SCALECOMPOSER: SCALE-CONDITIONED COMPOSITIONAL LEARNING FOR SOURCE-COUNT EXTRAPOLATION IN PHYSICAL FIELDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting physical fields under changing source configurations supports the design and operation of indoor air movement devices, but exhaustive characterization becomes impractical as configurations multiply. We study source-count extrapolation under the hypothesis that low-source supervision generalizes unevenly across spatial scales, and that organizing corrections by scale can improve extrapolation to unseen, higher-order configurations. We introduce ScaleComposer, a compositional framework that aggregates source and source-pair representations through separate attentions conditioned on location and scale. Using parameters shared across source counts, the framework projects decoded corrections onto prescribed frequency bands and combines them to refine a reference field. Results show that the model supports extrapolation from single- and two-source configurations to configurations with up to six sources on measured multi-ceiling-fan airflow and six controlled two-dimensional Navier–Stokes regimes spanning source overlap and advective dynamics. On measurements, ScaleComposer reduces mean absolute error by approximately 10% relative to the strongest evaluated baseline. Separate source and pair attention lowers measurement MAE by 4.7% relative to joint attention, while band restriction lowers it by 9.4% relative to unrestricted. Spectral restrictions reduce intermediate-band errors relative to an unrestricted variant. Our approach provides a promising solution to scalable field prediction by composing information learned from low-source cases.

1 INTRODUCTION

Predicting physical fields under changing source configurations is important for the design and operation of systems with multiple interacting sources, motivating learned simulators and PDE surrogates (Sanchez-Gonzalez et al., 2020; Kochkov et al., 2021; Wen et al., 2025). For example, the air-speed distribution generated by ceiling fans influences thermal comfort and air quality (Tartarini et al., 2026) and informs decisions about placement and operating settings (Raftery et al., 2019; Luo et al., 2021). Therefore, accurate predictions of the induced indoor airflow are essential to inform the decisions on device type, placement and operating settings. However, iterating candidates through spatial measurements or high-fidelity simulations becomes increasingly demanding as the number of sources and operating choices grows. This motivates learning from limited individual sources to predict physical fields generated by larger source sets. We formulate this problem as source-count extrapolation, using ceiling fans as the application. A model learns from background field and configurations with at most two active sources, and predicts steady-state mean air speed fields for configurations with more. More generally, this is set-to-field prediction in which test sets are larger than any set seen in training.

Increasing the source count changes both the number of inputs and the relations among them. Two active sources produce one pair, whereas six can produce fifteen simultaneously active pairs. Predicting fields for such larger source sets presents different challenges for simulation and learning. Reynolds-averaged Navier-Stokes (RANS) provides an established approach to predicting mean flows through turbulence closure (Menter, 1994; Duraisamy et al., 2019), but conventional simulations require solving the coupled flow equations for each new configuration. Exploring the growing configuration space therefore remains computationally demanding.

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.

Neural operators (Lu et al., 2021; Li et al., 2021; Kovachki et al., 2023) offer expressive nonlinear field mappings, while set architectures accommodate variable numbers of interacting inputs (Zaheer et al., 2017b; Lee et al., 2019). These capabilities make them promising for configuration-dependent prediction, but accommodating more inputs does not itself establish how information learned from smaller configurations should be combined (Xu et al., 2021; Yehudai et al., 2021). Under single- and two-source supervision, a model observes individual responses and isolated pairs, but receives no direct supervision on configurations in which several pair relations coexist. Predicting larger source sets consequently requires determining which learned information remains useful and how its contribution should change with the surrounding configuration. This motivates an explicit composition rule that reuses source and pair representations (Chang et al., 2017; Battaglia et al., 2018) while adapting their contributions to the spatial field being predicted.

We hypothesize that information learned from low-source configurations transfers unevenly across spatial scales. Here, we introduce ScaleComposer, a framework for source-count extrapolation that combines source and relational representations through conditioned aggregation and scale-separated residual reconstruction as below:

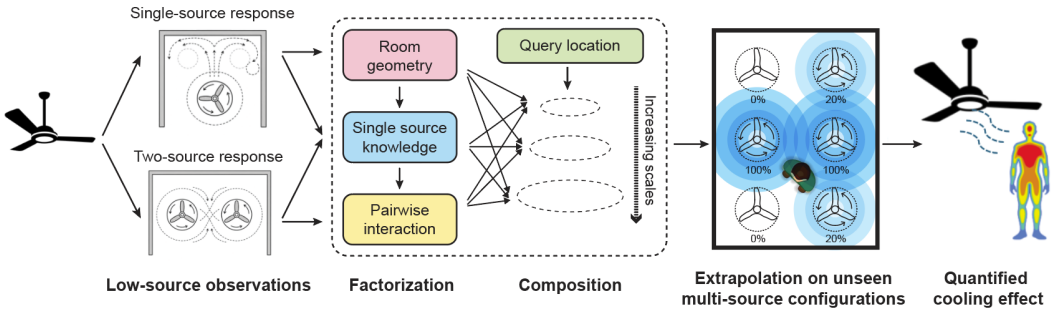

Figure 1: Conceptual overview: ScaleComposer factorizes low-source observations into source and relational representations, and recomposes them for unseen configurations containing more interacting sources. The composition is conditioned on the query location and organized across spatial scales, enabling prediction of configuration-dependent physical fields from limited supervision.

We demonstrate its capabilities on unseen tasks, which are challenging for data-driven models:

- ScaleComposer and its variant reliably model two complementary domains: (1) realistic ceiling-fan air-speed field measurements, where the unseen multi-source configurations contains multiple pairwise interactions under realistic sampling conditions, and (2) controlled two-dimensional incompressible Navier–Stokes experiments, where source overlap and advective dynamics are systematically varied to isolate the physical factors that govern when low-source interaction patterns transfer to higher-source configurations.
- ScaleComposer outperforms graph-based, attention-based, and Fourier-based baselines by almost 10%. Relational set representation and spectrally constrained correction contribute gains of 4.7% and 9.4% on measured data. Numerical results reveal metric and scale tradeoffs, with 97.7% of the six-source RMSE reduction relative to the reference occurring in the mean and lowest-frequency components.

2 Related Work

Relational Representations and Set-Size Generalization Deep Sets and Set Transformer encode unordered inputs through permutation-invariant aggregation (Zaheer et al., 2017b; Lee et al., 2019), while Interaction Networks share object and relation functions demonstrate generalization across object counts and configurations (Battaglia et al., 2016b). These architectures accommodate variable-sized inputs, but accurate prediction at unseen set sizes remains a challenge. Xu et al. (2021) connect algorithmic extrapolation to task-specific structure encoded in architectures or features, while Yehudai et al. (2021) identify changes in local-structure distributions as an obstacle to graph size generalization. Our model reuses source and symmetric pair representations

through configuration-dependent aggregation to examine whether learned relations remain useful as the source count increases.

Neural operators and multiscale representations Neural operators learn nonlinear maps between function spaces and can share parameters across spatial discretizations (Kovachki et al., 2023). GNOT combines heterogeneous attention with geometric gating for multiple input functions and multiscale prediction (Hao et al., 2023). Multiwavelet operators learn kernel projections on fixed bases across scales (Gupta et al., 2021). FNO parameterizes operators in Fourier space (Li et al., 2021); AFNO uses Fourier-domain token mixing (Guibas et al., 2021); SAOT combines local and global spectral representations (Zhou et al., 2026). Rather than spectral token mixing by itself, our construction uses scale to condition source retrieval and constrain candidate residual fields. Its evaluation concerns source-count extrapolation, which is distinct from transferring predictions between discretizations.

Prior-conditioned and residual field prediction Prior-conditioned prediction reuses existing model outputs, while residual learning estimates corrections to a reference field. Multi-fidelity DeepONet combines these strategies by coupling low- and high-fidelity operators through input augmentation and residual learning (Lu et al., 2022). DeltaPhi learns residuals between similar physical states to improve data-limited operator learning, where the state relationships also provide implicit data augmentation for data-limited operator learning (Yue et al., 2025). Our model uses frozen single-source predictions both to condition source and pair representations and to construct a non-additive reference. It learns scale-conditioned residual corrections to this reference, examining whether information learned from at most two active sources transfers to larger source sets.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

To predict the joint scalar field of a source configuration containing more active sources than any configuration used for training, we consider a fixed physical environment e, including the spatial domain $\Omega \subset \mathbb{R}^d$, geometry, and boundary conditions. Let $\mathcal{Z}$ denote the source-token space. Each localized source is described by

$$\mathbf{z_i} = (\mathbf{r_i}, a_i, \boldsymbol{\eta_i}) \in \mathcal{Z}, \tag{1}$$

where $\mathbf{r}_i$, a_i, and $\boldsymbol{\eta}_i$ denote source location, operating condition, and additional source or geometric attributes, respectively. A configuration space $\mathfrak{C}_e$ is a physically admissible finite unordered set

$$\mathcal{C} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\} \in \mathfrak{C}_e, \qquad \mathbf{z}_i = (\mathbf{r}_i, a_i, \boldsymbol{\eta}_i), \tag{2}$$

For the airflow application, the sources jointly drive a flow, and the prediction target is its temporally averaged scalar speed s. Measurements at different locations need not be simultaneous; comparable quasi-steady operating conditions are assumed.

The available low-source resources comprise single-source data $\mathcal{D}^{(1)}_{\text{prior}}$ used to construct the frozen priors and target-environment data $\mathcal{D}^{(\leq 2)}_e$ from configurations with at most two active sources, including background observations. We denote the configuration–target pairs used for residual-model training by

$$\mathcal{D}_{\leq 2} = \left\{(\mathcal{C}_m, \overline{y}_{\mathcal{C}_m}) : |\mathcal{C}_m| \leq 2\right\}_{m=1}^{N_{\text{train}}}. \tag{3}$$

Here $\overline{y}_{\mathcal{C}_m}$ contains the sampled scalar-speed targets for configuration $\mathcal{C}_m$. For measurements, all-fans-off observations supply the background input $b_e(q)$ and are excluded from $\mathcal{D}_{\leq 2}$ as residual-training targets.

The objective is to learn a predictor $\widehat{\mathcal{F}}_{\theta,e}(\mathcal{C})(\mathbf{x})$ from these low-source resources and apply it to configurations with $|\mathcal{C}| > 2$. Their measured fields are unavailable for training and model selection in the independent evaluation. We distinguish source-count extrapolation from joint source-count and relational extrapolation, where the latter additionally contains source attributes or pair relations outside those represented by the low-cardinality training configurations.

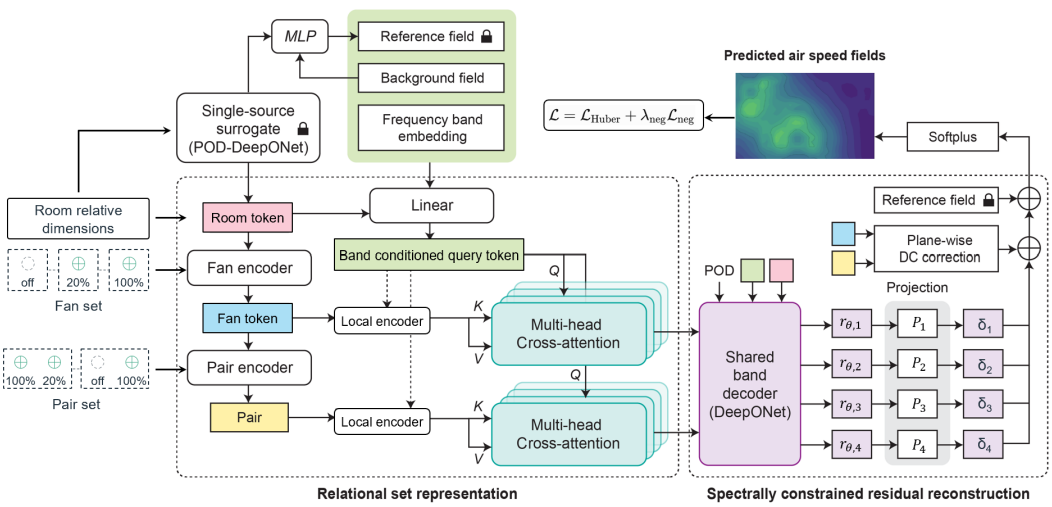

Figure 2: Overview of ScaleComposer for source-count extrapolation. Band-conditioned cross-attention aggregates source and source-pair representations, while spectrally constrained residual reconstruction refines a reference field built from single-source predictions and background observations. Trained on configurations with at most two active sources, the model predicts time-averaged scalar air-speed fields for unseen configurations with more active sources. Lock symbols indicate frozen components.

3.2 Relational Set Representation

Using ceiling fan as application, we represent each configuration by a set of fan tokens and pair tokens for unordered fan pairs. Fan tokens combine source attributes with information from a frozen single-fan surrogate $\mathrm{M1}$, while pair tokens encode interactions between fans. Query-dependent aggregation then selects information for each location and frequency band. This construction follows the shared-object and shared-relation principles of relational models (Battaglia et al., 2016a), while preserving the source-order symmetry of set representations (Zaheer et al., 2017a). Full transformations and feature definitions are given in Appendix A.2.

Input We use a fixed, non-additive reference field s_{ref} constructed from pretrained single-source surrogate $\mathrm{M1}$ and the measured background field $b_e(q)$. This reference provides a data-driven starting point for configuration-dependent residual reconstruction. $\mathrm{M1}$ denotes a POD-DeepONet with 24 fixed spatial modes. It supplies branch coefficients $\mathbf{h}_i^{\mathrm{M1}}$ and single-fan predictions converted to physical speed. A frozen composition model applies boundary correction to obtain $p_i(q)$ and MLP-based non-additive aggregation to obtain the reference field $s_{\text{ref}}(q)$.

Encoder The room token $\mathbf{t}_{\text{room}}$ provides room geometric context. For each frequency band $b \in \{1, 2, 3, 4\}$, a query encoder maps the prediction location q and a learned band embedding $\mathbf{e}_b$ to the query token $\mathbf{t}_{q,b}$. Each fan token $\mathbf{t}_i$ combines prescribed fan attributes $\mathbf{t}_i^{\text{attr}}$ with the response information $\mathbf{t}_i^{\text{prior}}$ encoded by the frozen single-fan surrogate. To make a relation depend on the properties of both participating sources, each unordered pair $i < j$ receives pair token $\mathbf{t}_{ij}$. Parameters are shared across items within each type, with separate encoders for fans and pairs. Re-indexing fans therefore permutes the token sets, and exchanging pair endpoints leaves $\mathbf{t}_{ij}$ unchanged. The pair tokens are latent relational features, without assuming that the physical response terminates at pairwise interactions. The fan and pair token sets represent the configuration. They are independent of query location and frequency band, and reused across both.

