Zhang et al., 2026; Zhao et al., 2026). Dynamic tokenization offers a natural alternative because it adapts token formation to each input while producing a compact

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

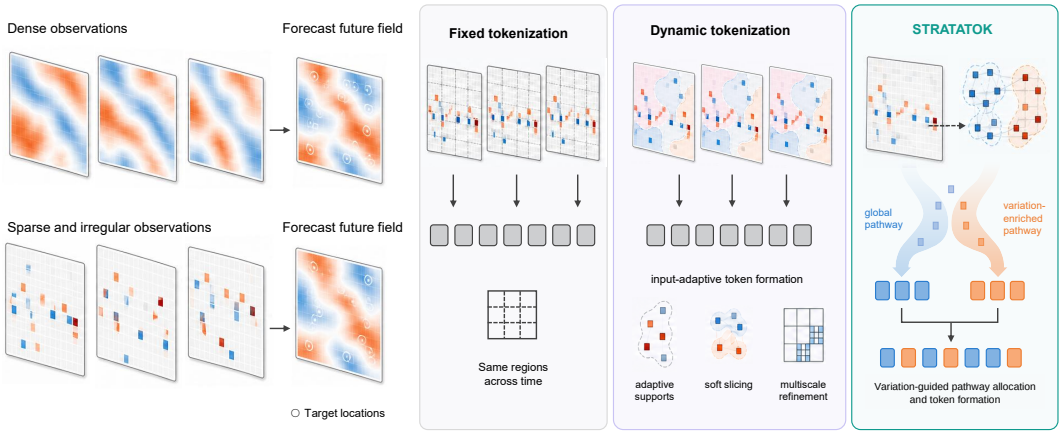


Figure 1: Schematic comparison of fixed, existing dynamic, and STRATATOK tokenization.

sequence for downstream processing (Ryoo et al., 2021; Luo et al., 2025). The central challenge is to infer this adaptation directly from the visible observations when large and irregular portions of the underlying field are unobserved.

Existing dynamic PDE tokenizers were developed mainly for fully observed discretizations. Transolver forms physics-aware tokens through learned soft assignments and mass-normalized aggregation (Wu et al., 2024). Transolver++ extends this construction with pointwise adaptive temperatures and Gumbel-softmax assignments (Luo et al., 2025). These learned slices adapt to encoded physical states, while variation across visible spatiotemporal neighborhoods does not explicitly govern observation-level allocation. MeshTok uses local activity to refine selected coarse patches (Zhao et al., 2026), while MATEY selects patch scales from local feature variance and combines coarse- and fine-scale tokens (Zhang et al., 2026). Both derive their refinement statistics from complete regular patches. Sparse and irregular observations leave such patches unevenly occupied or empty, making their activity and variance incomplete descriptions of the underlying local dynamics. Existing designs therefore provide either flexible soft aggregation without explicit local-variation guidance or explicit variation-guided refinement tied to fully observed patches. What remains missing is observation-level pathway allocation driven by variation estimated directly from visible spatiotemporal neighbors.

To address these limitations, we introduce STRATATOK, a variation-stratified dynamic tokenizer for forecasting PDE dynamics from sparse and irregular observations (Figure 1). Specifically, STRATATOK estimates local variation for each visible observation from its neighbors in joint space–time and uses this signal to divide its encoded representation between global and variation-enriched pathways. Learned soft assignments then distribute the representation across token slots within each pathway, followed by mass-normalized aggregation. The resulting nonnegative assignments admit a discrete transport interpretation: each observation supplies one unit of representation mass, and token representations are formed by mass-normalized pooling over received features. This design combines explicit variation guidance with flexible soft aggregation, producing complementary token representations of broad field structure and localized dynamics directly from sparse and irregular inputs.

Our contributions can be summarized as follows.

- We introduce STRATATOK, a variation-stratified dynamic tokenizer that forms compact spatiotemporal tokens directly from sparse and irregular observations without reconstructing the missing dense field.
- We introduce observation-level mass allocation between global and variation-enriched pathways, driven by variation estimated from visible spatiotemporal neighbors.
- Experiments across nine simulated PDE settings and four real-world systems show that STRATATOK achieves the strongest overall performance among six tokenizers. It remains strongest across nearly all tested observation densities and sampling patterns, and at reduced token budgets.

2 Related Work

Neural PDE solvers and geometric representations. DeepONet and the Fourier neural operator learn mappings between physical fields (Lu et al., 2021; Li et al., 2021; Kovachki et al., 2023), while graph simulators support particles and irregular meshes (Sanchez-Gonzalez et al., 2020; Pfaff et al., 2021; Brandstetter et al., 2022). GNOT, GINO, and CORAL further address irregular geometry through attention, latent grids, or neural fields (Hao et al., 2023; Li et al., 2023; Serrano et al., 2023). Latent Spectral Models, Universal Physics Transformers, and Poseidon learn compact or pretrained dynamics (Wu et al., 2023; Alkin et al., 2024; Herde et al., 2024). These methods primarily develop operator architectures or latent dynamics. Our focus is complementary: constructing compact tokens when the available input is itself sparse and irregular. RealPDEBench (Hu et al., 2026) extends PDE surrogate evaluation to paired experimental and numerical systems. We use four of its measured systems to test sparse tokenization on PIV velocity, combustion imaging, and controlled and uncontrolled cylinder wakes.

Forecasting from sparse and irregular observations. Finite Element Networks learn dynamics from arbitrary sparse locations, while SHRED forecasts from limited sensor histories (Lienen & Günnemann, 2022; Williams et al., 2024). Perceiver and Perceiver IO aggregate unstructured or sparse inputs through latent cross-attention (Jaegle et al., 2021; 2022). DINo couples coordinate-conditioned neural fields with latent dynamics, while continuous latent neural PDEs learn evolution from noisy and partially observed states (Yin et al., 2023; Iakovlev et al., 2023). CORAL removes fixed-mesh constraints through neural fields, and STRIDE couples sparse sensor histories with coordinate decoding (Serrano et al., 2023; Tong & Chen, 2026). These methods address irregular sampling, latent evolution, and flexible output queries. Their compression typically relies on latent queries or implicit codes rather than explicit local variation to organize complementary token representations. StrataTok targets this step directly.

Learned dynamic tokenization and latent compression. Learned tokenization reduces the number of representations passed to expensive downstream layers. TokenLearner extracts input-dependent tokens from dense features (Ryoo et al., 2021). Latent Spectral Models compress physical fields, Latent Neural Operators use cross-attention to enter latent space, and Universal Physics Transformers aggregate inputs into supernodes (Wu et al., 2023; Wang & Wang, 2024; Alkin et al., 2024). Transolver forms physics-aware tokens using learned pointwise soft assignments and mass-normalized aggregation (Wu et al., 2024). Transolver++ augments this construction with pointwise adaptive temperatures and Gumbel-softmax reparameterization (Luo et al., 2025). It provides our closest dynamic-tokenization comparator because its locally adaptive slice construction can be evaluated under the same sparse-observation interface and forecasting architecture. Within each pathway, StrataTok uses learned soft token formation. Its defining component is observation-level mass allocation between global and variation-enriched pathways, driven by variation estimated from visible neighbors.

Adaptive and multiscale physical representations. Learned mesh movers and DeepLag adapt points or particles to nonuniform physics (Hu et al., 2024; Ma et al., 2024). MATEY adapts patch sizes (Zhang et al., 2026), while MeshTok and AMR-Transformer use dense-field indicators for multiscale tokenization or mesh refinement (Zhao et al., 2026; Xu et al., 2025). StrataTok instead estimates local variation from visible irregular observations in space and time and forms a fixed-length sequence without dense reconstruction.