Attention Fans and pairs use separate multi-head cross-attention modules. This preserves their distinct feature semantics and separates the normalization of n fan items from that of $n(n-1)/2$ pair items. For type $\tau \in \{\text{fan}, \text{pair}\}$, frequency band b, head k, and item index a, it has separate affine query $\mathbf{q}_{b,k}^{\tau}$, key $\mathbf{k}_{b,a,k}^{\tau}$, and value projections $\mathbf{v}_{b,a,k}^{\tau}$. Each module has four heads of width $d_{\text{head}} =$

32 and uses scaled dot-product scores (Vaswani et al., 2017). The four heads are concatenated and passed through a bias-free projection to give $\mathbf{z}_b^\tau$. Projections are independent across types and shared across items, locations, and bands within each type. Separate relational aggregation is an inductive bias whose predictive benefit is evaluated through ablations detailed in Section 4. The resulting feature field $\mathbf{F}_b(q)$ provides a fixed-width interface between the source set and spatial reconstruction.

3.3 Spectrally constrained residual reconstruction

Learning spatial corrections from a limited set of configurations requires controlling which variations the model can express. A shared DeepONet controls their spatial representation, and fixed Fourier projectors constrain their frequency support. Architectural details are provided in Appendix A.2.

Decoder For each horizontal plane and frequency band, the decoder transforms the aggregated features into residual corrections $\widetilde{r}_b(q)$ using a DeepONet coefficient–basis factorization (Lu et al., 2021). The branch maps $\mathbf{F}_b$ to coefficients shared across the plane, while the trunk evaluates learned spatial basis functions using only $\bar{q}$. All bands share decoder parameters, with band identity entering through $\mathbf{F}_b$. M1 uses an analogous representation with fixed POD modes.

Spectral constraints We specify the admissible spatial variation within each component with fixed frequency projections. Motivated by Fourier representations for operator learning (Li et al., 2021) and frequency-dependent learning dynamics in ReLU networks (Rahaman et al., 2019), we constrain each decoded correction δ_b through fixed Fourier projection to restrict frequency support, complementing the decoder's finite-dimensional spatial representation. On a prescribed horizontal prediction grid, a signed plane-wise term c_{DC} adjusts the mean and gives the candidate residual r_{cand} together with δ_b. Then a scaled softplus readout maps s_{ref} and sum of r_{cand} to the non-negative air speed prediction $\widehat{s}(q)$.

3.4 Model training

We jointly optimize the model by supervising the final readout predictions $\widehat{s}_t(q)$ against the corresponding ground-truth speeds $y_t(q)$, where t indexes a training example. The prior map remains fixed throughout the training. Let $\mathcal{G}_t$ denote the complete grid and $\mathcal{M}_t \subseteq \mathcal{G}_t$ its supervised nodes. The training objective is the uniform average of $\mathcal{L}_t$ over training planes:

$$\mathcal{L}_t = \frac{1}{|\mathcal{M}_t|} \sum_{q \in \mathcal{M}_t} \rho_{0.1}\left(\frac{\widehat{s}_t(q) - y_t(q)}{s_\star}\right) + \frac{\lambda_{\mathrm{neg}}}{|\mathcal{G}_t|} \sum_{q \in \mathcal{G}_t} \left[\max\left(0, -\frac{v_t(q)}{s_\star}\right)\right]^2, \tag{4}$$

where $\rho_{0.1}$ is the Huber loss (Huber, 1964) with threshold 0.1 and $\lambda_{\mathrm{neg}} = 0.01$. The second term imposes a soft non-negativity constraint on the pre-readout field over the complete grid. Gradients propagate through the readout and Fourier projections, jointly learning all band components from the final-field targets. The selected model is applied to higher-count configurations without further parameter fitting. Further training details are specified in Appendix A.3.

4 Experiments

Empirical validation Ceiling fans enhance convective heat exchange with occupants, supporting thermal comfort through increased air movement. In spaces served by multiple fans, coordinating their operation requires understanding how their combined airflow changes across different configuration, which can be costly through measurements or high-fidelity simulations (e.g., CFD). We evaluate source-count extrapolation using measurements from a limited set of single- and two-fan configurations to predict air-speed fields under joint operation of up to six fans in real-world condition. The task is to predict spatial fields of time-averaged scalar air speed for configurations containing more active fans than those used for training. Measurements were acquired at heights of 0.1, 0.6, 1.1, and 1.7 m on a nominal 20×31 horizontal lattice, with coverage varying across configurations. Each target averages 100 samples collected at 1 Hz after a 30-second settling interval. An all-fans-off measurement supplies the nonzero ventilation background. Acquisition details are provided in Appendix A.1.1.

Numerical simulations We complemented the application-level evidence above with a controlled numerical benchmark. Our formulation draws on body-forced Navier-Stokes models with linear damping (Kochkov et al., 2021) and Brinkman solutions driven by localized Gaussian forces (Mondal et al., 2021). These established modelling ingredients motivate our source representation and damping term, with the advection-free cases recovering the Brinkman equations at steady state. We solved the non-dimensional, two-dimensional incompressible Navier–Stokes equations on the periodic domain $\Omega = [0, 2\pi)^2$:

$$\begin{aligned} \partial_t \mathbf{u} + \chi(\mathbf{u} \cdot \nabla)\mathbf{u} &= -\nabla p + \nu \nabla^2 \mathbf{u} - \alpha \mathbf{u} + \sum_{i \in S} \mathbf{b}_i, \\ \nabla \cdot \mathbf{u} &= 0, \end{aligned} \tag{5}$$

where $\mathbf{u}$ and p are velocity and pressure, ν and α are viscosity and damping coefficients, and $\mathbf{b}_i$ is the localized forcing from active source $i \in S$.

To examine how source overlap and flow dynamics affect extrapolation, candidate sources were arranged on a hexagon of two radius r to define low- and high-overlap geometries. Six regimes combined source overlap with weaker and stronger advective influence (Figure 5), obtained by varying viscosity at $\chi = 1$, and advection-off controls at $\chi = 0$. The prediction target was also time-averaged speed. Because taking the velocity magnitude remains non-additive even without advection, this control distinguishes nonlinear transport from the non-additivity of the scalar observation. Full specifications, numerical checks and an illustration are provided in Appendix A.1.2.

Baselines We compare ScaleComposer with retrieval-based residual learning and reference-augmented baselines spanning graph-based, attention-based, and Fourier-based architectures. DeltaPhi-DeepONet (Yue et al., 2025; Lu et al., 2021) serves as the retrieval-based baseline. The graph-based group comprises Ref+GNN (Scarselli et al., 2009), Ref+HGNN (Feng et al., 2019), and Ref+GAOT (Wen et al., 2025). The attention-based group includes Ref+GNOT (Hao et al., 2023), Ref+CViT (Wang et al., 2025), and Ref+Transolver++ (Luo et al., 2025). The Fourier-based group consists of Ref+GINO (Li et al., 2023) and Ref+FNO (Li et al., 2021). Here, Ref+ denotes a backbone augmented with the same reference field. These comparisons assess whether ScaleComposer improves source-count extrapolation beyond reference reuse alone.

Evaluation protocol We evaluate the proposed method across source counts and experimental domains to assess whether extrapolation performance remains consistent as interaction conditions change. Within each domain, models were trained on single- and two-source configurations and evaluated on configurations with three to six sources. We evaluate predictive performance using mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficient (PCC) (Schober et al., 2018). We interpret PCC jointly with MAE and RMSE to assess both spatial agreement and magnitude accuracy.

5 Results and Analysis

5.1 Model performance on source-count extrapolation

Average predictive performance On the test set in measurement domain Table 1, full ScaleComposer achieves the lowest mean MAE and RMSE and the highest mean PCC among the evaluated models. It reduces MAE by almost 10% compared to Ref+GNN. On the numerical dataset, ScaleComposer w/o Ref achieves the best MAE-PCC results. DeltaPhi-DeepONet matches the errors but exhibits an underperformed PCC (Appendix Table 6). Figure 3(d) shows that its MAE advantage over the selected baselines persists at all three evaluated source counts. Removing the reference field increases error, with the largest deterioration at six sources.

Distance beyond low-source supervision Training uses at most two active sources, so increasing the test source count extends prediction further beyond the training range. The same learned encoders and aggregation modules must accommodate more sources and simultaneous pairwise relationships, making extrapolation increasingly demanding. In Figure 3(a), full ScaleComposer achieves lower measurement MAE than the selected baselines at three, four, and six sources. Although errors increase across these configurations, ScaleComposer consistently retains its accuracy

Table 1: Full comparison with baseline methods on the measurements.

CATEGORY	MODEL	PERFORMANCE ON MEASUREMENTS		
		MAE ↓ (m/s)	RMSE ↓ (m/s)	PCC ↑
Prior retrieval	**DeltaPhi**-DeepONet	0.489 ± 0.096	0.624 ± 0.095	0.447 ± 0.070
Graph-based message passing or aggregation	Ref+**GNN**	0.203 ± 0.013	0.256 ± 0.004	0.834 ± 0.010
	Ref+**HGNN**	0.219 ± 0.031	0.278 ± 0.029	0.820 ± 0.021
	Ref+**GAOT**	0.224 ± 0.012	0.281 ± 0.007	0.800 ± 0.012
Attention-based feature aggregation	Ref+**GNOT**	0.213 ± 0.019	0.269 ± 0.016	0.832 ± 0.010
	Ref+**CViT**	0.235 ± 0.007	0.290 ± 0.007	0.787 ± 0.009
	Ref+**Transolver++**	0.231 ± 0.032	0.285 ± 0.027	0.819 ± 0.015
Fourier based	Ref+**FNO**	0.217 ± 0.010	0.283 ± 0.013	0.780 ± 0.013
	Ref+**GINO**	0.248 ± 0.024	0.331 ± 0.025	0.761 ± 0.018
Proposed method	**ScaleComposer**	$\mathbf{0.183 \pm 0.007}$	$\mathbf{0.241 \pm 0.006}$	$\mathbf{0.842 \pm 0.007}$
	ScaleComposer w/o Ref	0.301 ± 0.019	0.481 ± 0.027	0.541 ± 0.077

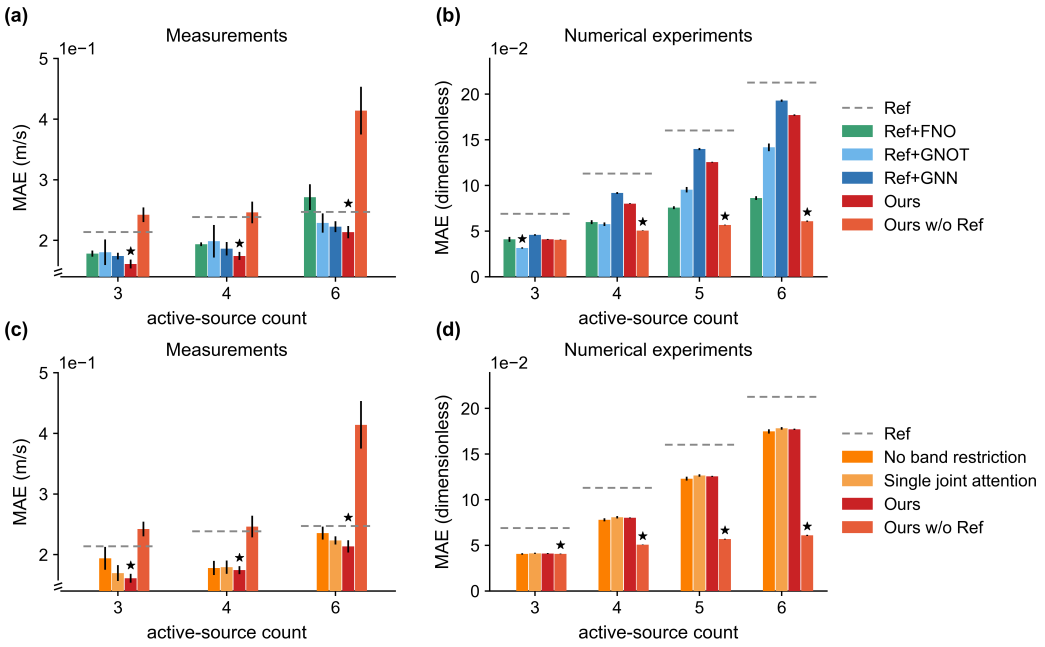

Figure 3: MAE by source count: selected baselines (a–b) and ablations (c–d). "w/o Ref" removes only output addition, and "No band restriction" uses shared union-band support, retaining Fourier projection.

advantage, supporting robust generalization across the tested source counts. Output reference addition is particularly beneficial for the measured task: it substantially reduces MAE at every evaluated count, with the largest improvement at six sources. The three configurations also differ in source layout and operating settings, so the observed increase in error reflects these changes together with the increase in source count.

Reference dependence in numerical task. On the numerical dataset, ScaleComposer without output reference addition maintains comparatively stable performance as the source count increases (Figure 3(b)). Although Ref+GNOT achieves the lowest MAE at three sources, the reference-free output variant performs best from four to six sources, with only modest error growth. At six sources, its MAE is less than half that of Ref+GNOT, Ref+GNN, and full ScaleComposer, while also remaining below Ref+FNO. These results demonstrate that the learned compositional predictor retains its

predictive accuracy further beyond the training source counts, with an increasingly substantial advantage over several competing methods. Reference addition is therefore a task-dependent design choice rather than a prerequisite for effective low-source information reuse.

Computational efficiency Computational efficiency is evaluated on a single 32GB NVIDIA RTX 5090 GPU. Shared encoders and decoder weights keep ScaleComposer below one million parameters, comparable to Ref+GNN and Ref+HGNN, while query-dependent processing and multiple band branches still require substantial computation. Its partial profiler FLOP count is among the highest reported. Six-source inference latency is similar to Ref+Transolver++ and higher than Ref+FNO and Ref+CViT, with training throughput similar to Ref+GAOT and Ref+Transolver++. Thus, the measurement accuracy gains involve a computational tradeoff despite the compact parameterization. Inference timings exclude frozen-prior construction. Details are given in Appendix A.4.1.

5.2 Contributions of relational aggregation and spectral constraints

Separate multi-head cross-attention. In Figure 3(c), separate attention aggregation achieves lower measurement MAE than single joint attention at all three evaluated source counts. Its advantage persists as more source pairs become simultaneously active, supporting the reuse of source and pair information through distinct aggregation pathways. The fan and pair modules use separate attention normalization sums, avoiding competition between the two item types as their relative numbers change. On the numerical task, the attention variants remain close across three to six sources (Figure 3(d)), possibly reflecting more complete low-source relation coverage. Separate attention aggregation therefore provides a consistent extrapolation benefit on measurements, while its numerical benefit is limited.