3 Method

StrataTok constructs input-adaptive spatiotemporal tokens directly from sparse and irregular observations. It estimates local variation from visible space–time neighbors and uses the score to divide each observation between two complementary representation pathways. Learned soft assignments aggregate the representations into token slots within each pathway, followed by mass-normalized pooling in every feature head. This design separates variation-guided pathway allocation from slot selection and provides a compact interface for the forecasting model. The resulting assignments also admit a discrete transport interpretation (Peyré & Cuturi, 2019). Figure 2 summarizes the complete pipeline.

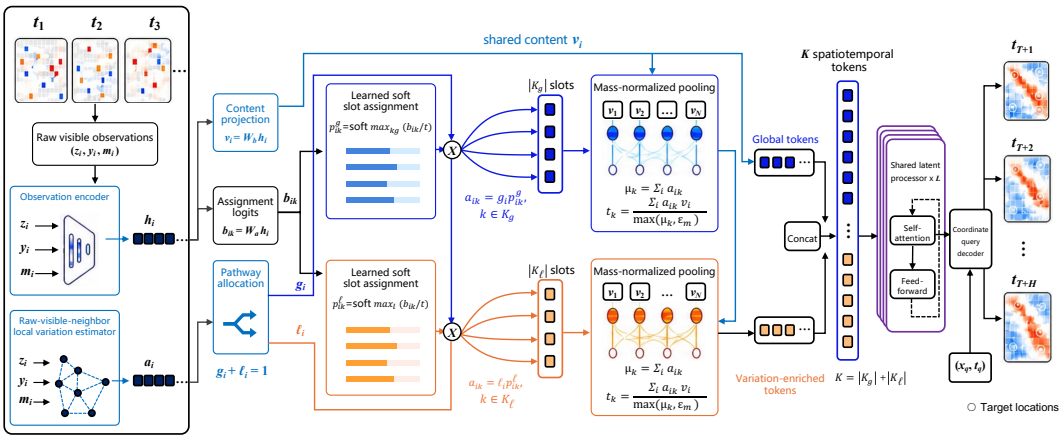


Figure 2: Overview of StrataTok, which uses local variation from visible observations to guide pathway allocation and mass-normalized spatiotemporal token formation for a shared forecaster.

3.1 Problem Formulation

We consider forecasting spatiotemporal physical fields from sparse and irregular observations. Let $\Omega \subset \mathbb{R}^{d_x}$ denote the spatial domain and $\mathcal{T}_{\text{in}}$ the observed time window. For one sample, the input is the finite set $\mathcal{E} = \{(\mathbf{z}_i, \mathbf{y}_i, \mathbf{m}_i)\}_{i=1}^{N}$ of visible observations. Here $\mathbf{z}_i = (\mathbf{x}_i, t_i) \in \Omega \times \mathcal{T}_{\text{in}}$ is a space–time location, $\mathbf{y}_i \in \mathbb{R}^{d_y}$ contains the observed physical quantities, and $\mathbf{m}_i \in \{0,1\}^{d_y}$ records channel availability. The number, locations, and available channels of the observations may vary across input times and samples. We write $\mathcal{V}$ for the valid observation indices and $N_v = |\mathcal{V}|$. In the computations below, $\mathbf{z}_i$ denotes the joint coordinate after robust per-sample normalization, and $\mathbf{y}_i$ denotes the physical quantities after channelwise normalization using training-split statistics. Given specified future target locations $\mathcal{Z}^q = \{(\mathbf{x}_j^q, t_j^q)\}_{j=1}^{N_q}$, the model is written as

$$\mathbf{T} = \mathcal{T}_\theta(\mathcal{E}), \qquad \widehat{\mathbf{Y}}^q = \mathcal{D}_\omega(\mathcal{P}_\psi(\mathbf{T}), \mathcal{Z}^q), \tag{1}$$

where $\mathcal{T}_\theta$ maps the variable-size observation set to K tokens $\mathbf{T} \in \mathbb{R}^{K \times D}$, $\mathcal{P}_\psi$ processes these tokens, and $\mathcal{D}_\omega$ predicts the physical states at the specified future target locations.

3.2 Local Variation from Sparse and Irregular Observations

The visible observations do not provide the regular stencils assumed by dense patch refinement. We therefore estimate variation around each observation in normalized joint space–time. Let $\mathcal{N}_\kappa(i)$ contain its κ nearest visible neighbors and let $\mathcal{C}_{ij}$ contain the physical channels visible at both i and j. We retain only comparable neighbors $\mathcal{N}_\kappa^+(i) = \{j \in \mathcal{N}_\kappa(i) : |\mathcal{C}_{ij}| > 0\}$. For each such neighbor, the channel-averaged physical difference is

$$\Delta_{ij} = \sqrt{\frac{1}{|\mathcal{C}_{ij}|} \sum_{c \in \mathcal{C}_{ij}} (y_{ic} - y_{jc})^2}. \tag{2}$$

We summarize the change per unit space–time distance and its amplitude using robust neighborhood medians,

$$s_i = \underset{j \in \mathcal{N}_\kappa^+(i)}{\text{median}} \frac{\Delta_{ij}}{\max(\|\mathbf{z}_i - \mathbf{z}_j\|_2, \epsilon)}, \qquad d_i = \underset{j \in \mathcal{N}_\kappa^+(i)}{\text{median}} \Delta_{ij}. \tag{3}$$

Their sample-normalized combination defines the local-variation score,

$$a_i = (1-\beta)\frac{s_i}{\max(\bar{s}, \epsilon)} + \beta \frac{d_i}{\max(\bar{d}, \epsilon)}. \tag{4}$$

Here, $\bar{s}$ and $\bar{d}$ denote the means of s_i and d_i, respectively, over valid observations, and $\epsilon = 10^{-6}$ is a numerical stabilizer. If $\mathcal{N}_\kappa^+(i)$ is empty, we set $s_i = d_i = 0$. The local slope term s_i measures

change per unit space–time distance, while the local amplitude term d_i records the size of the local change. Setting $\beta = 0$ gives the slope-only estimator. Values $\beta \in (0, 1]$ incorporate local spread for fields whose scale carries useful information. Both statistics use only visible input values. The construction is related to monitor functions in adaptive meshes (Huang & Russell, 2011), but it is defined directly on sparse and irregular observations.

3.3 Variation-guided pathway allocation

The score controls the allocation of each observation between a global pathway and a variation-enriched pathway. We first normalize and reshape the score within each sample,

$$q_i = \mathrm{clip}\left(\frac{a_i}{\max(\overline{a}, \epsilon)}, 0, \rho\right), \qquad r_i = \lambda q_i^{\eta}. \tag{5}$$

Here, $\overline{a}$ denotes the mean local-variation score over valid observations. When $\overline{s} \geq \epsilon$ and $\overline{d} \geq \epsilon$, Equation 4 gives $\overline{a} = 1$ by construction. The second normalization in Equation 5 is retained for near-degenerate samples in which either statistic falls below the numerical stabilizer. The resulting pathway weights are

$$g_i = \frac{1}{1 + r_i}, \qquad \ell_i = \frac{r_i}{1 + r_i}. \tag{6}$$

The weights satisfy $g_i + \ell_i = 1$. Observations with little local variation place more weight on the global pathway, while larger scores direct more weight to the variation-enriched pathway. We fix the clipping cap at $\rho = 4$. The positive scale λ, positive contrast η, neighborhood size, and pathway sizes are selected on validation data.

3.4 Mass-normalized token formation

Within each pathway, learned soft assignment and mass-normalized pooling follow the physical-slice construction of Transolver (Wu et al., 2024). StrataTok precedes this operation with observation-level mass allocation between the global and variation-enriched pathways.

Let $\mathcal{K}_g$ and $\mathcal{K}_\ell$ be disjoint sets of global and variation-enriched slots with $|\mathcal{K}_g| + |\mathcal{K}_\ell| = K$. These slots serve as learned latent destinations whose spatial support is determined by the input-dependent assignments. An observation encoder maps $(\mathbf{z}_i, \mathbf{y}_i, \mathbf{m}_i)$ to a feature $\mathbf{h}_i \in \mathbb{R}^D$. The implementation applies the following construction independently in each attention head, and we suppress the head index for clarity. Linear projections of $\mathbf{h}_i$ produce a head-specific content vector $\mathbf{v}_i$ and slot logits b_{ik}. Let $\tau > 0$ denote the learned head-specific temperature that controls assignment sharpness. Within pathway γ, a softmax converts the slot logits into a distribution,

$$p_{ik}^{\gamma} = \frac{\exp(b_{ik}/\tau)}{\sum_{u \in \mathcal{K}_\gamma} \exp(b_{iu}/\tau)}, \qquad \gamma \in \{g, \ell\}. \tag{7}$$

Write $w_i^g = g_i$ and $w_i^\ell = \ell_i$. The assignment weight from observation i to slot k is

$$\alpha_{ik} = w_i^{\gamma} p_{ik}^{\gamma}, \qquad k \in \mathcal{K}_\gamma, \quad \gamma \in \{g, \ell\}. \tag{8}$$

Each slot accumulates mass μ_k and forms a normalized pooled representation,

$$\mu_k = \sum_{i \in \mathcal{V}} \alpha_{ik}, \qquad \widetilde{\mathbf{t}}_k = \frac{\sum_{i \in \mathcal{V}} \alpha_{ik} \mathbf{v}_i}{\max(\mu_k, \epsilon_m)}. \tag{9}$$

Here, $\epsilon_m > 0$ is a numerical stabilizer. Local variation determines how much weight enters each pathway through g_i and ℓ_i. The learned distributions p_{ik}^g and p_{ik}^ℓ then select slots within the corresponding pathway. This ordering allows local variation to organize token formation without fixing the spatial support of any token.

3.5 Transport interpretation and structural properties

Equations 8 and 9 give a discrete transport interpretation. Each observation is a source atom with one unit of representation mass, and each token slot is a learned destination. Here, representation mass refers to the nonnegative allocation weight α_{ik}. The matrix $\boldsymbol{\pi} = \boldsymbol{\alpha}/N_v$ transports the uniform

source measure over visible observations to the slot masses μ_k/N_v. The coupling factors into an observation-level pathway allocation followed by a within-pathway slot assignment. When $\mu_k \geq \epsilon_m$, $\widetilde{\mathbf{t}}_k$ is the barycentric projection of the features received by slot k; otherwise the stabilizer contracts that representation toward the origin. Feed-forward assignments define this mass-conserving coupling and are trained end to end through the forecasting objective.

The construction has three structural properties. Each valid observation conserves its unit mass, the global pathway retains a positive aggregate mass, and every per-head slot representation obeys a convex-hull bound:

$$\sum_{k=1}^{K} \alpha_{ik} = 1, \qquad M^g := \sum_{i \in \mathcal{V}} g_i \geq \frac{N_v}{1 + \lambda \rho^{\eta}}, \qquad \widetilde{\mathbf{t}}_k \in \operatorname{conv}(\{\mathbf{0}\} \cup \{\mathbf{v}_i : \alpha_{ik} > 0\}). \tag{10}$$

The unit-mass identity and convex-hull containment follow from row-normalized assignment and mass-normalized pooling. The bounded dual-path allocation preserves a positive global contribution and orders the two pathways according to variation measured in visible neighborhoods. The second inequality gives the explicit lower bound under the bounded variation score. For $\lambda = \eta = 1$, a tighter Jensen bound gives an aggregate global mass of at least $N_v/2$. Since ℓ_i increases with local variation while g_i decreases, the variation-enriched pathway receives an assignment-weighted mean score no lower than the global pathway. Finally, each per-head pooled representation $\widetilde{\mathbf{t}}_k$, before head concatenation, dropout, and layer normalization, lies in the convex hull of the origin and its received content vectors. Its scale therefore cannot grow merely because more observations are assigned to the slot. When $\mu_k \geq \epsilon_m$, the origin has zero weight and the representation is their normalized weighted average. Proofs are given in Appendix B.3.

The per-head pooled representations are concatenated, passed through dropout and layer normalization to form the final tokens $\mathbf{t}_k$, and slots with $\mu_k \leq \epsilon_m$ are masked. These tokens then enter the shared latent processor and coordinate decoder from Equation 1. The complete model is trained end-to-end with forecast error and token-usage regularization. Additional implementation details and dataset-specific configurations are provided in Appendix B.2.

4 Experiments

We evaluate StrataTok across simulated and real-world physical systems, sparse and irregular observation processes, and different token budgets. The evaluation includes (1) matched comparisons with five tokenization baselines across nine simulated PDE settings and four RealPDEBench systems; (2) robustness analyses across visible-observation ratios and spatiotemporal sampling patterns; (3) controlled routing ablations; and (4) comparisons under matched and reduced token budgets.

4.1 Experimental setup

Datasets. Our simulated data comprise two independently sampled radial-dam-break SWE sets generated with the PDEBench solver (Takamoto et al., 2022); compressible CFD-rand, Euler CRP, Burgers 1D, and Advection 1D from PDEBench; spherical SWE and Navier–Stokes from PDEArena (Gupta & Brandstetter, 2023); and 2D Burgers from PDENNEval (Wei et al., 2024). RealPDEBench (Hu et al., 2026) provides real-world measurements of fluid–structure interaction, combustion, cylinder wake, and controlled cylinder wake. The full collection spans one- and two-dimensional domains, planar and spherical geometries, scalar and multichannel fields, PIV velocity, and OH^* chemiluminescence.

Baselines. We compare StrataTok with five tokenization methods under a shared forecasting architecture. FixedPower and AdaptivePower construct tokens from coordinate-based supports. Transolver++ (Luo et al., 2025) forms locally adaptive soft physical slices using pointwise assignment temperatures and Gumbel-softmax assignment. MeshTok (Zhao et al., 2026) and MATEY (Zhang et al., 2026) construct tokens through adaptive multiscale refinement, with MATEY using its AdapMix construction. For a controlled comparison, every method receives exactly the same sparse and irregular observations and uses the same observation encoder, latent processor, coordinate decoder, model capacity, and training budget. This work focuses on tokenization mechanisms. Results for the end-to-end forecasting baselines DINo (Yin et al., 2023) and STRIDE (Tong & Chen, 2026) are reported in Appendix A4.

Implementation. Each task takes sparse, irregularly sampled observations as input and predicts states at specified target locations over consecutive future time steps. Within each setting, all methods share the visible inputs, backbone, model capacity, optimizer, and training budget. Shared and method-specific hyperparameters are selected on validation data. We report geometric-mean RMSE over all predicted frames. Each run uses a single NVIDIA RTX A5000 GPU. Complete configurations, data splits, and evaluation details are provided in Appendix B.

4.2 Main results

Table 1: Performance across simulated PDE settings and RealPDEBench systems with sparse and irregular observations. We report geometric-mean RMSE, where lower is better. Percentages show StrataTok's relative RMSE reduction over each baseline. The best result is bold with a dark blue background, and the second-best result has a light blue background.