Band constraints. The no-band restriction control replaces each fixed Fourier projector P_b with $P_U = \sum_b P_b$, while retaining four band-conditioned correction fields, the shared decoder, parameter count, and total frequency support. Band restriction lowers measurement MAE by 9.4% relative to this control overall, and the improvement persists at every evaluated source count in Figure 3(c). On the numerical benchmark, however, aggregate MAEs remain similar (Figure 3(d)). It increases MAE but improve RMSE and PCC (Table 8). This disagreement across aggregate metrics suggests that band constraints redistribute prediction error across spatial scales, motivating the scale-resolved analysis in Section 5.3.

DC correction. The plane-wise DC correction c_{DC} complements the band-projected corrections δ_b. Both DC-only and w/o DC increase measurement MAE; removing DC raises it from 0.183 to 0.218 m/s with little change in PCC (Table 7). Numerical DC removal shows a similar pattern, but is a post-hoc intervention (Table 8). The measured ablations therefore support retaining both DC and band-projected corrections, while PCC alone understates the benefit of DC correction.

5.3 Scale-resolved extrapolation errors

We examine the frequency band decomposition of model predictions on numerical dataset (Table 9 in Appendix A.5). In the numerical results, the error of reference field increases with source count in every band. From 3-source to six-source, the largest growth are in the spatial mean (DC, 8.76 times) and coarsest band (B_1, 5.81 times). At six sources, reductions in these two error components account for 97.7% of ScaleComposer's total MSE reduction relative to the reference field. The improvement is therefore concentrated in the mean air speed and broad spatial structure.

Band restriction helps limit harmful corrections. "No band restriction" uses the union of all four bands but retains the four conditioned branches and parameter count. This control has lower B_1 and B_4 errors but higher total MSE than ScaleComposer. Across source counts, its B_2 error exceeds that of the reference field by 6.2-11.1% and exceeds that of the proposed by 24.0-36.3%. Thus, the benefit of per-branch restriction is to constrain this tradeoff. Fine-scale reference errors, B_3 and B_4, also persist in the matched advection-off regime, where the velocity dynamics are linear but the scalar-speed target remains non-additive. Such fine-scale errors therefore do not, by themselves, indicate nonlinear advective interactions.

5.4 Practical implications and limitations

Figure 4 compares ScaleComposer and CFD predictions with the chamber measurements for a three-source configuration. Both reproduce the broad low-speed region in the upper part of the plane and the higher-speed regions at the lower-left and central-right. CFD predicts sharper low-speed pockets surrounded by concentrated high-speed regions, whereas ScaleComposer yields a smoother field but misses some measured local extrema. For this configuration, ScaleComposer achieves lower MAE and RMSE and higher PCC than CFD.

ScaleComposer is trained directly on measured time-averaged scalar speed and does not reconstruct the velocity components (u, v, w). The scalar targets already reflect the combined effects of three-dimensional motion, allowing the model to learn their configuration-dependent response directly. However, the comparison does not establish whether additional vector supervision would improve accuracy.

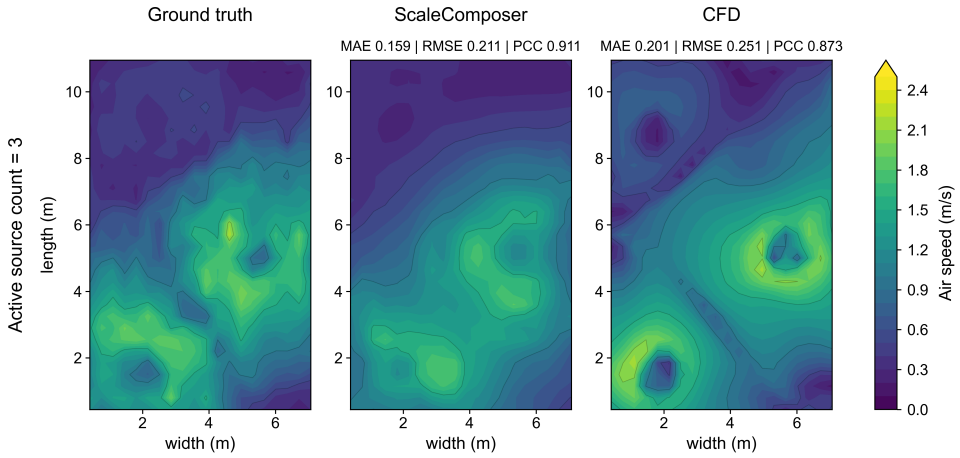

Figure 4: Air speed distributions for a configuration with three active sources.

6 Conclusion

ScaleComposer addresses an unseen compositional task: predicting the joint field of multiple active source pairs from supervision limited to at most one active pair per configuration. Shared source and pair encoders, separate scale-conditioned retrieval, and a common spatial decoder provide a reusable prediction rule as the active set grows. On measured airflow fields, the reference-assisted model achieves the best reported mean accuracy, with ablations supporting separate aggregation and band restriction. On controlled numerical fields, omitting output reference addition improves MAE and RMSE with lower PCC. Together, these findings support compositional field reconstruction while treating the reference as a setting-dependent output choice.

The evidence covers the evaluated configurations with up to six sources. Low-source observations do not uniquely determine arbitrary higher-order responses, and pair representations do not imply pairwise-additive physics. Further evaluation should test new joint configurations reserved from model development and examine how reference quality and spatial scales affect extrapolation. The present scalar-speed formulation also remains subject to measurement resolution and comparable averaging conditions and does not recover velocity directions or turbulence statistics.

A Appendix

A.1 Dataset details

A.1.1 Chamber measurements

Experimental setting. The empirical study reuses previously collected ceiling-fan air-speed measurements described by an earlier surrogate-modeling study (Zhang et al., 2026). We summarize the acquisition conditions relevant to the present prediction task here. The measurements were obtained in Lab 1 of the Human Centric Design Paired Comparison laboratory at The GEAR, Singapore. The principal floor area measures approximately $7.48\,\mathrm{m} \times 11.40\,\mathrm{m}$, with a $2.60\,\mathrm{m}$ ceiling; facade-side buffer spaces make the full room geometry depart from a simple rectangle.

The room contains six individually controlled Haiku L-Series fans (FR112C-U1H, Big Ass Fans), each with a diameter of $1.32\,\mathrm{m}$. The detailed acquisition description places the blade plane $2.24\,\mathrm{m}$ above the floor, corresponding to a mounting distance of $0.36\,\mathrm{m}$ below the ceiling. Fan settings were commanded through the building management system, with active operation between 20% and 100% of the rated maximum speed of 201 rpm.

Environmental conditions and background. Before acquisition, the room and adjacent buffer spaces were conditioned to a nominal temperature of $27\,^{\circ}\mathrm{C}$. Mechanical ventilation remained present, and supply airflow was monitored through the building management system. The campaign-level mean and standard deviation were $27.43 \pm 0.62\,^{\circ}\mathrm{C}$ for indoor temperature, $45.61 \pm 1.97\%$ for relative humidity, and $575.56 \pm 94.52\,\mathrm{m^3\,h^{-1}}$ for supply airflow. Measurements from different configurations were acquired in separate sessions under these controlled conditions. The all-fans-off measurement therefore represents the ventilated background, not an identically zero air-speed field.

Instrumentation. Air speed was recorded with omnidirectional hot-wire probes mounted at 0.1, 0.6, 1.1, and $1.7\,\mathrm{m}$ above the floor. The primary measurement system used Kanomax 0975-31 probes connected to 1590-00 and 1591-00 multichannel hubs. The reported probe range is 0.01–$25.0\,\mathrm{m\,s^{-1}}$, with a resolution of $0.01\,\mathrm{m\,s^{-1}}$ and an accuracy specified as the larger of 3% of the reading and $0.1\,\mathrm{m\,s^{-1}}$. The dataset records the probe type for each observation (Buckle et al., 2026).s

Spatial coverage and temporal averaging. The nominal measurement lattice contains $20 \times 31 = 620$ horizontal locations. The detailed dataset description gives width-wise and length-wise spacings of 0.375 and $0.370\,\mathrm{m}$, respectively. At four measurement heights, a complete configuration would contain 2480 location-height observations. Background and most multi-fan configurations were sampled across the full lattice. Single-fan and window-side pair configurations used denser sampling near the active fans and sparser coverage farther away. Consequently, the nominal lattice size does not imply that every configuration has the same number of observed points.

At each tripod placement, experimenters left the room before a 30-second settling interval. The retained window comprised 100 consecutive samples acquired at $1\,\mathrm{Hz}$; the settling interval was excluded from the reported mean. If $v_{\mathcal{C},h,p,t}$ is the recorded instantaneous scalar speed at location p, the target associated with this protocol is

$$y_{\mathcal{C},h,p} = \frac{1}{100} \sum_{t=1}^{100} v_{\mathcal{C},h,p,t}. \tag{6}$$

This quantity is a temporal mean of speed. The probes do not provide the velocity direction, so it is not a measured mean velocity vector or the magnitude of such a vector. Traversing the room also means that a reconstructed spatial field combines local averaging windows acquired at different times, under the assumption of comparable quasi-steady operating conditions.

Configuration coverage. The full collection contains 14 test configurations. Here we use the configurations listed in Table 2. The collection includes three-, four-, and six-source fields, but no five-source configuration. Its low-source measurements also do not cover every singleton and

pair appearing in the larger configurations. Any evaluation based on these measurements must therefore distinguish an increase in source count from changes in the source locations and pair relations represented by the available low-source support.

Table 2: Configurations and same original test IDs are used as the published measurement inventory (Zhang et al., 2026; Buckle et al., 2026). The active counts are obtained by counting nonzero fan settings.

Usage	Active fans	Test IDs	Operating patterns
Training	0	4	All fans off; mechanical-ventilation background.
	2	12, 13	Fans 1 and 4 together, both at 20% or both at 100%.
Testing	3	11	Fans 1, 3, and 5 at 100%, 20%, and 100%; other fans off.
	4	9	Fans 1, 2, 4, and 5 at 100%, 20%, 20%, and 100%.
	6	10	Fans 1, 3, 4, and 6 at 100%. Fan 2 and 5 at 20%.

A.1.2 CONTROLLED NUMERICAL EXPERIMENTS

We used a non-dimensional, two-dimensional flow benchmark to assess generalization from single- and two-source configurations to configurations containing three to six sources. The benchmark separates source overlap, nonlinear transport, and the scalar observation of a vector flow. Its periodic domain avoids introducing wall and inlet/outlet effects and admits a Fourier discretization. it is a controlled composition problem, rather than a calibrated model of three-dimensional room airflow.

Governing equations and damping. On $\Omega = [0, 2\pi)^2$, the velocity $\mathbf{u}$ and pressure p satisfy

$$\begin{gathered}\frac{\partial \mathbf{u}}{\partial t} + \chi(\mathbf{u}\cdot\nabla)\mathbf{u} = -\nabla p + \nu\nabla^2\mathbf{u} - \alpha\mathbf{u} + \sum_{i\in S}\mathbf{b}_i(\mathbf{x}),\\ \nabla\cdot\mathbf{u} = 0, \qquad \mathbf{u}(\mathbf{x},0) = 0.\end{gathered} \tag{7}$$

where $(\mathbf{u}\cdot\nabla)\mathbf{u}$ denotes nonlinear advection, $-\nabla p$ the pressure-gradient term, $\nu\nabla^2\mathbf{u}$ viscous diffusion, $-\alpha\mathbf{u}$ linear damping, and $\sum_{i\in S}\mathbf{b}_i(\mathbf{x})$ localized body forcing from the active-source subset S. Here ν is the nondimensional kinematic viscosity, $\alpha = 1$ is the damping rate, and $\chi \in \{0, 1\}$ switches advection off or on. Pressure enforces incompressibility, with zero spatial mean fixing its gauge. The applied body force need not itself be divergence-free. Advection couples the evolving velocity components, whereas viscosity diffuses momentum and damps spatial variations.

Linear drag provides a simple momentum-relaxation term, as used in quasi-two-dimensional fluid models (Mecke et al., 2024). Its specific role here follows directly from the periodic equations. Writing $\langle\cdot\rangle_\Omega$ for a spatial average gives

$$\frac{\mathrm{d}\langle\mathbf{u}\rangle_\Omega}{\mathrm{d}t} = -\alpha\langle\mathbf{u}\rangle_\Omega + \left\langle\sum_{i\in S}\mathbf{b}_i\right\rangle_\Omega. \tag{8}$$

The localized forces generally have a nonzero spatial mean. Viscosity does not damp this uniform mode, which would accelerate indefinitely under a constant nonzero mean force if $\alpha = 0$. Positive damping bounds that response without subtracting the mean force. Setting $\alpha = 1$ fixes a common unit relaxation time across regimes; it is a benchmark choice, not a fitted wall-friction coefficient.

Localized sources and geometric overlap. Each time-independent source is a smooth, directed body force,

$$\begin{aligned}\mathbf{b}_i(\mathbf{x}) &= a_i G_\sigma^{\mathrm{per}}(\mathbf{x} - \mathbf{c}_i)\mathbf{d}_i,\\ G_\sigma^{\mathrm{per}}(\mathbf{r}) &= \sum_{\mathbf{m}\in\mathbb{Z}^2}\exp\left[-\frac{\|\mathbf{r} + 2\pi\mathbf{m}\|_2^2}{2\sigma^2}\right].\end{aligned} \tag{9}$$

We fixed $\sigma = 0.45$ and drew each active amplitude independently from $\mathcal{U}[0.8, 1.2]$; inactive amplitudes were zero. The Gaussian avoids a singular point force, and wrapping preserves periodic smoothness. Its width spans approximately 4.6 solver-grid intervals and is held fixed so that overlap

is varied through separation. The implemented Gaussian is unnormalized and uses image offsets $m_x, m_y \in \{-1, 0, 1\}$.

The six candidate centers are $\mathbf{c}_i = (\pi, \pi) + r(\cos\theta_i, \sin\theta_i)$, where $\theta_i = 2\pi i/6$ and $i = 0, \ldots, 5$. The radii $r = 1.8$ and 0.65 define low and high forcing overlap. The mean normalized inner product of the six Gaussian profiles over all 15 unordered pairs is 0.00734 and 0.34590, respectively. These quantify overlap of the applied forcing, not of the resulting velocity fields. The fixed unit directions are $\mathbf{d}_i = (\cos(\theta_i + \delta_i), \sin(\theta_i + \delta_i))$, with $\boldsymbol{\delta} = (0.4, 1.2, -0.7, 0.9, -0.3, 1.7)$ radians. The oblique offsets break the rotational symmetry of the source arrangement.