Dataset	FixedPower	AdaptivePower	Transolver++	MeshTok	MATEY	StrataTok
Radial SWE A	0.2709↓ 41.08%	0.2241↓ 28.76%	0.2836↓ 43.70%	0.2025↓ 21.18%	0.2785↓ 42.68%	**0.1596**
Radial SWE B	0.2495↓ 32.59%	0.2167↓ 22.39%	0.3705↓ 54.60%	0.1968↓ 14.55%	0.1996↓ 15.75%	**0.1682**
PDEArena spherical SWE	0.6332↓ 0.23%	0.6339↓ 0.34%	0.6323↓ 0.10%	0.6343↓ 0.40%	0.6392↓ 1.17%	**0.6317**
CFD-rand	0.3398↓ 35.77%	0.2925↓ 25.37%	0.2501↓ 12.73%	0.2993↓ 27.08%	0.2905↓ 24.86%	**0.2183**
Euler CRP	0.6866↓ 7.13%	0.6607↓ 3.48%	0.6378↓ 0.01%	0.6713↓ 5.01%	0.6720↓ 5.10%	**0.6377**
Burgers 1D	0.2872↓ 9.34%	0.2719↓ 4.24%	0.2724↓ 4.39%	0.2845↓ 8.47%	0.2947↓ 11.63%	**0.2604**
PDENNEval 2D Burgers	0.1783↓ 84.61%	0.0497↓ 44.74%	0.0590↓ 53.49%	0.1707↓ 83.91%	0.1801↓ 84.76%	**0.0275**
Advection 1D	0.2776↓ 24.46%	0.2312↓ 9.31%	0.2146↓ 2.28%	0.2565↓ 18.25%	0.2542↓ 17.49%	**0.2097**
PDEArena Navier–Stokes	0.5046↓ 4.14%	0.4973↓ 2.74%	0.5251↓ 7.88%	0.4900↓ 1.29%	0.5021↓ 3.68%	**0.4837**
Fluid–structure interaction	0.5721↓ 49.08%	0.6157↓ 52.69%	0.3114↓ 6.47%	0.3816↓ 23.67%	0.3833↓ 24.01%	**0.2913**
Combustion imaging	0.8105↓ 18.14%	0.8181↓ 18.90%	0.7554↓ 12.16%	0.7662↓ 13.41%	0.7599↓ 12.68%	**0.6635**
Cylinder wake	0.4091↓ 35.84%	0.4043↓ 35.09%	0.2716↓ 3.38%	0.3289↓ 20.19%	0.3200↓ 17.98%	**0.2624**
Controlled cylinder wake	0.3990↓ 36.36%	0.3702↓ 31.42%	0.2632↓ 3.53%	0.3281↓ 22.62%	0.3391↓ 25.12%	**0.2539**

We first compare StrataTok with five alternative tokenizers across nine simulated PDE settings and four RealPDEBench systems. The baselines cover coordinate-based supports, the locally adaptive soft slicing of Transolver++, and the adaptive multiscale refinement used by MeshTok and MATEY. As shown in Table 1, StrataTok attains the lowest geometric-mean RMSE in all thirteen settings. Relative to Transolver++, its geometric mean is lower in every setting, with near-ties on spherical SWE (0.10%) and Euler CRP (0.01%). The reductions reach 54.60% on Radial SWE B, 53.49% on PDENNEval 2D Burgers, 43.70% on Radial SWE A, and 12.73% on CFD-rand. The larger gains span distinct dynamics rather than a single benchmark family.

We next examine whether StrataTok extends from simulated PDE trajectories to real-world measurements. The four rows below the dashed separator in Table 1 report the RealPDEBench systems. StrataTok ranks first on fluid–structure interaction, combustion imaging, cylinder wake, and controlled cylinder wake. Relative to the strongest comparator on each system, it reduces RMSE by 6.47%, 12.16%, 3.38%, and 3.53%, respectively. Its margins over MeshTok and MATEY range from 12.68% to 25.12%. These rankings span PIV velocity, reacting-flow imagery, and controlled and uncontrolled wake dynamics, extending the evidence for variation-stratified tokenization to sparse and irregular measurements from real physical systems.

The design combines visible-neighbor variation estimates, conservative pathway allocation, and mass-normalized pooling. The following experiments test changes in observation density and sampling, routing alternatives, and reduced token counts.

4.3 Robustness to observation density and sampling

We next evaluate StrataTok as the proportion and spatiotemporal distribution of visible observations change. Figure 3a–d reports RMSE as visibility ranges from 0.3125% to 20%. StrataTok maintains the lowest RMSE throughout the sweeps on Radial SWE A, PDENNEval 2D Burgers,

and RealPDEBench combustion, with reductions over the strongest competitor reaching 24.86%, 84.34%, and 14.59%, respectively. On CFD-rand, MATEY is lower at the two most sparse levels; from 1.25% visibility onward, STRATATOK ranks first and reaches a maximum reduction of 26.21%. The results show that the advantage persists over a broad range of observation densities, with the boundary appearing only under the most extreme CFD sparsity.

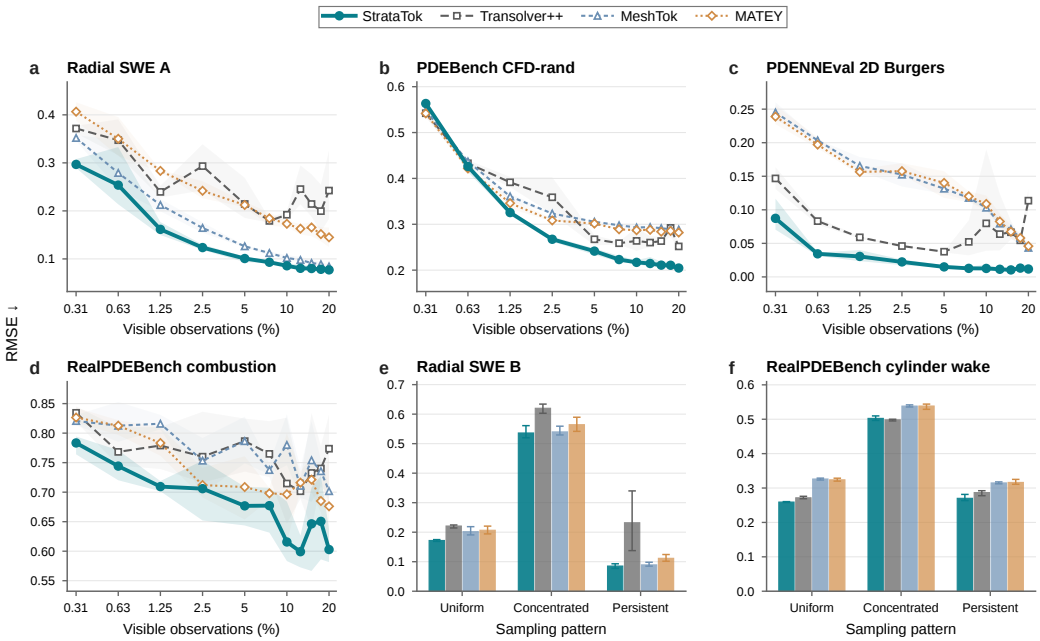


Figure 3: Robustness to observation density and sampling. **a–d,** RMSE across visible-observation ratios. **e–f,** RMSE under uniform, concentrated, and persistent sampling on Radial SWE B at 1.25% visibility and RealPDEBench cylinder wake at approximately 3% visibility. Curves and bars show geometric-mean RMSE across runs; shaded bands and error bars span the corresponding ranges. Lower is better.