Dynamical regimes and matched controls. For each geometry, we used $(\chi, \nu) = (1, 0.15)$, $(1, 0.025)$, and $(0, 0.025)$, giving six regimes, as illustrated in Figure 5. Lower viscosity reduces diffusive damping relative to transport. In the single- and two-source pilot, the maximum final-time ratio $\|(\mathbf{u} \cdot \nabla)\mathbf{u}\|_{L^2(\Omega)} / \|\nu \nabla^2 \mathbf{u}\|_{L^2(\Omega)}$ was 0.318 and 0.378 in the two high-viscosity geometries, compared with 2.21 and 2.99 in their low-viscosity counterparts. Accordingly, "weak" and "strong" denote relative advective influence, not a laminar–turbulent classification. Varying viscosity also changes momentum diffusion, so this contrast alone does not isolate advection. Each advection-off control instead matches its $\nu = 0.025$ counterpart in source centers, directions, amplitude realizations, damping, initial condition, and observation window, changing only χ. Amplitude distributions, but not all realizations, are shared between the other regimes.

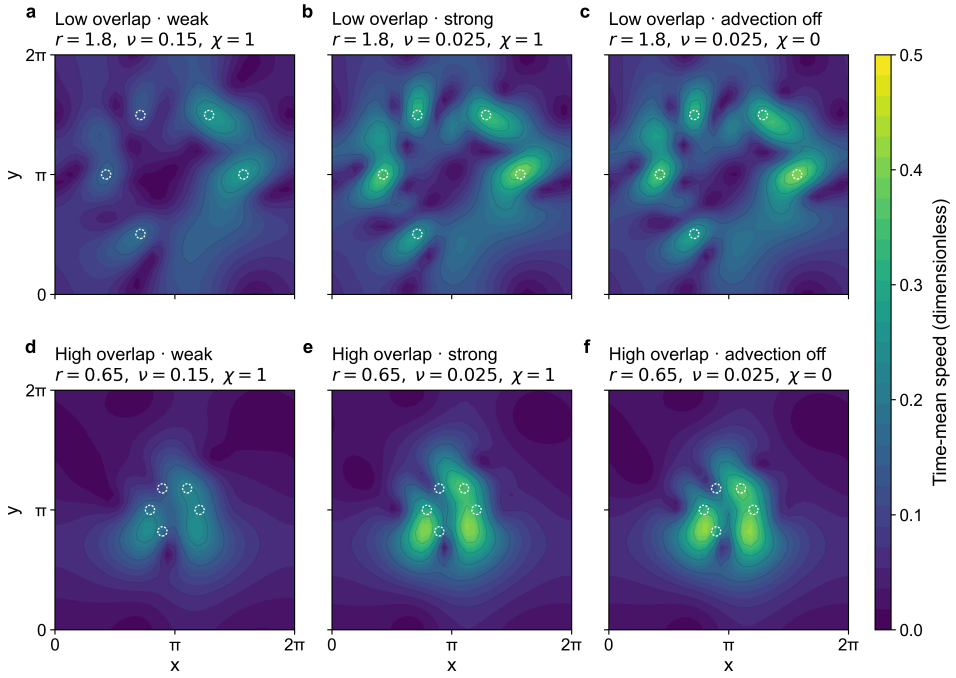

Figure 5: Illustration of the six numerical regimes at a fixed source count $N_s = 5$. Rows vary the degree of source overlap, from low ($r = 1.8$) to high ($r = 0.65$), while columns vary the transport regime: weak ($\nu = 0.15, \chi = 1$), strong ($\nu = 0.025, \chi = 1$), and advection-off ($\nu = 0.025, \chi = 0$). White dashed circles mark source locations, and contours denote the time-mean dimensionless speed.

Prediction target and interpretation of the control. The target is the time average of instantaneous speed,

$$\bar{s}_S(\mathbf{x}) = \frac{1}{T} \int_{t_0}^{t_0+T} \|\mathbf{u}_S(\mathbf{x}, t)\|_2 \, \mathrm{d}t, \qquad t_0 = 12, \quad T = 6, \tag{10}$$

rather than the magnitude of time-averaged velocity. With $\chi = 0$, the velocity dynamics are linear and isolated-source velocities superpose under matched conditions. Nevertheless, the norm implies $\bar{s}_S(\mathbf{x}) \leq \sum_{i \in S} \bar{s}_i(\mathbf{x})$, generally with strict inequality. The advection-off control therefore retains

nonadditivity from vector cancellation and the speed observation while removing nonlinear transport. It is not an additive-scalar ground truth.

Numerical implementation and verification. We used a float64 Fourier vorticity solver on a 64×64 grid, strict two-thirds truncation for quadratic advection, and second-order exponential Heun integration with $\Delta t = 0.025$. Diffusion and damping are integrated through the exponential linear propagator; the uniform velocity mode is retained separately. Trapezoidal averaging over $t \in [12, 18]$ precedes selection of every second node for the common 32×32 observation grid.

The settings were checked before formal high-source label generation using sources $\{0\}$ and $\{0, 1\}$ at amplitude 1.2 in every regime. The largest relative L^2 change in the pilot target arrays was 1.02×10^{-3} for refinement from 64^2 to 96^2 and 7.50×10^{-7} when replacing the averaging interval by $[18, 30]$. A separate transient time-step check supported second-order convergence. Across all 1,512 formal configurations, maximum final-time divergence, advective grid CFL, and relative half-window discrepancy were below 3.50×10^{-17}, 0.350, and 2.05×10^{-6}, respectively. The near-stationary labels do not establish turbulent generalization, and grid-refinement checks were limited to the stated low-source pilot.

Training and validation enumerate all single-source and two-source subsets; evaluation enumerates all three- through six-source subsets of the same six candidates, using separate amplitude random-number streams. Thus extrapolation concerns active-source count within fixed candidate geometries and an unchanged amplitude distribution.

A.2 Feature definitions and architecture

This section specifies the feature construction for the ceiling-fan application and describes how configuration-level tokens are transformed into scale-conditioned residual fields. All source-dependent modules share parameters across sources, and all pair-dependent modules share parameters across unordered pairs. Their input and output widths do not depend on the number of active fans.

A.2.1 Notation, units, and frozen inputs

Let the room extents along the physical coordinate axes be (L_x, L_y, H), a query be $q = (\mathbf{x}, h) = (x, y, h)$, and room-relative coordinates be $\bar{q} = (x/L_x, y/L_y, h/H)$. In the stored room descriptor, width corresponds to L_x and length to L_y; its storage order is $[L_y, L_x, H, \ldots]$. Horizontal coordinates are measured from a fixed room corner, and vertical coordinates are measured from the floor. The fan descriptor is

$$\mathbf{f}_i = [a_i, \alpha_i, D_i, z_i, \mu_i, x_i, y_i], \tag{11}$$

where a_i is normalized operating strength, $\alpha_i \in \{0, 1\}$ is the activity indicator, D_i is diameter, and μ_i is the ceiling-to-blade mounting distance. The fan centre is (x_i, y_i, z_i), with z_i denoting blade height. All geometric lengths are in metres, and scalar air speeds are in $\mathrm{m\,s^{-1}}$. The operating-strength encoding is $a_i = N_i/(176\,\mathrm{rpm})$, where N_i is rotational speed. Throughout the residual model, $s_\star = 1\,\mathrm{m\,s^{-1}}$.

Define the active set $\mathcal{I} = \{i : \alpha_i = 1\}$ and the unordered pair set $\mathcal{P} = \{\{i, j\} : i < j,\ i, j \in \mathcal{I}\}$. For $n = |\mathcal{I}|$, there are $n(n-1)/2$ valid pairs. The implemented capacity is six fan slots and fifteen pair slots. All item-wise modules share parameters across items of the same type; their feature widths do not change with the active fan count.

The frozen inputs are the M1 case representation, the boundary-corrected single-fan fields, and the configuration-conditioned non-additive reference. We distinguish raw M1 predictions $m_i(q)$ from the boundary-corrected fields $p_i(q)$ used by the residual model. The latter are the outputs of the frozen reference model, whose selected checkpoint applies the boundary correction in inverse-softplus space. The reference $s_{\mathrm{ref},\mathcal{C},h}(\mathbf{x})$ is its output. This is evaluated by the frozen non-additive reference map rather than defined by summing the local priors. For compactness, write $s_{\mathrm{ref}}(q)$ for this reference below.

The background $b_e(q)$ is a separately measured all-fans-off field, aligned to the prediction grid by direct coordinate matching. It is an explicit input, with no implicit zero default. Neither the current multi-fan target field nor its spectrum is used to construct prediction features. The frozen reference, including its boundary and interaction modules, remains fixed during residual training.

A.2.2 CONFIGURATION-LEVEL FAN AND ROOM FEATURES

The eight-dimensional fan descriptor and six-dimensional boundary descriptor are

$$\begin{aligned}\boldsymbol{\phi}_i^{\mathrm{fan}} &= \left[\frac{x_i}{L_x}, \frac{y_i}{L_y}, \frac{z_i}{H}, \frac{D_i}{\sqrt{L_x L_y}}, \frac{\mu_i}{H}, a_i, \alpha_i, \frac{D_i}{H}\right], \\ \boldsymbol{\phi}_i^{\mathrm{B}} &= \left[\frac{x_i}{L_x}, \frac{L_x - x_i}{L_x}, \frac{y_i}{L_y}, \frac{L_y - y_i}{L_y}, \frac{z_i}{H}, \frac{H - z_i}{H}\right].\end{aligned} \tag{12}$$

The room descriptor and room token are

$$\boldsymbol{\phi}^{\mathrm{room}} = [L_x, L_y, H, L_x/L_y, H/L_x, H/L_y, 1], \qquad \mathbf{t}_{\mathrm{room}} = E_{\mathrm{room}}(\boldsymbol{\phi}^{\mathrm{room}}). \tag{13}$$

The first three room entries remain in metres; only the specified ratios are dimensionless. The room token supplies shared geometric context.

Every token encoder in this appendix has the form

$$E(\mathbf{u}) = W_2\,\mathrm{SiLU}(W_1\mathbf{u} + \mathbf{b}_1) + \mathbf{b}_2, \qquad E : \mathbb{R}^{d_{\mathrm{in}}} \to \mathbb{R}^{128}, \tag{14}$$

with hidden width 256. Different encoder names denote separate parameters, and $[\cdot;\cdot]$ denotes concatenation. The frozen M1 representation $\mathbf{h}_i^{\mathrm{M1}} \in \mathbb{R}^{24}$ contains the branch coefficients of the single-fan POD-DeepONet. The fan token $\mathbf{t}_i$ is

$$\begin{aligned}\mathbf{t}_i^{\mathrm{attr}} &= E_{\mathrm{attr}}(\boldsymbol{\phi}_i^{\mathrm{fan}}), \\ \mathbf{t}_i^{\mathrm{prior}} &= E_{\mathrm{prior}}([\mathbf{h}_i^{\mathrm{M1}}; \mathbf{f}_i]), \\ \mathbf{t}_i &= \alpha_i E_{\mathrm{fan}}([\mathbf{t}_i^{\mathrm{attr}}; \mathbf{t}_i^{\mathrm{prior}}; \boldsymbol{\phi}_i^{\mathrm{B}}; \mathbf{t}_{\mathrm{room}}]).\end{aligned} \tag{15}$$

The respective input widths are 8, 31, and 390. In E_{prior}, the seven fan entries retain their supplied values and units, with the activity entry replaced by the authoritative binary mask. This trainable encoder applies no additional input standardization and receives the M1 coefficients without rescaling.

M1 uses its own frozen feature transformation before supplying these coefficients. Let $N_i^{\mathrm{M1}} = 176\max(a_i, 0)$ be its reconstructed RPM value, expressed numerically in revolutions per minute. Its nine case features, in order, are

$$\begin{aligned}\boldsymbol{\chi}_i = \Big[&\frac{D_i}{\sqrt{L_x L_y}}, \frac{H}{D_i}, \frac{D_i}{h_\star}, \frac{z_i}{D_i}, \frac{H - z_i}{D_i}, \\ &\max\left(0, 0.2 - \frac{H - z_i}{D_i}\right), \frac{D_i - D_{\min}}{D_{\max} - D_{\min}}, \frac{N_i^{\mathrm{M1}} - N_{\min}}{N_{\max} - N_{\min}}, \\ &\log\left(1 + \frac{\pi D_i N_i^{\mathrm{M1}}}{60 v_\star}\right)\Big].\end{aligned} \tag{16}$$

Here $h_\star = 1.7\,\mathrm{m}$ and $v_\star = 10\,\mathrm{m\,s^{-1}}$. The saved training ranges are $[D_{\min}, D_{\max}] = [1.2192, 4.2672]\,\mathrm{m}$ and $[N_{\min}, N_{\max}] = [28, 240]$ RPM. M1 recomputes ceiling-to-blade distance as $H - z_i$. Saved case-feature means and scales standardize $\boldsymbol{\chi}_i$ before the frozen $9 \to 128 \to 128 \to 24$ SiLU branch. The checkpoint's dropout probability is 0.03, disabled during frozen evaluation. Neither training ranges nor standardization statistics are refitted on the current configuration.

To evaluate M1 at q, its fixed mean field and 24 POD modes are interpolated using inverse-square distance weights over 12 nearest support points. The spatial feature coordinates are horizontal radial distance divided by D_i, height divided by $h_\star$, and radial distance divided by the fan-to-wall distance along the query direction. The interpolated mean and coefficient-weighted modes pass through softplus. M1 then converts this dimensionless field to air-speed units using its fixed operating-speed calibration, before the frozen boundary correction.

A.2.3 SYMMETRIC PAIR FEATURES AND MASKS

For each active unordered pair, define

$$\overline{D}_{ij} = \frac{D_i + D_j}{2}, \qquad d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \qquad d_e = \sqrt{L_x^2 + L_y^2}, \tag{17}$$

and the normalized horizontal wall proximity

$$w_i = \min\left\{ \frac{x_i}{L_x}, \frac{L_x - x_i}{L_x}, \frac{y_i}{L_y}, \frac{L_y - y_i}{L_y} \right\}. \tag{18}$$

The ten-dimensional pair descriptor is

$$\begin{aligned} \boldsymbol{\phi}_{ij}^{\mathrm{pair}} = \Big[& \frac{|x_i - x_j|}{\overline{D}_{ij}}, \frac{|y_i - y_j|}{\overline{D}_{ij}}, \frac{d_{ij}}{\overline{D}_{ij}}, \frac{d_{ij}}{d_e}, \frac{\max(D_i, D_j)}{\min(D_i, D_j)}, \\ & |a_i - a_j|, \frac{a_i + a_j}{2}, \min(a_i, a_j), \min(w_i, w_j), \frac{w_i + w_j}{2} \Big]. \end{aligned} \tag{19}$$

The corresponding token is

$$\mathbf{t}_{ij} = \alpha_i \alpha_j E_{\mathrm{pair}}([\mathbf{t}_i + \mathbf{t}_j; |\mathbf{t}_i - \mathbf{t}_j|; \boldsymbol{\phi}_{ij}^{\mathrm{pair}}]). \tag{20}$$

Its input width is 266 and its output width is 128. Every component is invariant under endpoint exchange $i \leftrightarrow j$.

The equations describe valid items. In padded computation, the implementation uses safe diameters before evaluating features: active diameters are lower-bounded by $10^{-6}\,\mathrm{m}$, and inactive slots use a finite dummy diameter of $1\,\mathrm{m}$. Inactive M1 coefficients and local priors are replaced by zeros. Fan, pair, and local-item tokens are then masked, and activity masks are applied again to attention gates. Thus inactive items contribute neither values nor normalization mass. Empty fan or pair sets have zero pooled representations and zero masked means.

A.2.4 QUERY-RELATIVE GEOMETRY AND LOCAL PRIOR FEATURES

The nine-dimensional query descriptor and the corresponding query token are

$$\begin{aligned} \boldsymbol{\phi}^q(q) &= \left[\frac{x}{L_x}, \frac{y}{L_y}, \frac{h}{H}, \frac{x}{L_x}, \frac{L_x - x}{L_x}, \frac{y}{L_y}, \frac{L_y - y}{L_y}, \frac{h}{H}, \frac{H - h}{H} \right], \\ \mathbf{t}_q &= E_q(\boldsymbol{\phi}^q(q)). \end{aligned} \tag{21}$$

The repeated coordinates preserve the implemented coordinate-and-boundary encoding; they do not supply independent geometric observations. For each fan, define

$$\begin{aligned} \boldsymbol{\delta}_i(q) &= \left[\frac{x - x_i}{D_i}, \frac{y - y_i}{D_i}, \frac{h - z_i}{D_i} \right], \\ \rho_i(q) &= \frac{\sqrt{(x - x_i)^2 + (y - y_i)^2}}{D_i}, \qquad \zeta_i(q) = \frac{h - z_i}{D_i}. \end{aligned} \tag{22}$$

The local fan descriptor is $[\boldsymbol{\delta}_i(q); p_i(q)/s_\star]$. The seven-dimensional local pair descriptor is

$$\begin{aligned} \boldsymbol{\psi}_{ij}(q) = \Big[& \frac{p_i(q) + p_j(q)}{s_\star}, \frac{|p_i(q) - p_j(q)|}{s_\star}, \frac{p_i(q) p_j(q)}{s_\star^2}, \\ & \min(\rho_i, \rho_j), \max(\rho_i, \rho_j), \frac{|\zeta_i| + |\zeta_j|}{2}, \big||\zeta_i| - |\zeta_j|\big| \Big], \end{aligned} \tag{23}$$

where the arguments of ρ_i and ζ_i are q. This descriptor is symmetric in the pair endpoints. The product of scalar priors encodes their local co-occurrence, not a measured interaction or a velocity inner product.

A.2.5 SCALE-CONDITIONED LOCAL ITEMS AND RETRIEVAL

Each band $b \in \{1, 2, 3, 4\}$ has a learned embedding $\mathbf{e}_b \in \mathbb{R}^8$, which conditions the geometric query token $\mathbf{t}_q$. A shared query adapter constructs

$$\mathbf{t}_{q,b}(q) = E_{\text{adapt}}([\mathbf{t}_q; \mathbf{e}_b; s_{\text{ref}}(q)/s_\star; b_e(q)/s_\star]), \tag{24}$$

with input width 138 and output width 128. The local item encoders combine this query with configuration tokens:

$$\begin{aligned} \mathbf{t}_{i,b}^{\text{loc,fan}}(q) &= \alpha_i E_{\text{loc,fan}}([\mathbf{t}_i; \mathbf{t}_{q,b}; \boldsymbol{\delta}_i(q); p_i(q)/s_\star]), \\ \mathbf{t}_{ij,b}^{\text{loc,pair}}(q) &= \alpha_i \alpha_j E_{\text{loc,pair}}([\mathbf{t}_{ij}; \mathbf{t}_{q,b}; \boldsymbol{\psi}_{ij}(q)]). \end{aligned} \tag{25}$$

The displacement $\boldsymbol{\delta}_i(q)$ is relative to fan i and normalized by D_i. The symmetric descriptor $\boldsymbol{\psi}_{ij}(q)$ combines sums, absolute differences, and products of normalized local priors with query-to-fan distances. Conditioning items on $\mathbf{t}_{q,b}$ allows the content supplied by a source or relation to vary with both location and band. The respective input widths are 260 and 263; both outputs have width 128. Configuration tokens are reused across queries and bands, whereas these local items are recomputed for each query and band.

Let $\tau \in \{\text{fan}, \text{pair}\}$ denote the retrieval branch, a an item, and $k \in \{1, 2, 3, 4\}$ an attention head. The item subscript a is distinct from the operating strength a_i. Each branch has separate affine query, key, and value projections:

$$\begin{aligned} \mathbf{q}_{b,k}^{\tau}(q) &= Q_k^{\tau}(\mathbf{t}_{q,b}(q)), \\ \mathbf{k}_{b,a,k}^{\tau}(q) &= K_k^{\tau}(\mathbf{t}_{a,b}^{\text{loc},\tau}(q)), \\ \mathbf{v}_{b,a,k}^{\tau}(q) &= V_k^{\tau}(\mathbf{t}_{a,b}^{\text{loc},\tau}(q)). \end{aligned} \tag{26}$$

Each head has width 32, and all projections are shared across bands and queries within their branch. Both keys and values therefore depend on the query and band through the local items.

With activity mask $m_a = \alpha_i$ for fan i and $m_a = \alpha_i \alpha_j$ for pair $\{i, j\}$, the score and gate are

$$\ell_{b,a,k}^{\tau}(q) = \frac{\mathbf{q}_{b,k}^{\tau}(q)^\top \mathbf{k}_{b,a,k}^{\tau}(q)}{\sqrt{32}}, \qquad g_{b,a,k}^{\tau}(q) = m_a \sigma(\ell_{b,a,k}^{\tau}(q)). \tag{27}$$

Geometry enters through the query and local-item encoders; the score has no additional additive geometry-bias network. Both branches use the same sigmoid normalization rule, with separate parameters and separate sums:

$$\begin{aligned} G_{b,k}^{\tau} &= \sum_a g_{b,a,k}^{\tau}, \qquad & \mathbf{c}_{b,k}^{\tau} &= \frac{\sum_a g_{b,a,k}^{\tau} \mathbf{v}_{b,a,k}^{\tau}}{1 + G_{b,k}^{\tau}}, \\ \omega_{b,k}^{\tau} &= \frac{G_{b,k}^{\tau}}{1 + G_{b,k}^{\tau}}, \qquad & \mathbf{z}_b^{\tau} &= W_O^{\tau}[\mathbf{c}_{b,1}^{\tau}; \mathbf{c}_{b,2}^{\tau}; \mathbf{c}_{b,3}^{\tau}; \mathbf{c}_{b,4}^{\tau}]. \end{aligned} \tag{28}$$

Query arguments are suppressed in this equation. The output projection W_O^{τ} has no bias. Its output has width 128, and $\boldsymbol{\omega}_b^{\tau} = [\omega_{b,1}^{\tau}, \ldots, \omega_{b,4}^{\tau}]$ has width 4. The constant 1 in the denominator acts as a zero-value null contribution. With one valid item, the pooled head is $g\mathbf{v}/(1 + g)$ and remains score-dependent; with no valid items, both outputs are zero. No attention dropout is applied. For $G_{b,k}^{\tau} > 0$, each aggregate is a gated mean multiplied by $\omega_{b,k}^{\tau}$, retaining a saturating dependence on total gate mass. If all head values have norm at most M, then $\|\mathbf{c}_{b,k}^{\tau}\| \le \omega_{b,k}^{\tau} M$ before output projection.

A.2.6 SHARED DEEPONET DECODER

At each grid location, the decoder receives the 528-channel feature field

$$\mathbf{F}_b(q) = [\mathbf{t}_{q,b}; \mathbf{z}_b^{\text{fan}}; \mathbf{z}_b^{\text{pair}}; \boldsymbol{\omega}_b^{\text{fan}}; \boldsymbol{\omega}_b^{\text{pair}}; \mathbf{t}_{\text{room}}; \mathbf{e}_b]. \tag{29}$$

The DeepONet branch processes the complete horizontal feature plane at a fixed height. It first resamples that plane to a 4×4 sensor grid using bilinear interpolation with aligned endpoints. Sensor endpoints coincide with the supplied prediction-grid endpoints. A shared affine $528 \to 24$

map and GELU act at each sensor. The resulting tensor is flattened in channel-first order into 384 entries and passed through a $384 \to 120 \to 64$ branch network, with GELU between its affine layers. This produces one coefficient vector $\mathbf{c}_b^{\mathrm{dec}}(\mathcal{C}, e, h) \in \mathbb{R}^{64}$ per plane and band.

The trunk T_θ uses only the normalized coordinate $[x/L_x, y/L_y, h/H]$. Its affine widths are $3 \to 64 \to 64 \to 64$, with GELU after the first two layers. Writing $\mathcal{S}_{4\times4}$ for sensor sampling and absorbing the sensor lift into B_θ, we obtain the raw correction $\widetilde{r}_b(q)$ predicted by the decoder for band b at location q

$$\begin{aligned} \mathbf{c}_b^{\mathrm{dec}} &= B_\theta(\mathcal{S}_{4\times4}\mathbf{F}_b), \qquad \mathbf{t}(q) = T_\theta(\bar{q}), \\ \widetilde{r}_b(q) &= \frac{s_\star}{\sqrt{p}} \sum_{j=1}^{p} c_{b,j}^{\mathrm{dec}} t_j(q), \qquad p = 64. \end{aligned} \tag{30}$$

All affine branch, sensor-lift, and trunk layers include biases, but the final inner product has no additional scalar bias. The same coefficient vector is reused across output locations in a plane. All four bands share the sensor lift, branch, and trunk parameters. The learned band query enters the branch feature field and upstream retrieval, not the trunk. The branch thus summarizes a query-conditioned feature field rather than only configuration-level tokens.

A.2.7 Final prediction

DC correction. Let $\bar{\mathbf{t}}_{\mathrm{fan}}$ and $\bar{\mathbf{t}}_{\mathrm{pair}}$ be the means of valid configuration tokens, with an empty-set mean defined as zero. For a grid field v, write $\langle v \rangle$ for its unweighted mean over the complete plane and $\mathrm{RMS}(v) = \sqrt{\langle v^2 \rangle}$. The DC head takes the 389-dimensional vector

$$\begin{aligned} \mathbf{u}_{\mathrm{DC}} = \Big[& \bar{\mathbf{t}}_{\mathrm{fan}}; \bar{\mathbf{t}}_{\mathrm{pair}}; \mathbf{t}_{\mathrm{room}}; h/H; \\ & \langle s_{\mathrm{ref}} \rangle / s_\star; \mathrm{RMS}(s_{\mathrm{ref}})/s_\star; \langle b_e \rangle / s_\star; \mathrm{RMS}(b_e)/s_\star \Big]. \end{aligned} \tag{31}$$

It uses affine widths $389 \to 64 \to 32 \to 1$, with SiLU after the first two layers, and produces $c_{\mathrm{DC}} = s_\star D_{\mathrm{DC}}(\mathbf{u}_{\mathrm{DC}})$. This correction is constant within each horizontal plane. Its field statistics use the frozen reference and permitted background, never the target residual mean.

Grid and boundary convention. On measurement dataset, the spatial decoder and projection operate on complete Cartesian planes with $\Delta x = \Delta y = 0.35\,\mathrm{m}$. Arrays have shape (N_y, N_x) and row-major ordering, with x varying along the final axis. D2/D3 planes have shape $(15, 12)$; D4 planes have shape $(31, 20)$. The grid origins (x_0, y_0) are $(0.385, 0.435)\,\mathrm{m}$ for D2, $(0.235, 0.435)\,\mathrm{m}$ for D3, and $(0.4125, 0.45)\,\mathrm{m}$ for D4. The evaluation heights are 0.1, 0.6, 1.1, and $1.7\,\mathrm{m}$; the D4 two-fan training support uses the first two heights.

Fourier projection. For a raw plane v, the operator $\mathcal{E}$ concatenates v with its horizontal reversal, then concatenates the result with its vertical reversal. This repeated-endpoint even extension has shape $(2N_y, 2N_x)$. It reflects the sampled-grid edges, which need not coincide with physical room walls. Let $\mathcal{R}$ crop the inverse transform to the original $N_y \times N_x$ block. The band operator is

$$P_b v = \mathcal{R}\mathcal{F}^{-1}[H_b \odot \mathcal{F}(\mathcal{E}v)], \tag{32}$$

using orthonormal two-dimensional real FFT and inverse FFT operations. Complex coefficients are retained during filtering. For radial spatial frequency $\nu = \sqrt{\nu_x^2 + \nu_y^2}$ in cycles per metre, the fixed masks select

$$\begin{aligned} H_1 &: \ 0 < \nu < 1/4, & H_2 &: \ 1/4 \le \nu < 1/2, \\ H_3 &: \ 1/2 \le \nu \le 1/1.4, & H_4 &: \ 1/1.4 < \nu \le 1/0.7. \end{aligned} \tag{33}$$

The frequency grid uses transform lengths $2N_x$ and $2N_y$ with the original physical spacing. The masks exclude DC and radial frequencies above $1/0.7$ cycles per metre; they do not partition every grid mode.

Readout The candidate residual and final prediction are

$$r_{\mathrm{cand}}(q) = c_{\mathrm{DC}}(\mathcal{C}, e, h) + \sum_{b=1}^{4} (P_b \widetilde{r}_b)(q), \tag{34}$$

$$\widehat{s}(q) = \varepsilon \,\mathrm{softplus}\left(\frac{s_{\mathrm{ref}}(q) + r_{\mathrm{cand}}(q)}{\varepsilon}\right), \qquad \varepsilon = 0.01\,\mathrm{m\,s^{-1}}, \tag{35}$$

where $\mathrm{softplus}(u) = \log(1 + \exp u)$ is evaluated stably. The readout is smooth and approaches the identity when $v = s_{\mathrm{ref}} + r_{\mathrm{cand}}$ is positive and large relative to ε. A zero candidate residual returns the softplus-transformed reference, rather than the reference exactly. Since this nonlinear map can mix frequencies, the zero-mean and frequency support constraints apply to the projected candidate components before readout, not to the final physical increment $\widehat{s} - s_{\mathrm{ref}}$.