Figure 3e–f varies the sampling pattern. Spatially concentrated observations substantially increase the error, especially on Radial SWE B. On that system, STRATATOK ranks first under all three patterns and reduces RMSE by 0.67–14.80% relative to the strongest competitor. On RealPDEBench cylinder wake, it ranks first under uniform and persistent sampling and remains within 0.86% of Transolver++ under concentrated sampling. The two sweeps probe complementary changes: the visibility study reduces the amount of observed information, whereas the sampling-pattern study changes where that information is available. Because local variation is recomputed from each visible input, pathway allocation adapts to both changes without requiring a sampling-specific partition.

4.4 ABLATION STUDY

Table 2 compares six routing constructions with the remaining forecasting model fixed. The single-bank control uses Transolver's learned soft assignment and mass-normalized pooling (Wu et al., 2024); its local-variation variant additionally conditions assignment logits on the same visible-neighborhood score. Fixed two-bank and sample-level controls respectively use constant observation shares and one input-level share. The shuffled control permutes observation-level shares, while STRATATOK preserves their alignment with local-variation scores.

STRATATOK achieves the lowest RMSE on all three tasks. The local-variation single-bank control remains above STRATATOK (0.1836 versus 0.1596, 0.2395 versus 0.2183, and 0.7546 versus 0.6635), isolating the gain from using local variation for observation-level pathway allocation. Fixed, sample-

Table 2: Routing ablation on representative continuous multi-frame forecasting tasks.

Tokenizer construction	Radial SWE A	CFD-rand	RealPDEBench combustion
Single learned bank	0.1779	0.2413	0.7188
Single bank + local variation	0.1836	0.2395	0.7546
Two banks, fixed allocation	0.1860	0.2413	0.7573
Sample-level allocation	0.1802	0.2439	0.7414
Shuffled observation-level allocation	0.2007	0.2382	0.7560
STRATATOK (aligned)	**0.1596**	**0.2183**	**0.6635**

level, and shuffled allocation also remain above STRATATOK, supporting the aligned dual-path construction.

4.5 TOKEN BUDGET

Table 3: Performance across token budgets on four sparse and irregular PDE forecasting settings. Values are geometric-mean RMSE, where lower is better. Dark and light blue mark the best and second-best results.

Tokenizer	**Radial SWE A**		**CFD-rand**		**RealPDEBench combustion**		**PDENNEval 2D Burgers**	
	$K = 16$	$K = 32$	$K = 24$	$K = 48$	$K = 96$	$K = 192$	$K = 16$	$K = 32$
FixedPower	0.2645	0.2709	0.3640	0.3398	0.8213	0.8105	0.1741	0.1783
AdaptivePower	0.2302	0.2241	0.3207	0.2925	0.8196	0.8181	0.0524	0.0497
Transolver++	0.2812	0.2836	0.2748	0.2501	0.7750	0.7554	0.0650	0.0590
MeshTok	0.2368	0.2025	0.4256	0.2993	0.7230	0.7662	0.2018	0.1707
MATEY	0.2825	0.2785	0.4227	0.2905	0.7611	0.7599	0.1969	0.1801
STRATATOK	**0.1615**	**0.1596**	**0.2402**	**0.2183**	**0.6904**	**0.6635**	**0.0268**	**0.0275**

Table 3 compares all six tokenizers at two token counts, and STRATATOK ranks first at both budgets in every setting. Relative to the strongest same-budget comparator, it reduces RMSE by 12.16–44.74% at the full token count and by 4.50–48.94% after halving the count. The half-budget STRATATOK also surpasses the strongest full-budget baseline in all four settings, with reductions of 20.24% on Radial SWE A, 3.98% on CFD-rand, 8.60% on RealPDEBench combustion, and 46.16% on PDENNEval 2D Burgers. Overall, variation-stratified token formation uses token capacity more effectively. It lowers forecasting error at matched budgets and retains comparable or better accuracy than the strongest full-budget baseline with half as many token slots. The largest full-budget reduction, 44.74%, occurs on PDENNEval 2D Burgers at $K = 32$. On that task, the two STRATATOK budgets are nearly tied: RMSE is 0.0268 at $K = 16$ and 0.0275 at $K = 32$. The other three settings favor the full token count.

5 CONCLUSION

We introduced STRATATOK, a variation-stratified dynamic tokenizer for PDE forecasting from sparse and irregular observations. It uses visible-neighbor variation to allocate observations across global and variation-enriched pathways before learned assignment and mass-normalized pooling. The resulting assignments admit a discrete transport interpretation with mass conservation and a per-head convex-hull bound. Across nine simulated settings and four RealPDEBench systems, STRATATOK ranks first among six tokenizers. Robustness and routing experiments further support local-variation-guided allocation for compact sparse-PDE representations.

REPRODUCIBILITY STATEMENT

Source code and experiment configurations are publicly available at STRATATOK.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript revision and the review of experimental design and results.

REFERENCES

Benedikt Alkin, Andreas Fürst, Simon Schmid, Lukas Gruber, Markus Holzleitner, and Johannes Brandstetter. Universal physics transformers: A framework for efficiently scaling neural operators. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619:533–538, 2023.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations*, 2022.

Jayesh K. Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized PDE modeling. *Transactions on Machine Learning Research*, 2023.

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.

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, pp. 72525–72624, 2024.

Peiyan Hu, Yue Wang, and Zhi-Ming Ma. Better neural PDE solvers through data-free mesh movers. In *International Conference on Learning Representations*, pp. 4550–4576, 2024.

Peiyan Hu, Haodong Feng, Hongyuan Liu, Tongtong Yan, Wenhao Deng, Tianrun Gao, Rong Zheng, Haoren Zheng, Chenglei Yu, Chuanrui Wang, Kaiwen Li, Zhi-Ming Ma, Dezhi Zhou, Xingcai Lu, Dixia Fan, and Tailin Wu. RealPDEBench: A benchmark for complex physical systems with real-world data. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=y3oHMcoItR. Oral presentation.

Weizhang Huang and Robert D. Russell. *Adaptive Moving Mesh Methods*, volume 174 of *Applied Mathematical Sciences*. Springer New York, 2011. doi: 10.1007/978-1-4419-7916-2.

Valerii Iakovlev, Markus Heinonen, and Harri Lähdesmäki. Learning space-time continuous latent neural PDEs from partially observed states. In *Advances in Neural Information Processing Systems*, volume 36, pp. 26372–26395, 2023.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021.

Nikola B. Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew M. 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.

Rémi Lam, Álvaro Sánchez-González, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421, 2023.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. 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.

Marten Lienen and Stephan Günnemann. Learning the dynamics of physical systems from sparse observations with finite element networks. In *International Conference on Learning Representations*, 2022.

Phillip Lippe, Bas Veeling, Paris Perdikaris, Richard Turner, and Johannes Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67398–67433, 2023.

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*, 3:218–229, 2021.

Huakun Luo, Haixu Wu, Hang Zhou, Lanxiang Xing, Yichen Di, Jianmin Wang, and Mingsheng Long. Transolver++: An accurate neural solver for PDEs on million-scale geometries. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41432–41449. PMLR, 2025.

Qilong Ma, Haixu Wu, Lanxiang Xing, Shangchen Miao, and Mingsheng Long. DeepLag: Discovering deep lagrangian dynamics for intuitive fluid prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25904–25938. PMLR, 2023.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021.

Michael S. Ryoo, A. J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. TokenLearner: Adaptive space-time tokenization for videos. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12786–12797, 2021.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8459–8468. PMLR, 2020.