A.2.8 Module dimensions, parameter sharing, and execution

Table 3 summarizes all trainable token encoders. Each encoder is shared across the relevant items, queries, and bands, while different rows have independent parameters.

Table 3: Token-encoder dimensions. Every row uses a hidden width of 256, SiLU, and an output width of 128.

MODULE	INPUT WIDTH	INPUT COMPONENTS
E_{attr}	8	Fan attributes
E_{prior}	31	M1 coefficients (24), fan descriptor (7)
E_{room}	7	Room dimensions and aspect ratios
E_{fan}	390	Attribute, prior, boundary, room features
E_{pair}	266	Token sum, absolute difference, pair features
E_q	9	Query coordinates and boundary distances
E_{adapt}	138	Query (128), band (8), reference, background
$E_{\mathrm{loc,fan}}$	260	Fan token, band query, geometry, scalar prior
$E_{\mathrm{loc,pair}}$	263	Pair token, band query, local pair features

The residual core has 883,761 trainable parameters, including 75,216 in the DeepONet decoder. This count excludes the frozen M1 and reference networks. Token encoders, band embeddings, retrieval projections, the decoder, and the DC head are trainable; Fourier masks and the final readout operation are fixed. The last affine layers of the DeepONet branch and DC head are initialized with zero weights and biases, giving a zero initial candidate residual.

Inference proceeds as follows:

1. Evaluate or retrieve the frozen M1 coefficients, boundary-corrected single-fan fields, non-additive reference, and aligned background.
2. Construct room, fan, and symmetric pair tokens, together with their valid-item masks.
3. At every grid location and band, construct the adapted query and local items, retrieve fan and pair context, and assemble the feature plane.
4. Encode each feature plane at the fixed sensor grid, evaluate the shared coordinate trunk, and decode the four signed raw fields.
5. Apply the fixed band operators, add the DC correction and reference, and apply the scaled-softplus readout.

No parameters are fitted at inference. Shared item encoders, symmetric pair features, and masked set aggregation make the residual independent of source enumeration. Invariance of the complete prediction also requires the frozen reference to obey the same symmetry. Pair representations do not impose pairwise additivity on the physical response or establish identifiable pair contributions.

These feature definitions and grid conventions specify the real ceiling-fan application. The source descriptors, reference construction, and periodic grid conventions of the controlled two-dimensional experiments are specified separately.

A.3 TRAINING AND BASELINE DETAILS

We train the proposed model with the full-feature DeepONet branch and coordinate-only trunk described in the main text. This is the D0 configuration in the experiment records. Feature definitions, network dimensions, and frequency masks are specified in the preceding architecture appendix.

A.3.1 DATA AND FROZEN PRIORS.

Each training example is a complete configuration–height plane. Measurements are split by dataset and configuration, keeping all heights together, with split seed 1734. Training contains 112 planes from 29 configurations, including four two-fan D4 support planes; validation contains 28 planes from seven configurations. D4 losses and metrics exclude missing and flagged imputed nodes. D2/D3 supervise all cells of their archived processed grids, whose files have no per-node imputation flags. All-fans-off measurements provide background inputs and are excluded as residual-training targets.

M1 was pretrained on the separate D1 single-fan dataset. The measurement reference uses a separately fitted non-additive composition model. Both remain frozen, including their parameters, POD modes, and normalization statistics. Reference fitting excluded the residual validation configurations from gradient updates, while using them for checkpoint selection. For each residual run, all trainable encoders, retrieval modules, the DeepONet decoder, and the DC head are fitted jointly from scratch. Their initial spatial and DC corrections are zero. The same frozen priors are shared across seeds and comparison methods.

A.3.2 TRAINING OBJECTIVE AND OPTIMIZATION.

Following the main-text notation, the pre-readout field is v_t, where $\delta_{t,b} = P_b \widetilde{r}_{t,b}$ projects the raw decoder correction. For plane t, let $\mathcal{G}_t$ denote the complete grid and $\mathcal{M}_t$ its supervised nodes. We minimize

$$\begin{aligned}\mathcal{L}_t = \frac{1}{|\mathcal{M}_t|} \sum_{q \in \mathcal{M}_t} \rho_{0.1}\left(\frac{\widehat{s}_t(q) - y_t(q)}{s_\star}\right) \\ + \frac{0.01}{|\mathcal{G}_t|} \sum_{q \in \mathcal{G}_t} \left[\max\left(0, -\frac{v_t(q)}{s_\star}\right)\right]^2,\end{aligned} \tag{36}$$

where $\rho_{0.1}$ is the Huber loss with threshold 0.1. The speed scale is $s_\star = 1\,\mathrm{m\,s^{-1}}$ for measurements and 1 in numerical units. Planes receive equal weight. The second term penalizes negative pre-readout predictions without requiring labels at unsupervised nodes. No target-spectrum or representation-regularization loss is used. The residual encoders, band embeddings, both attention modules, DeepONet decoder, and DC head are optimized jointly. No bandwise target spectra or additional decomposition losses are used.

We use AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, moment coefficients $(0.9, 0.999)$, and gradient-norm clipping at 1. Training lasts 100 epochs, with one complete plane per update and shuffled plane order. There is no learning-rate schedule, early stopping, query subsampling, or token dropout. This gives $6{,}720$ updates for measurements and $7{,}560$ per numerical regime. Validation is evaluated after every epoch. Measurement checkpoints minimize pooled supervised-node RMSE. Numerical checkpoints minimize configuration-average MAE, requiring an improvement greater than 10^{-12}. Both rules retain the earlier checkpoint on ties.

A.3.3 SETUP FOR THE NUMERICAL EXPERIMENTS

A separate model is fitted in each of the six regimes, using 126 training, 42 validation, and 84 high-count configurations. Every node of each 32×32 field is supervised. The numerical reference is the sum of nonnegative quadratic scalar priors,

$$p_i(a_i, q) = \max\{a_i c_{i,1}(q) + a_i^2 c_{i,2}(q), 0\}, \qquad s_{\mathrm{ref}}(q) = \sum_i p_i(a_i, q). \tag{37}$$

Their coefficients are fitted only to the 36 single-source training fields, with ridge coefficient 10^{-10}, and then frozen. Background is zero. Known forcing descriptors populate the source features. The

24-dimensional prior slot contains ten complex Fourier coefficients plus source direction, normalized width, and amplitude. No measurement-domain M1 or composition checkpoint is transferred. The numerical solver is periodic, but the residual model retains its ordinary geometric features and even-extension projector. Band cutoffs are unchanged numerically, using spacing $2\pi/32$ in nondimensional coordinates.

A.3.4 BASELINE INPUTS AND PREDICTION.

External operators receive the same permitted source descriptors, geometry, priors, reference, and background through an invariant grid adapter trained separately for each model. At each query, shared source and symmetric-pair MLPs produce 24 features per item. Masked sums are divided by 6 and 15, respectively. Ten additional coordinate, room, reference, background, and count channels yield a 58-channel input field. Current query targets and observation masks are excluded from this adapter.

Table 4 summarizes the prediction paths and selected measurement architectures. Each field in the second column passes through the common readout $\mathcal{R}_\varepsilon(v) = \varepsilon\,\mathrm{softplus}(v/\varepsilon)$, with $\varepsilon = 0.01$ in the applicable units. The operators retain their own spatial backbones. The proposed model's retrieval modules, DC head, and fixed output-band projectors are specific to our method. Residual-output layers are initialized to zero.

Table 4: Baseline adaptations. Architecture are the selected or preset measurement configurations. Here $g_\theta(\mathbf{x})$ denotes a learned scalar-valued field: a complete pre-readout prediction for the Direct models and a correction for the other learned models.

BASELINE	BEFORE READOUT	ARCHITECTURE
DeltaPhi-DeepONet	$\widetilde{y}_{\mathrm{aux}} + g_\theta$	Same DeepONet backbone, with retrieved auxiliary inputs
Reference-GNN	$s_{\mathrm{ref}} + g_\theta$	Complete source graph; state width 96; two edge–node–global message-passing blocks
Reference-HGNN	$s_{\mathrm{ref}} + g_\theta$	All nonempty active-source subsets as hyperedges; state width 96; two node–hyperedge–node–global blocks
Reference-GAOT	$s_{\mathrm{ref}} + g_\theta$	8×8 latent grid with 32 channels and 2×2 patches;
Reference-GNOT	$s_{\mathrm{ref}} + g_\theta$	Width 96; three blocks; four heads; four experts per MoE sublayer
Reference-CViT	$s_{\mathrm{ref}} + g_\theta$	32×32 input grid with 4×4 patches; width 128; four heads; three encoder blocks and one decoder block
Reference-Transolver++	$s_{\mathrm{ref}} + g_\theta$	Width 128; six blocks; eight heads; 32 slices
Reference-FNO	$s_{\mathrm{ref}} + g_\theta$	Width 48; four blocks; modes 8 per axis; padding 4
Reference-GINO	$s_{\mathrm{ref}} + g_\theta$	16×16 latent grid; FNO width 32 and four layers; configured $n_{\mathrm{modes}} = (12, 12)$; input/output GNO radii 0.20; kernel MLP hidden widths $(64, 64)$

A.3.5 COMPARISON PROTOCOLS AND REPORTING.

All baselines use the loss function in Equation 36, the stated optimizer settings, and 100 formal training epochs. Architecture and checkpoint selection use configuration-average validation MAE, averaging heights within each measured configuration. Formal seeds are $1234, 17, 42, 2026, 3407$. They follow the same split, training order, and pooled-RMSE checkpoint rule. We report mean and sample standard deviation across runs, with paired comparisons restricted to shared seeds. Fitting uses an NVIDIA RTX 5090 with AMP and TF32 disabled. Frozen priors and all learned parameters remain fixed at inference.

A.4 FULL PERFORMANCE METRICS

This section reports the complete quantitative results underlying the comparisons presented in the main text. We provide results separately for the measurement and the PDE-derived numerical experiments, as the two datasets differ in their data-generation process, field structure, and extrapolation setting. The full baseline comparisons are reported first, followed by ablations of the main architectural components of ScaleComposer. Trained-model results are reported as means $\pm$ sample standard deviations across five random seeds. Bold values indicate the best reported mean in each column. Lower MAE and RMSE and higher PCC indicate better performance.

A.4.1 COMPARISONS TO BASELINES

Table 1 reports the full comparison on the measurement. The baselines cover complementary mechanisms for representing and aggregating configuration-dependent information, including prior retrieval, graph-based message passing, attention-based feature aggregation, and Fourier-based operators. This comparison complements the source-count-specific results in the main text by providing all evaluation metrics for the same test configurations. Table 5 reports the computational efficiency comparison with baseline methods on the measurements.

Table 6 gives the corresponding comparison on the PDE-derived numerical experiments. These experiments evaluate the same source-count extrapolation setting under controlled numerical regimes, allowing the behavior of different representation mechanisms to be examined independently of measurement-specific effects. Differences between the numerical and measurement results are discussed in the main text. MAE and RMSE are dimensionless here. For each seed, metrics are averaged over 84 test configurations within each of six regimes, then equally over the six regimes.

Table 5: Computational efficiency comparison with baseline methods on the measurements.

CATEGORY	MODEL	#PARAMS.(M)	FLOPs (G)	TRAINING (epoch/s)	LATENCY (ms)
Prior retrieval	**DeltaPhi**-DeepONet	0.998	0.327	0.947 ± 0.042	5.132
Graph-based message passing or aggregation	Ref+**GNN**	0.839	4.443	0.637 ± 0.047	6.080
	Ref+**HGNN**	0.940	14.596	0.548 ± 0.009	7.498
	Ref+**GAOT**	1.135	0.523	0.471 ± 0.007	7.422
Attention-based feature aggregation	Ref+**GNOT**	2.73	3.558	0.315 ± 0.020	6.738
	Ref+**CViT**	0.909	0.512	0.762 ± 0.027	3.891
	Ref+**Transolver++**	1.062	1.551	0.452 ± 0.038	8.827
Fourier based	Ref+**FNO**	2.388	0.171	0.995 ± 0.012	3.155
	Ref+**GINO**	0.732	0.616	0.056 ± 0.001	152.651
Proposed method	**ScaleComposer**	0.884	14.548	0.492 ± 0.033	8.640
	ScaleComposer w/o Ref	0.884	14.548	0.471 ± 0.019	8.448

FLOPs are partial profiler counts in billions per forward pass. Training throughput excludes validation and I/O. Latency is measured in milliseconds per 31×20 plane with six sources and excludes frozen-prior construction.

A.4.2 ABLATION RESULTS

We further isolate the contributions of the main design choices in ScaleComposer. The ablations examine three aspects of the model: incorporation of the reference field, the mechanism used to aggregate source and interaction representations, and the spectrally constrained residual reconstruction. As above, we report the measurement and numerical benchmarks separately to distinguish effects that are consistent across datasets from those that depend on the underlying field-generation process. These results complement the baseline comparison by identifying which components of ScaleComposer contribute to its performance.

Table 7 reports the ablation results on the measurements. Table 8 reports the same architectural ablations on the PDE-derived numerical experiments. Comparing the two benchmarks helps identify which components provide robust benefits under source-count extrapolation and which are sensitive to differences between measured airflow fields and the controlled numerical system. The numerical ablations therefore serve primarily as a diagnostic test of the proposed inductive biases.

Table 6: Full comparison with baseline methods on the PDE-derived numerical experiments.

CATEGORY	MODEL	PERFORMANCE ON NUMERICAL EXPERIMENTS		
		MAE ↓	RMSE ↓	PCC ↑
Prior retrieval	**DeltaPhi**-DeepONet	0.047 ± 0.004	0.068 ± 0.003	0.730 ± 0.004
Graph-based message passing or aggregation	Ref+**GNN**	0.079 ± 0.001	0.103 ± 0.001	$\mathbf{0.850 \pm 0.000}$
	Ref+**HGNN**	0.097 ± 0.017	0.125 ± 0.017	0.813 ± 0.010
	Ref+**GAOT**	0.080 ± 0.002	0.108 ± 0.002	0.828 ± 0.001
Attention-based feature aggregation	Ref+**GNOT**	0.053 ± 0.001	0.069 ± 0.002	0.836 ± 0.005
	Ref+**CViT**	0.067 ± 0.002	0.093 ± 0.002	0.830 ± 0.003
	Ref+**Transolver++**	0.072 ± 0.006	0.095 ± 0.007	0.687 ± 0.033
Fourier-based	Ref+**FNO**	0.054 ± 0.002	0.073 ± 0.003	0.749 ± 0.010
	Ref+**GINO**	0.054 ± 0.005	0.072 ± 0.006	0.705 ± 0.039
Proposed method	**ScaleComposer**	0.070 ± 0.000	0.095 ± 0.001	0.839 ± 0.003
	ScaleComposer w/o Ref	$\mathbf{0.047 \pm 0.000}$	$\mathbf{0.066 \pm 0.001}$	0.798 ± 0.009

Table 7: Ablation study on the measurements.

ABLATIONS	VARIANTS	PERFORMANCE ON MEASUREMENTS		
		MAE (m/s) ↓	RMSE (m/s) ↓	PCC ↑
Use of the prior field	Reference field (Ref)	0.233	0.285	0.820
	Direct-DeepONet	0.426 ± 0.033	0.549 ± 0.027	0.457 ± 0.060
Attention mechanism	Single joint attention	0.192 ± 0.009	0.253 ± 0.009	0.819 ± 0.016
	Set Transformer	0.197 ± 0.007	0.255 ± 0.008	0.821 ± 0.011
Fourier projection	DC-only	0.227 ± 0.009	0.281 ± 0.006	0.800 ± 0.007
	w/o DC	0.218 ± 0.002	0.268 ± 0.004	0.839 ± 0.007
	w/o band restriction	0.202 ± 0.012	0.262 ± 0.012	0.805 ± 0.018
Proposed method	**ScaleComposer**	$\mathbf{0.183 \pm 0.007}$	$\mathbf{0.241 \pm 0.006}$	$\mathbf{0.842 \pm 0.007}$
	ScaleComposer w/o Ref	0.301 ± 0.019	0.481 ± 0.027	0.541 ± 0.077

A.5 Band-wise errors under source-count extrapolation

We compared the frozen reference field without residual reconstruction, the proposed framework, and its retrained control without per-branch band restriction in all six numerical regimes. The control retained four conditioned branches and parameter count, but replaced each projector P_b with the common union $P_U = \sum_{b=1}^{4} P_b$. Both models used the same seeds, the same low-source data, 100 epochs, and validation-MAE checkpoint selection. For each complete 32×32 field, we decomposed the final prediction error using the model's even-extension projectors:

$$E_b = |\mathcal{G}|^{-1} \, \|P_b(\widehat{s} - y)\|_2^2 . \tag{38}$$

Errors are averaged equally over configurations within each regime (see Table 9). There are $40, 30, 12, 2$ evaluation configurations per regime at $N = 3, 4, 5, 6$. The band boundaries are $0, 0.25, 0.5, 1/1.4, 1/0.7$ cycles per non-dimensional coordinate unit. DC and the remaining tail complete the MSE budget.

Reference error increased in every band from three to six sources, with 8.76-fold growth in DC and 5.81-fold growth in B_1, compared with 2.78-fold growth in B_4. DC and B_1 accounted for 97.7% of MSE reduction by the proposed model at six sources. The principal degradation and correction therefore concerned mean and coarse-scale structure.

Removing band restriction raised B_2 MSE by 24.0–36.3%, placing it 6.2–11.1% above the reference at every source count. Although B_1 and B_4 errors decreased, total MSE increased by 1.3–8.3%. These results are consistent with band restriction limiting damaging B_2 corrections, with tradeoffs in other bands.

Table 8: Ablation study on the PDE-derived numerical experiments.

ABLATIONS	VARIANTS	PERFORMANCE ON NUMERICAL EXPERIMENTS		
		MAE ↓	RMSE ↓	PCC ↑
Use of the prior field	Reference field (Ref)	0.101	0.129	0.827
	Direct-DeepONet	0.090 ± 0.073	0.149 ± 0.159	0.509 ± 0.226
Attention mechanism	Single joint attention	0.071 ± 0.001	0.096 ± 0.001	0.837 ± 0.004
	Set Transformer	0.072 ± 0.001	0.099 ± 0.001	0.828 ± 0.002
Fourier projection	DC-only	0.074 ± 0.001	0.107 ± 0.001	0.827 ± 0.000
	w/o DC	0.103 ± 0.001	0.123 ± 0.000	0.837 ± 0.002
	w/o band restriction	0.069 ± 0.001	0.097 ± 0.001	0.830 ± 0.002
Proposed method	**ScaleComposer**	0.070 ± 0.000	0.095 ± 0.001	$\mathbf{0.839 \pm 0.003}$
	ScaleComposer w/o Ref	$\mathbf{0.047 \pm 0.000}$	$\mathbf{0.066 \pm 0.001}$	0.798 ± 0.009

Table 9: Band-wise MSE in units of 10^{-3} (non-dimensional speed squared). Each cell lists the reference field without residual reconstruction, the proposed framework, and its retrained control without per-branch band restriction from top to bottom.

SOURCES	MODEL	DC	BAND B_1	BAND B_2	BAND B_3	BAND B_4	TAIL
3	Ref	5.2408	2.6902	0.9466	0.3755	0.1693	0.0107
	ScaleComposer	1.2740	1.5342	0.7716	0.3727	0.1718	0.0111
	w/o band restriction	1.3339	1.5337	1.0518	0.3806	0.1697	0.0106
4	Ref	13.4230	5.7191	1.6093	0.6726	0.2861	0.0161
	ScaleComposer	6.1600	3.9690	1.3582	0.6760	0.2911	0.0166
	w/o band restriction	6.3236	3.8477	1.7526	0.6831	0.2858	0.0161
5	Ref	26.2818	9.7945	2.2117	1.0248	0.3801	0.0204
	ScaleComposer	15.5648	7.5740	1.9124	1.0270	0.3875	0.0209
	w/o band restriction	15.8327	7.2782	2.3759	1.0391	0.3794	0.0204
6	Ref	45.9221	15.6247	2.9002	1.5342	0.4707	0.0242
	ScaleComposer	31.2306	12.9902	2.4845	1.5278	0.4820	0.0245
	w/o band restriction	31.6961	12.5331	3.0805	1.5528	0.4695	0.0242

Fine-scale reference errors ($B_3 + B_4$) were higher in the strong than weak regimes, but persisted in the matched advection-off controls. High overlap reduced their fraction of non-DC error at six sources. These findings do not support a uniform fine-scale failure mechanism.

A.6 NOMENCLATURE

Problem, configurations, and observations

e, $\Omega \subset \mathbb{R}^d$	Physical environment, spatial domain, and spatial dimension d.
$\mathcal{Z}$, $\mathbf{z}_i$	Source-token space and source descriptor $\mathbf{z}_i = (\mathbf{r}_i, a_i, \boldsymbol{\eta}_i)$.
$\mathbf{r}_i$, $\boldsymbol{\eta}_i$	Source location and additional source or geometric attributes.
$\mathfrak{C}_e$, $\mathcal{C}$	Admissible configuration space and finite unordered source configuration in environment e.
$\mathcal{I}$, $\mathcal{P}$	Active-fan index set and unordered active-pair set; $\mathcal{P} = \{\{i, j\} : i < j,\ i, j \in \mathcal{I}\}$.
n, N, N_s	Active-source count; $n = \lvert\mathcal{I}\rvert$ and $\lvert\mathcal{P}\rvert = n(n-1)/2$. N and N_s occur in numerical reporting.
$\mathcal{D}^{(1)}_{\text{prior}}$	Single-source data used to construct the prior.
$\mathcal{D}^{(\leq 2)}_e$	Target-environment supervision with at most two active sources, including background observations.
$\mathcal{D}_{\leq 2}$	Dataset of configuration–target pairs defined in Equation 3.
m, N_{train}	Configuration-example index and number of entries in $\mathcal{D}_{\leq 2}$.
$\overline{y}_{\mathcal{C}_m}$	Sampled scalar-speed targets for configuration $\mathcal{C}_m$.
$\widehat{\mathcal{F}}_{\theta,e}$	Learned configuration-to-field predictor with parameters θ in environment e.
s, y, $\widehat{s}$	Target scalar-speed field, observed target, and model prediction. Targets average instantaneous speed over time.
$v_{\mathcal{C},h,p,t}$	Instantaneous measured scalar speed; p indexes measurement location and t the retained sample.
$y_{\mathcal{C},h,p}$	Measured target: the mean of 100 retained scalar-speed samples for a configuration, height, and location.
$\mathcal{G}_t$, $\mathcal{M}_t$	Complete prediction grid and supervised-node subset for training plane t.

	Geometry and source attributes
L_x, L_y, H	Room width, length, and height along the physical coordinate axes; lengths are in metres.
$q = (\mathbf{x}, h)$	Query position; $\mathbf{x} = (x, y)$ is horizontal position and h is height above the floor.
$\bar{q}$	Normalized query coordinates $(x/L_x, y/L_y, h/H)$.
x_i, y_i, z_i	Fan-center coordinates; z_i is blade height above the floor.
D_i, μ_i	Fan diameter and ceiling-to-blade mounting distance.
N_i, a_i, α_i	Fan speed in rpm, normalized strength $a_i = N_i/(176\,\mathrm{rpm})$, and binary activity indicator.
$\mathbf{f}_i$	Seven-entry fan descriptor $[a_i, \alpha_i, D_i, z_i, \mu_i, x_i, y_i]$.
$\overline{D}_{ij}$, d_{ij}, d_e	Mean pair diameter, horizontal fan separation, and horizontal room diagonal.
w_i	Minimum normalized horizontal distance from fan i to a room boundary.
$\boldsymbol{\delta}_i(q)$	Three-component query position relative to the fan center, normalized by D_i.
$\rho_i(q)$, $\zeta_i(q)$	Horizontal query-to-fan distance and signed vertical offset, each normalized by D_i.
	Frozen priors and speed scales
M1	Frozen single-fan POD-DeepONet with 24 fixed spatial modes in the measurement model.
$\mathbf{h}_i^{\mathrm{M1}}$	24-dimensional vector of frozen M1 branch coefficients.
$m_i(q)$, $p_i(q)$	Raw M1 single-fan prediction and boundary-corrected single-fan speed field.
$p_i(a_i, q)$	Numerical single-source prior: $\max\{a_i c_{i,1}(q) + a_i^2 c_{i,2}(q), 0\}$ with frozen fitted coefficients.
$s_{\mathrm{ref}}(q)$	Frozen reference field: learned non-additive composition for measurements; $\sum_i p_i(a_i, q)$ numerically.
$b_e(q)$	Background speed field; measured with all fans off, and zero in the numerical benchmark.
$s_\star$	Speed normalization: $1\,\mathrm{m\,s^{-1}}$ for measurements and 1 in numerical units.
$\boldsymbol{\chi}_i$, N_i^{M1}	Nine-entry M1 feature vector and reconstructed rotational speed $176\max(a_i, 0)$ in rpm.
$h_\star$, $v_\star$	M1 feature scales: $1.7\,\mathrm{m}$ and $10\,\mathrm{m\,s^{-1}}$, respectively.
$D_{\min}, D_{\max}, N_{\min}, N_{\max}$	Frozen M1 training-range endpoints for diameter and rotational speed.

Descriptors, tokens, and local items

$E_\cdot$	Named trainable encoders or adapters. Token encoders have output width 128; distinct names denote distinct parameters.
$W_1, W_2, \mathbf{b}_1, \mathbf{b}_2$	Affine weights and biases of the token encoder; d_{in} denotes its input width.
$\boldsymbol{\phi}_i^{\mathrm{fan}}$, $\boldsymbol{\phi}_i^{\mathrm{B}}$	Fan and boundary descriptors of widths 8 and 6.
$\boldsymbol{\phi}^{\mathrm{room}}$	Seven-entry room descriptor; its first three entries retain physical lengths.
$\boldsymbol{\phi}_{ij}^{\mathrm{pair}}$	Ten-entry symmetric pair descriptor.
$\boldsymbol{\phi}^q(q)$	Nine-entry normalized coordinate-and-boundary query descriptor.
$\boldsymbol{\psi}_{ij}(q)$	Seven-entry symmetric local pair descriptor combining priors and query-relative geometry.
$\mathbf{t}_i^{\mathrm{attr}}$, $\mathbf{t}_i^{\mathrm{prior}}$	128-dimensional trainable tokens encoding fan attributes and frozen-prior information.
$\mathbf{t}_i$, $\mathbf{t}_{ij}$	128-dimensional configuration-level fan and symmetric pair tokens, reused across queries and bands.
$\mathbf{t}_{\mathrm{room}}$	128-dimensional room token defined by $E_{\mathrm{room}}(\boldsymbol{\phi}^{\mathrm{room}})$.
$\mathbf{t}_q$	128-dimensional geometric query token.
$\mathbf{e}_b$	Learned eight-dimensional embedding for frequency band b.
$\mathbf{t}_{q,b}$	128-dimensional query token adapted using band, reference, and background information.
$\mathbf{t}_{i,b}^{\mathrm{loc,fan}}$	128-dimensional local fan item, conditioned on query location and band.
$\mathbf{t}_{ij,b}^{\mathrm{loc,pair}}$	128-dimensional local pair item, conditioned on query location and band.
$\bar{\mathbf{t}}_{\mathrm{fan}}$, $\bar{\mathbf{t}}_{\mathrm{pair}}$	Means of valid configuration tokens; empty-set means are zero.

Indexing and elementary operations

i, j	Source indices; $i < j$ selects an unordered pair. The decoder also uses j as a basis index.
b, τ, a, k	Band, item type, item index, and head index. Here $b, k \in \{1, 2, 3, 4\}$ and $\tau \in \{\mathrm{fan}, \mathrm{pair}\}$.
$[\cdot;\cdot]$, $\odot$	Feature concatenation and elementwise multiplication.
$\sigma(\cdot)$	Logistic sigmoid used in attention gates; distinct from Gaussian width σ in the numerical benchmark.