Louis Serrano, Lise Le Boudec, Armand Kassaï Koupaï, Thomas X. Wang, Yuan Yin, Jean-Noël Vittaut, and Patrick Gallinari. Operator learning with neural fields: Tackling PDEs on general geometries. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1596–1611, 2022.

Yanjie Tong and Peng Chen. From sparse sensors to continuous fields: STRIDE for spatiotemporal reconstruction. *arXiv preprint arXiv:2602.04201*, 2026.

Tian Wang and Chuang Wang. Latent neural operator for solving forward and inverse PDE problems. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Duncan Watson-Parris, Yuhan Rao, Dirk Olivié, Øyvind Seland, Peer Nowack, Gustau Camps-Valls, Philip Stier, Shahine Bouabid, Maura Dewey, Emilie Fons, Jessenia Gonzalez, Paula Harder, Kai Jeggle, Julien Lenhardt, Peter Manshausen, Maria Novitasari, Lucile Ricard, and Carla Roesch. ClimateBench v1.0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14(10):e2021MS002954, 2022.

Ping Wei, Menghan Liu, Jianhuan Cen, Ziyang Zhou, Liao Chen, and Qingsong Zou. PDENNEval: A comprehensive evaluation of neural network methods for solving PDEs. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5181–5189, 2024. doi: 10.24963/ijcai.2024/573.

Jan P. Williams, Olivia Zahn, and J. Nathan Kutz. Sensing with shallow recurrent decoder networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 480(2298): 20240054, 2024. doi: 10.1098/rspa.2024.0054.

Haixu Wu, Tengge Hu, Huakun Luo, Jianmin Wang, and Mingsheng Long. Solving high-dimensional PDEs with latent spectral models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37417–37438. PMLR, 2023.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for PDEs on general geometries. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53681–53705. PMLR, 2024.

Zeyi Xu, Jinfan Liu, Kuangxu Chen, Ye Chen, Zhangli Hu, and Bingbing Ni. AMR-Transformer: Enabling efficient long-range interaction for complex neural fluid simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5804–5813, 2025.

Yuan Yin, Matthieu Kirchmeyer, Jean-Yves Franceschi, Alain Rakotomamonjy, and Patrick Gallinari. Continuous PDE dynamics forecasting with implicit neural representations. In *International Conference on Learning Representations*, 2023.

Pei Zhang, M. Paul Laiu, Matthew Norman, Doug Stefanski, and John Gounley. MATEY: Multiscale adaptive transformer models for spatiotemporal physical systems. *Machine Learning: Science and Technology*, 7(3):035016, 2026. doi: 10.1088/2632-2153/ae67d1.

Yanshun Zhao, Xiaoyu Peng, Jiamin Jiang, Congcong Zhu, and Jingrun Chen. MeshTok: Efficient multi-scale tokenization for scalable PDE transformers. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

A Outlook

The neighborhood metric is Euclidean after coordinatewise median and interquartile-range normalization, which removes marginal scale differences. Physics-scaled or learned metrics could incorporate space–time anisotropy and characteristic propagation speeds when spatial and temporal sampling scales differ substantially.

The evaluation covers several simulated equation families and four RealPDEBench systems with real-world measurements, spanning PIV velocity, combustion imaging, and controlled and uncontrolled cylinder wakes. Future evaluations can extend this coverage to other sensing modalities and physical systems.

B Reproducibility details

B.1 Data and split construction

Radial SWE A and B are independent samples from the same radial-dam-break distribution, generated with the pinned PDEBench `RadialDamBreak2D` finite-volume solver (Takamoto et al., 2022). Each set contains 240 unique trajectories on a 128×128 grid, with 21 states on $t \in [0, 1]$. The square domain is $[-2.5, 2.5]^2$, gravity and outer depth are one, and initial momentum is zero. For each trajectory, the dam center is sampled uniformly from $[-0.2, 0.2]^2$, its radius from $[0.30, 0.65]$, and its inner depth from $[1.6, 2.4]$. Sets A and B are independent fixed realizations, each followed by a 160/40/40 trajectory split. The observation-density study uses set A; the sampling-pattern study uses set B, with one persistent irregular sensor layout shared across context frames.

PDEArena spherical SWE uses 200/50/50 train/validation/test trajectories and three field channels. From five context steps, each model jointly predicts 16 consecutive future frames using 1,024 visible events over the context window and 512 target locations per predicted frame. Latitude and longitude are embedded as spherical Cartesian coordinates. CFD-rand uses 800/200/200 trajectories and the density, pressure, and x-velocity channels on a one-dimensional spatial coordinate. From five context steps, each model jointly predicts 36 consecutive future frames using 512 visible events over the context window and 512 target locations per predicted frame.

Our RealPDEBench evaluation uses 51 FSI, 30 combustion, 91 cylinder-wake, and 96 controlled cylinder-wake source trajectories (Hu et al., 2026). Complete source trajectories are split before extracting eight 21-frame forecasting windows. The source-level train/validation/test splits are 35/7/9, 21/4/5, 62/10/19, and 66/10/20, respectively. For FSI, we retain the first 2,173 frames and subsample the 128×128 PIV fields to 64×64. Combustion uses OH^* chemiluminescence intensity; the other systems use measured velocity fields. Every tokenizer observes ten context frames at approximately 3% visibility, forms 192 tokens, and predicts the next eleven frames jointly. This gives 1,229 visible events for FSI and combustion and 2,458 for the two wake systems. Normalization uses training trajectories only.

The main comparison aggregates three matched training runs in every setting. Every method uses the same run identifiers within a row. Frozen split manifests, file hashes, and resolved configurations accompany the code release.

B.2 Model and optimization configurations

Joint space–time coordinates are centered by their coordinatewise median and scaled by their interquartile range over the visible observations. Degenerate scales are replaced by one, and target coordinates use the same transformation. The observation encoder receives dyadic Fourier features of the normalized coordinates, zero-filled visible values, and the corresponding channel mask. It consists of two linear layers with a GELU activation.

The tokenizer uses multihead content and assignment projections. Each head forms content vectors and slot logits in its own feature subspace. Learned temperatures are initialized to 0.5 and constrained to the interval $[0.1, 5]$. Softmax normalization is performed separately over the global and variation-enriched slots, after which Equations 8 and 9 are applied per head. Head outputs are concatenated, passed through dropout and layer normalization, and slots with mass at most ϵ_m are masked. The

relative local-variation score is clipped with the fixed cap $\rho = 4$ in every configuration, and the pooling stabilizer is fixed at $\epsilon_m = 10^{-8}$.

Controlled tokenizer baselines. FixedPower and AdaptivePower use soft power-diagram memberships

$$p_{ik} \propto \exp\left(\frac{u_k - \|\mathbf{D}(\mathbf{z}_i - \mathbf{c}_k)\|_2^2}{\tau}\right),$$

where $\mathbf{D}$ is a learned positive diagonal coordinate scale and τ is the assignment temperature. Mass-normalized pooling then aggregates the encoded visible events. FixedPower uses deterministic Sobol centers $\mathbf{c}_k$ and $u_k = 0$. AdaptivePower predicts centers and zero-mean power offsets from slot states that attend to the visible input. The sparse MeshTok adapter partitions joint space–time into B_t temporal bins and coarse spatial cells, then replaces the highest-scoring occupied cells with spatial children. Its score combines visible-only local variation with within-cell feature anomaly. The sparse MATEY adapter uses the same coarse-to-fine layout and selects refinements from within-cell feature variance using threshold γ. Both adapters retain coarse and refined tokens in one fixed-length sequence and attach cell center and scale to the pooled features.

Tables A1–A3 report the resolved configurations used for the comparisons in Table 1. Within each row, every tokenizer uses the same observation protocol, forecasting architecture, optimizer, and training budget. Here, d_x is the coordinate dimension, d_y is the number of observed physical channels, N_v is the total number of visible observations in the input window, and N_q is the number of target locations per predicted frame.

Table A1: Forecasting protocols for the reported main comparisons. Split counts are complete physical trajectories or simulation configurations in train/validation/test order and precede window extraction.

Dataset	**Split**	d_x	d_y	**Context**	**Forecast**	N_v	N_q	**Batch**
Radial SWE A	160/40/40	2	1	5	16	1,024	512	4
Radial SWE B	160/40/40	2	1	5	16	1,024	512	4
PDEArena spherical SWE	200/50/50	3	3	5	16	1,024	512	4
PDEBench CFD-rand	800/200/200	1	3	5	36	512	512	2
Euler CRP	1144/143/143	2	5	5	16	1,024	512	4
Burgers 1D	2048/512/512	1	1	10	11	512	256	11
Advection 1D	4096/512/512	1	1	10	11	512	256	11
PDENNEval 2D Burgers	800/100/100	2	2	5	16	1,024	512	4
PDEArena Navier–Stokes	100/25/25	2	3	5	9	1,024	512	4
RealPDEBench FSI	35/7/9	2	2	10	11	1,229	512	4
RealPDEBench combustion	21/4/5	2	1	10	11	1,229	512	4
RealPDEBench cylinder wake	62/10/19	2	2	10	11	2,458	512	4
RealPDEBench controlled cylinder	66/10/20	2	2	10	11	2,458	512	4

All reported configurations use two latent-processor layers, one decoder layer, zero dropout, AdamW, gradient clipping at 1, and a token-count balance coefficient of 0.003. "Cosine" denotes an eight-epoch linear warm-up followed by cosine decay to 0.05 of the initial learning rate. Constant schedules have no warm-up. Early-stopping patience is 24 epochs. The minimum checkpoint epoch is 12 for the simulated settings, except for Burgers 1D and Advection 1D, where it is 15 and 20, respectively; it is 8 for the RealPDEBench settings.

Table A2: Shared architecture and optimization configurations. d is the hidden width, H the number of attention heads, K the number of token slots, F the number of Fourier bands, and λ_{ent} the assignment-entropy coefficient.

Dataset	d	H	K	F	**Epochs**	**LR**	**Weight decay**	λ_{ent}	**Schedule**
Radial SWE A	256	8	32	6	120	3×10^{-4}	10^{-4}	0	Cosine
Radial SWE B	128	8	32	4	120	3×10^{-4}	10^{-4}	0.001	Constant
PDEArena spherical SWE	128	8	32	4	120	9×10^{-4}	10^{-4}	0	Constant
PDEBench CFD-rand	128	8	48	6	120	3×10^{-4}	10^{-4}	0.001	Cosine
Euler CRP	128	8	128	6	120	6×10^{-4}	10^{-4}	0.001	Constant
Burgers 1D	128	8	32	4	120	10^{-3}	0	0.001	Constant
Advection 1D	192	8	16	6	120	3×10^{-4}	10^{-4}	0.001	Constant
PDENNEval 2D Burgers	256	8	32	6	120	3×10^{-4}	10^{-4}	0	Cosine
PDEArena Navier–Stokes	256	8	32	6	120	3×10^{-4}	10^{-4}	0.001	Constant
RealPDEBench systems	192	2	192	6	100	6×10^{-4}	10^{-4}	0.003	Constant

Table A3 gives the active controls that are specific to token construction. The columns k_{nn}, f_ℓ, λ, η, and β are the STRATATOK local-neighbor count, variation-path slot fraction, pathway scale, pathway contrast, and slope–spread mixing weight. The final three columns give the preferred number of temporal bins, refinement fraction, and variance threshold used by the sparse MeshTok and MATEY adapters; γ applies only to MATEY. The adapters use native input coordinates in every setting.

Table A3: Tokenizer-specific configurations used in the final evaluation.

Dataset	k_{nn}	f_ℓ	λ	η	β	B_t	f_{ref}	γ
Radial SWE A	32	0.500	1.000	1.00	0.50	5	0.750	0.50
Radial SWE B	16	0.375	1.000	1.00	0.50	5	0.750	0.05
PDEArena spherical SWE	8	0.500	0.250	0.50	0.50	5	0.750	0.20
PDEBench CFD-rand	8	0.625	1.000	1.00	0.50	5	0.125	0.20
Euler CRP	16	0.750	1.000	2.00	0.50	5	0.750	0.20
Burgers 1D	8	0.500	0.500	1.00	0.50	10	0.750	0.80
Advection 1D	8	0.375	4.000	1.00	0.50	10	0.500	0.05
PDENNEval 2D Burgers	8	0.500	2.000	1.00	0.75	5	0.250	0.20
PDEArena Navier–Stokes	32	0.500	4.000	0.50	0.25	5	0.750	0.50
RealPDEBench FSI	16	0.125	0.500	1.00	0.50	5	0.250	0.20
RealPDEBench combustion	16	0.125	0.500	1.00	0.50	5	0.250	0.20
RealPDEBench cylinder wake	16	0.125	0.500	1.00	0.50	5	0.250	0.20
RealPDEBench controlled cylinder	16	0.125	0.125	1.00	0.50	5	0.250	0.20

The observation-density and sampling-pattern studies retain the corresponding full-budget architecture above and vary only the visible-observation count or sampling process. The token-budget study keeps all other settings fixed and uses $K = 16/32$ for Radial SWE A and PDENNEval 2D Burgers, $K = 24/48$ for CFD-rand, and $K = 96/192$ for RealPDEBench combustion. Complete resolved YAML files accompany the implementation.

The stabilized normalized assignment entropy used by those configurations is

$$p^{\epsilon}_{ik} = \max(\bar{\alpha}_{ik}, 10^{-8}), \qquad \mathcal{L}_{\text{ent}} = -\frac{1}{N_v \log K} \sum_{i\in\mathcal{V}} \sum_{k=1}^{K} p^{\epsilon}_{ik} \log p^{\epsilon}_{ik}, \tag{11}$$

where $\bar{\alpha}_{ik} = H^{-1} \sum_h \alpha_{ihk}$. The 10^{-8} floor stabilizes the logarithm and is confined to this auxiliary loss.

The token-count balance term uses a straight-through hard version of these head-averaged assignments. Write $\bar{\boldsymbol{\alpha}}_i = (\bar{\alpha}_{i1}, \ldots, \bar{\alpha}_{iK})$. For each valid observation, let

$$\mathbf{h}_i = \mathrm{onehot}\left(\arg\max_k \bar{\alpha}_{ik}\right), \qquad \widehat{\boldsymbol{\alpha}}_i = \bar{\boldsymbol{\alpha}}_i + \mathrm{stopgrad}(\mathbf{h}_i - \bar{\boldsymbol{\alpha}}_i), \tag{12}$$

$$m_k = \sum_{i \in \mathcal{V}} \widehat{\alpha}_{ik}, \qquad \bar{m} = \frac{1}{K}\sum_{k=1}^{K} m_k. \tag{13}$$

Thus the forward assignment is one-hot while its gradient follows the soft assignment. The per-sample balance loss is

$$\mathcal{L}_{\mathrm{bal}} = \frac{1}{K}\sum_{k=1}^{K} \frac{(m_k - \bar{m})^2}{\bar{m}^2 + 10^{-8}}. \tag{14}$$

The computation pools all K slots from both $\mathcal{K}_g$ and $\mathcal{K}_\ell$ before minibatch averaging. The reported models minimize

$$\mathcal{L} = \mathcal{L}_{\mathrm{forecast}} + \lambda_{\mathrm{bal}}\mathcal{L}_{\mathrm{bal}} + \lambda_{\mathrm{ent}}\mathcal{L}_{\mathrm{ent}}, \qquad \lambda_{\mathrm{bal}} = 0.003, \tag{15}$$

where λ_{ent} is zero unless otherwise reported in Table A2.