	Separate multi-head cross-attention
$Q_k^\tau, K_k^\tau, V_k^\tau$	Affine query, key, and value projections, shared across queries and bands within item type τ.
$\mathbf{q}_{b,k}^\tau$, $\mathbf{k}_{b,a,k}^\tau$, $\mathbf{v}_{b,a,k}^\tau$	Attention query, key, and value vectors, each of width $d_{\mathrm{head}} = 32$.
m_a	Item activity mask: α_i for a fan and $\alpha_i\alpha_j$ for a pair.
$\ell_{b,a,k}^\tau$	Scaled dot-product attention score.
$g_{b,a,k}^\tau$	Masked sigmoid gate $m_a\sigma(\ell_{b,a,k}^\tau)$.
$G_{b,k}^\tau$, $\omega_{b,k}^\tau$	Total gate mass $\sum_a g_{b,a,k}^\tau$ and normalized mass $G_{b,k}^\tau/(1+G_{b,k}^\tau)$.
$\mathbf{c}_{b,k}^\tau$	32-dimensional pooled head vector; weighted values are normalized by $1 + G_{b,k}^\tau$.
W_O^τ, $\mathbf{z}_b^\tau$	Bias-free projection of concatenated heads and its 128-dimensional attention output.
$\boldsymbol{\omega}_b^\tau$	Four-dimensional vector of normalized gate masses.
	Shared decoder and DC correction
$\mathbf{F}_b(q)$	528-channel feature field supplied to the shared decoder.
$\mathcal{S}_{4\times4}$	Resampling of a complete feature plane to the decoder sensor grid.
B_θ, T_θ	Shared DeepONet branch and coordinate-only trunk.
$\mathbf{c}_b^{\mathrm{dec}}$, $\mathbf{t}(q)$	Plane-level decoder coefficient vector and spatial basis vector, each with $p = 64$ entries.
$\widetilde{r}_b(q)$	Raw decoded correction before Fourier projection.
$\delta_b(q)$	Band-projected correction $P_b\widetilde{r}_b(q)$; the training appendix also writes $\delta_{t,b}$.
$\mathbf{u}_{\mathrm{DC}}$, D_{DC}	389-dimensional DC-head input and the trainable DC-head network.
c_{DC}	Signed correction constant within each horizontal plane; $c_{\mathrm{DC}} = s_\star D_{\mathrm{DC}}(\mathbf{u}_{\mathrm{DC}})$.
$r_{\mathrm{cand}}(q)$	Candidate residual $c_{\mathrm{DC}} + \sum_b \delta_b(q)$ before readout.
$v(q)$, $v_t(q)$	Pre-readout field $s_{\mathrm{ref}}(q) + r_{\mathrm{cand}}(q)$, with t identifying a training plane.
$\mathcal{R}_\varepsilon$, ε	Scaled-softplus readout $\varepsilon\log(1 + \exp(v/\varepsilon))$; $\varepsilon = 0.01$ in the applicable speed units.

Fourier projections and scale-resolved errors

N_x, N_y	Grid sizes along horizontal axes; a plane array has shape (N_y, N_x).
$\Delta x, \Delta y$	Model-grid spacing: $0.35\,\mathrm{m}$ for processed measurement planes and $2\pi/32$ numerically.
$\mathcal{E}$, $\mathcal{R}$	Repeated-endpoint even extension and cropping back to the original plane.
$\mathcal{F}$, $\mathcal{F}^{-1}$	Orthonormal two-dimensional real Fourier transform and its inverse.
ν_x, ν_y, ν	Horizontal and radial spatial frequencies; $\nu = \sqrt{\nu_x^2 + \nu_y^2}$, in cycles per metre or per nondimensional coordinate unit.
H_b, P_b	Fixed frequency mask and projector for band b; $P_b = \mathcal{R}\mathcal{F}^{-1}[H_b \odot \mathcal{F}\mathcal{E}(\cdot)]$.
B_1, B_2, B_3, B_4	Four frequency bands with boundaries 0, $1/4$, $1/2$, $1/1.4$, and $1/0.7$; endpoint conventions follow Equation 33.
P_U	Union-band projector $\sum_{b=1}^{4} P_b$; retains the same total support while removing dedicated per-band support.
E_b	MSE of projected final-prediction error: $\lvert\mathcal{G}\rvert^{-1}\lVert P_b(\widehat{s} - y)\rVert_2^2$.
DC, tail	Zero-frequency mean component and the remaining high-frequency modes outside the four bands.
$\langle v \rangle$, $\mathrm{RMS}(v)$	Unweighted complete-plane mean and root mean square $\sqrt{\langle v^2 \rangle}$.

Training, metrics, and model labels

$\mathcal{L}_t$, t	Loss for training plane t; the training objective averages these losses uniformly over planes.
$\rho_{0.1}$, λ_{neg}	Huber loss with threshold 0.1 on normalized error and nonnegativity penalty weight 0.01.
MAE, RMSE, MSE	Mean absolute error, root mean squared error, and mean squared error; units are speed, speed, and speed squared.
PCC	Pearson correlation coefficient; dimensionless.
Ref, Ref+	Frozen reference field without learned correction, and a baseline backbone augmented with that reference.
w/o output Ref, w/o Ref	ScaleComposer variant omitting output reference addition while retaining prior/reference inputs.
No band restriction, w/o band restriction	Control replacing every dedicated P_b with the shared P_U; Fourier projection is retained.
DC-only, w/o DC	Variants retaining only the DC correction or omitting it. Numerical w/o DC is a post-hoc intervention.

Controlled numerical benchmark

$\Omega = [0, 2\pi)^2$	Periodic two-dimensional numerical domain; benchmark quantities are nondimensional.
$\mathbf{u}$, p, t	Velocity field, pressure with zero spatial mean, and nondimensional simulation time.
S, $\mathbf{b}_i(\mathbf{x})$	Active-source index subset and localized directed body force from source i.
ν, α, χ	Kinematic viscosity, linear damping rate $\alpha = 1$, and advection switch $\chi \in \{0, 1\}$.
G_σ^{per}, σ	Periodized unnormalized Gaussian profile and its width $\sigma = 0.45$.
$\mathbf{m} \in \mathbb{Z}^2$	Periodic-image offset; the implemented Gaussian uses $m_x, m_y \in \{-1, 0, 1\}$.
a_i	Numerical forcing amplitude: active values are drawn from $\mathcal{U}[0.8, 1.2]$; inactive values are zero.
$\mathcal{U}[0.8, 1.2]$	Uniform distribution on the specified amplitude interval.
$\mathbf{c}_i$, r, θ_i	Candidate source center, arrangement radius (1.8 or 0.65), and angular position $2\pi i/6$.
$\mathbf{d}_i$, δ_i	Unit forcing direction and prescribed angular offset in radians.
$\bar{s}_S(\mathbf{x})$	Time average of instantaneous speed $\|\mathbf{u}_S(\mathbf{x}, t)\|_2$ for active subset S.
t_0, T, Δt	Averaging start 12, averaging duration 6, and solver step 0.025, in numerical time units.
$\langle\cdot\rangle_\Omega$	Spatial average over the periodic numerical domain.

Abbreviations and contextual distinctions

POD, DeepONet, MLP	Proper orthogonal decomposition, deep operator network, and multilayer perceptron.
FFT, PDE, CFD	Fast Fourier transform, partial differential equation, and computational fluid dynamics.
RANS	Reynolds-averaged Navier–Stokes.
p, t	Context distinguishes pressure / decoder rank / measurement location, and time / sample index / training-plane index.
ν, σ	Context distinguishes viscosity from spatial frequency, and Gaussian width from the sigmoid function $\sigma(\cdot)$.
δ_i, $\boldsymbol{\delta}_i$, δ_b	Angular source offset, normalized query displacement, and band-projected correction, respectively.

AI USE STATEMENT

In this work, we used generative AI tools for cleaning and reformatting the dataset, supporting qualitative data analysis, and interpreting results. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, or proposing or refine hypotheses. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code, sourcing/searching for information, and editing a research paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 2 authors. LLM-drafted parts were reviewed by at least 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The model formulation is presented in Section 3.1 with feature definitions, network dimensions, and Fourier-band specifications in Appendix A.2. Section 4 describes the experimental setting and evaluation metrics. Measurement acquisition and numerical data generation, including solver verification, are documented in Appendices A.1.1 and A.1.2, respectively. Appendix A.3 details the data splits, frozen-prior construction, optimization settings, checkpoint selection, baseline implementations, random seeds, and computing hardware. An anonymous project overview is available at `https://anonymous.4open.science/r/ScaleComposer-932B`.

REFERENCES

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4502–4510, 2016a. doi: https://doi.org/10.48550/arXiv.1612.00222.

Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. *CoRR*, abs/1612.00222, 2016b. URL http://arxiv.org/abs/1612.00222.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks, 2018. URL https://arxiv.org/abs/1806.01261.

Roman Buckle, Fangyu Liu, Yiting Zhang, Wei Liang, Shanrui Shi, Yue Lei, Yue He, Eikichi Ono, Shohei Miyata, and Adrian Chong. Mapping airflow interactions in a ceiling fan array, 2026.

Michael B. Chang, Tomer Ullman, Antonio Torralba, and Joshua B. Tenenbaum. A compositional object-based approach to learning physical dynamics. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1612.00341.

Karthik Duraisamy, Gianluca Iaccarino, and Heng Xiao. Turbulence modeling in the age of data. *Annual Review of Fluid Mechanics*, 51:357–377, 2019. doi: 10.1146/annurev-fluid-010518-040547. URL https://doi.org/10.1146/annurev-fluid-010518-040547.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3558–3565, 2019. doi: 10.1609/aaai.v33i01.33013558. URL https://ojs.aaai.org/index.php/AAAI/article/view/4235.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers, 2021.

Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. Multiwavelet-based operator learning for differential equations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24048–24062. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c9e5c2b59d98488fe1070e744041ea0e-Paper.pdf.

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 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12556–12569. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/hao23c.html.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi: 10.1214/aoms/1177703732.

Dmitrii Kochkov, Jamie A. Smith, Ayya Alieva, Qing Wang, Michael P. Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21):e2101784118, 2021. doi: 10.1073/pnas.2101784118. URL https://doi.org/10.1073/pnas.2101784118.

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. URL http://jmlr.org/papers/v24/21-1524.html.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3744–3753. PMLR, 2019.

Zongyi Li, Nikola 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. URL https://arxiv.org/abs/2010.08895.

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

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. doi: 10.1038/s42256-021-00302-5. URL https://www.nature.com/articles/s42256-021-00302-5.

Lu Lu, Raphael Pestourie, Steven G. Johnson, and Giuseppe Romano. Multifidelity deep neural operators for efficient learning of partial differential equations with application to fast inverse design of nanoscale heat transport, 2022. URL https://arxiv.org/abs/2204.06684.

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*, 2025. URL https://arxiv.org/abs/2502.02414.

Maohui Luo, Hui Zhang, Paul Raftery, Linxuan Zhou, Thomas Parkinson, Edward Arens, Yingdong He, and Elaina Present. Detailed measured air speed distribution in four commercial buildings with ceiling fans. *Building and Environment*, 200:107979, 2021. doi: 10.1016/j.buildenv.2021.107979. URL https://doi.org/10.1016/j.buildenv.2021.107979.

Joscha Mecke, Yongxiang Gao, Gerhard Gompper, and Marisol Ripoll. Chiral active systems near a substrate: Emergent damping length controlled by fluid friction. *Communications Physics*, 7:332, 2024. doi: 10.1038/s42005-024-01817-0. URL https://www.nature.com/articles/s42005-024-01817-0.

F R Menter. Two-equation eddy-viscosity turbulence models for engineering applications. *AIAA J.*, 32(8):1598–1605, August 1994. doi: 10.2514/3.12149.

Debasmita Mondal, Ameya G. Prabhune, Sriram Ramaswamy, and Prerna Sharma. Strong confinement of active microalgae leads to inversion of vortex flow and enhanced mixing. *eLife*, 10:e67663, 2021. doi: 10.7554/eLife.67663. URL https://doi.org/10.7554/eLife.67663.

Paul Raftery, Jay Fizer, Wenhua Chen, Yingdong He, Hui Zhang, Edward Arens, Stefano Schiavon, and Gwelen Paliaga. Ceiling fans: Predicting indoor air speeds based on full scale laboratory measurements. *Building and Environment*, 155:210–223, 2019. doi: 10.1016/j.buildenv.2019.03.040. URL https://doi.org/10.1016/j.buildenv.2019.03.040.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

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, pp. 8459–8468. PMLR, 2020. URL https://proceedings.mlr.press/v119/sanchez-gonzalez20a.html.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605. URL https://ieeexplore.ieee.org/document/4700287.

Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients: Appropriate use and interpretation: Appropriate use and interpretation. *Anesth. Analg.*, 126(5):1763–1768, May 2018. doi: 10.1213/ANE.0000000000002864.

Federico Tartarini, Adrian Chong, Toby Cheung, Ollie Jay, Michael G. Kent, Jiayu Li, Aleksandra Lipczynska, Ardeshir Mahdavi, Wilmer Pasut, Zbigniew Popiolek, Paul Raftery, Nicholas Ravanelli, Christian Taber, Kwok Wai Tham, Arnold J. Wilkins, and Stefano Schiavon. Ten questions concerning air movement as a primary cooling strategy: "fan-first" cooling. *Building and Environment*, 305:115218, 2026. ISSN 0360-1323. doi: https://doi.org/10.1016/j.buildenv.2026.115218. URL https://www.sciencedirect.com/science/article/pii/S036013232601019X.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Sifan Wang, Jacob H. Seidman, Shyam Sankaran, Hanwen Wang, George J. Pappas, and Paris Perdikaris. CViT: Continuous vision transformer for operator learning. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2405.13998.

Shizheng Wen, Arsh Kumbhat, Levi Lingsch, Sepehr Mousavi, Yizhou Zhao, Praveen Chandrashekar, and Siddhartha Mishra. Geometry aware operator transformer as an efficient and accurate neural surrogate for PDEs on arbitrary domains. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2505.18781.

Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S. Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=UH-cmocLJC.

Gilad Yehudai, Ethan Fetaya, Eli Meirom, Gal Chechik, and Haggai Maron. From local structures to size generalization in graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 11975–11986. PMLR, 2021.

Xihang Yue, Yi Yang, and Linchao Zhu. DeltaPhi: Physical states residual learning for neural operators in data-limited PDE solving. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3391–3401, 2017a. doi: https://doi.org/10.48550/arXiv.1703.06114.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017b.

Yiting Zhang, Wei Liang, Yue Lei, Roman Buckle, Kian Wee Chen, Ippei Izuhara, Shohei Miyata, Eikichi Ono, and Adrian Chong. Configuration-aware representation for generalizable indoor airflow field surrogates under ceiling fan reconfiguration. *Building and Environment*, 305:115167, 2026. doi: 10.1016/j.buildenv.2026.115167. URL https://doi.org/10.1016/j.buildenv.2026.115167.

Chenhong Zhou, Jie Chen, and Zaifeng Yang. SAOT: An enhanced locality-aware spectral transformer for solving PDEs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28928–28936, 2026. doi: 10.1609/aaai.v40i34.40128.