B.3 Proofs of the structural properties

For a valid observation i, the two within-pathway distributions satisfy $\sum_{k\in\mathcal{K}_g} p^g_{ik} = 1$ and $\sum_{k\in\mathcal{K}_\ell} p^\ell_{ik} = 1$. Since $g_i + \ell_i = 1$, Equation 8 gives $\sum_k \alpha_{ik} = 1$. Consequently $\pi_{ik} = \alpha_{ik}/N_v$ has source marginal $1/N_v$, target marginal μ_k/N_v, and total mass one.

The clipped score obeys $0 \le q_i \le \rho$. Hence $r_i = \lambda q_i^\eta \le \lambda\rho^\eta$ and $g_i = (1 + r_i)^{-1} \ge (1 + \lambda\rho^\eta)^{-1}$. Summing over observations proves the global-mass bound in Equation 10. For the canonical choice $\lambda = \eta = 1$, clipping can only reduce the sample mean of $a_i/\overline{a}$, so $N_v^{-1}\sum_i q_i \le 1$. Convexity of $(1 + q)^{-1}$ then yields $N_v^{-1}\sum_i g_i \ge 1/2$.

The local share $\ell(q) = \lambda q^\eta/(1 + \lambda q^\eta)$ is nondecreasing, while $g(q) = 1 - \ell(q)$ is nonincreasing. The rearrangement inequality therefore places the assignment-weighted mean of q_i in the variation-enriched pathway no lower than its unweighted mean and the corresponding global mean no higher, whenever both pathways have positive mass. Finally, define $\widetilde{\mu}_k = \max(\mu_k, \epsilon_m)$. The coefficients $\alpha_{ik}/\widetilde{\mu}_k$ are nonnegative and sum to $\mu_k/\widetilde{\mu}_k \le 1$. Assigning the remaining coefficient $1 - \mu_k/\widetilde{\mu}_k$ to the origin writes $\widetilde{\mathbf{t}}_k$ as a convex combination of the origin and the received content vectors. If $\mu_k \ge \epsilon_m$, the remaining coefficient is zero and $\widetilde{\mathbf{t}}_k$ is the barycentric projection of the received vectors. This proves the final property in Equation 10.

B.4 Metric aggregation

For method a, run s, and physical trajectory r, we first average squared error over target coordinates and channels. Windows from one source trajectory remain in the same physical group. The reported run-level RMSE is

$$e_{a,s,r} = \frac{1}{|\Omega_r|}\sum_{(q,c)\in\Omega_r} \left(\widehat{y}^{(a,s)}_{rqc} - y_{rqc}\right)^2, \tag{16}$$

$$R_{a,s} = \sqrt{\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \frac{1}{|\mathcal{R}_g|}\sum_{r\in\mathcal{R}_g} e_{a,s,r}}. \tag{17}$$

Here $\mathcal{R}_g$ contains the windows from physical group g. Tables report $\mathrm{GM}_s(R_{a,s})$. Relative reductions compare these geometric means, and Figure 3 spans the minimum and maximum matched-run RMSE.

B.5 Evaluation provenance

Machine-readable manifests record run identifiers, split and configuration hashes, checkpoint-selection rules, and analysis hashes. Training and checkpoint selection use validation data; reported

metrics are computed on the test split. The routing ablation reported in Section 4.4 uses the same run identifiers, forecasting architecture, and training schedule across all five constructions on each dataset. Their configuration files differ only in the routing construction.

C Additional controls

C.1 Transolver++ token-construction implementation

The main comparison implements the token-construction mechanism introduced by Transolver++ (Luo et al., 2025) within the shared sparse forecasting architecture. The implementation uses a shared slice projection, predicts pointwise assignment temperatures from encoded observations, and applies Gumbel-softmax assignment to form locally adaptive soft slices. All remaining encoder, processor, decoder, optimization, and evaluation settings follow the matched protocol described in Appendix B.

C.2 End-to-end sparse forecasting comparison

We additionally compare the complete forecasting system with DINo and STRIDE under the same sparse-input and target-query protocol (Yin et al., 2023; Tong & Chen, 2026). DINo couples latent neural-ODE evolution with coordinate-conditioned implicit decoding, while STRIDE uses temporal sequence encoding with coordinate-conditioned decoding.

Table A4: End-to-end sparse forecasting comparison with DINo and STRIDE. Values are RMSE; lower is better.

Dataset	StrataTok	DINo	STRIDE
Radial SWE A	**0.1596**	0.2152	0.2983
Radial SWE B	**0.1682**	0.1982	0.2651
PDEArena spherical SWE	0.6317	**0.6115**	0.6202
PDEBench CFD-rand	**0.2183**	0.4653	0.4545
Euler CRP	**0.6377**	0.7457	0.7401
Burgers 1D	**0.2604**	0.3127	0.3180
PDENNEval 2D Burgers	**0.0275**	0.1119	0.1270
Advection 1D	**0.2097**	0.3003	0.2942
PDEArena Navier–Stokes	**0.4837**	0.5121	0.4905
RealPDEBench FSI	**0.2913**	0.5712	0.5933
RealPDEBench combustion	0.6635	**0.6140**	0.6879
RealPDEBench cylinder wake	**0.2624**	0.3849	0.3992
RealPDEBench controlled wake	**0.2539**	0.4766	0.5341

StrataTok records the lowest RMSE in 11 of the 13 settings. It is lower than STRIDE in 12 settings and lower than DINo in 11. DINo leads on spherical SWE and combustion, while STRIDE's only lead is on spherical SWE. Across the remaining simulated and real-world systems, the tokenizer-centered forecasting model remains competitive with both task-specific sparse forecasting architectures.

C.3 Observation robustness and routing-control details

The shuffled control deterministically permutes event-level variation-bank shares within each sample while preserving their distribution and total mass. The fixed-marginal control replaces each event share by the sample-average share, preserving the sample's total allocation to both banks but removing event-level alignment. A static dual-bank control assigns one half of every event's mass to each bank. These controls isolate event-level alignment, additional slots, sample-level bank mass, and unrestricted event routing.

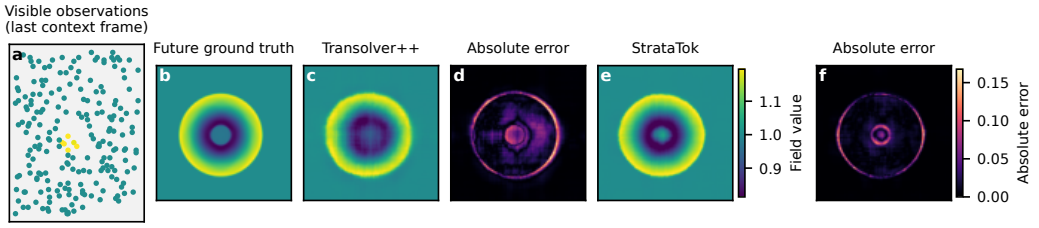


Figure A1: Qualitative prediction on a Radial SWE A test trajectory. The first panel shows visible observations in the final context frame. Ground truth and both predictions share one field-value scale, while the error maps share one absolute-error scale.