SPARSEWAKE: A BENCHMARK FOR MULTI-SOURCE STATE ESTIMATION FROM LOCAL FLOW MEASUREMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Local flow sensing can help underwater robots locate a target of interest, but the measured signal is shaped by nearby sources and may not be dominated by the target. A controlled testbed with source-level records would allow us to diagnose where localization fails and to isolate the effect of source competition, guiding the design of local perception for future underwater agents. Here, we introduce SparseWake, a benchmark for state estimation from sparse, body-oriented velocity histories generated by a physics-informed dynamic model of fish wakes. It defines three task families of increasing competition. A single-neighbor task (canonical) establishes a reference, controlled mixtures add interference around a fixed nearest target, and symmetric source sampling (common-prior) selects the nearest source only after the scene is generated. Alongside the mixed observations, SparseWake retains each source's flow contribution, so that predictions can be diagnosed against source-level ground truth and individual contributions can be edited without changing the target. To demonstrate its use, we evaluate four architectures and find a shared failure mode, in which predictions far from every source dominate squared localization error. We further characterize the roles of temporal and measurement structure, and use paired source interventions to show that fitted estimators respond to relative source contribution beyond overall signal scale. We package the observations, source records, splits, baselines, and evaluation tools as a controlled testbed for developing local perception in underwater agents.

1 INTRODUCTION

Swimming fish generate flow disturbances that carry information about their motion and position. Fish sense their hydrodynamic surroundings through the lateral line (Dijkgraaf, 1963; Liao, 2006), which has inspired artificial flow sensors for underwater robots, offering a way to infer nearby sources from measurements on the robot's own body (Yang et al., 2006; Zhai et al., 2021). For an estimator, recovering a neighbor's position from such measurements is an inverse problem: the available observations are local, while the source lies elsewhere in the surrounding flow. With multiple fish nearby, the task is to locate the fish closest to the observer using flow measurements that contain contributions from all surrounding fish. However, the closest fish does not always contribute the strongest signal, so locating it requires more than identifying the strongest contributor. When an estimate is inaccurate, its distance from the target tells us how large the error is, but not where the predicted position lies relative to the other fish. The prediction could be close to a different fish, or far from every fish in the scene. We therefore examine predictions relative to all sources to distinguish errors that place the estimate near another fish from those that fail to localize any fish source.

We introduce SparseWake to study this problem through controlled hydrodynamic sensing tasks. The generator combines body potential flow with the physics-informed dynamic wake model from Zhou et al. (2025), parameterized using high-fidelity three-dimensional direct numerical simulations. Six body-fixed probes record local velocity histories. Within each history, wake phase advances while relative positions and headings remain fixed; the dataset varies those geometries across scenes. Alongside the observations, SparseWake keeps the source states and the information needed to reconstruct each source's contribution. These records serve two purposes: they locate a prediction

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.

relative to every source, and they allow individual flow contributions to be changed while source positions and target labels remain unchanged.

Three task families separate the effects of temporal information from those of source competition. The single-neighbor task establishes how history length and sensor layout affect estimation, including sensitivity to measurement corruption. Controlled mixtures then add farther sources around a designated nearest target, allowing us to measure the effect of added interference on the same localization task. Finally, the common-prior task samples all sources symmetrically and selects the nearest only after the scene has been generated. The target therefore has no privileged source-specific sampling distribution. Successful localization requires recovering the nearest source from the mixture.

The common-prior experiments reveal a consistent spatial pattern of failure. Across four architectures and both test populations, predictions more than one-quarter body length from every source account for 90.4–95.1% of squared position error. Most of the error therefore comes from estimates that do not accurately localize any source, rather than estimates that accurately localize the wrong neighbor. MLP, TCN, GRU, and Transformer models also reach closely grouped five-fish position errors on the second test population under the shared training presets. Their similar aggregate performance is accompanied by the same dominant failure pattern.

Paired interventions show how the fitted estimators respond to changes in source competition. We amplify one source's contribution and rescale the resulting mixture to preserve its original total root-mean-square amplitude. On these edited inputs, doubling the target contribution lowers localization error, whereas doubling the strongest original rival's contribution raises it. The direction is consistent across every evaluated MLP/TCN model, source count, and test population. Separate global-amplitude controls distinguish this response from a change in overall signal scale. Relative source strength thus affects localization error even when total measurement amplitude is held constant.

2 Related work

Hydrodynamic perception and source localization. Biological studies establish the relevance of local flow sensing to fish behavior. Lateral-line function affects responses to surrounding flow and tailbeat synchronization during schooling (Liao, 2006; Mekdara et al., 2021), while canal responses encode information about source location (Ćurčić-Blake & van Netten, 2006). Artificial lateral-line arrays turn this sensing principle into a source-estimation problem. Yang et al. (2006) demonstrated hydrodynamic imaging with an artificial array, and subsequent work developed nonlinear model-based localization (Abdulsadda & Tan, 2013) and neural estimators, including multi-source localization (Pu et al., 2022; Wolf et al., 2020). For neighboring swimmers, body-surface pressure and shear-stress signatures contain information about relative position, while pressure signals also vary with swimming phase (Li et al., 2022). SparseWake builds on this physical basis using the dynamic wake model of Zhou et al. (2025). Its experiments examine how temporal observations support localization and how estimation fails when a geometrically selected target shares the measurement with other sources.

Physical benchmarks and inverse learning. Physical learning resources differ in what must be inferred from the available observations. PDEBench and The Well provide datasets for learning physical field dynamics (Takamoto et al., 2022; Ohana et al., 2024), while AirfRANS focuses on steady flow around airfoils (Bonnet et al., 2022). For inverse problems, DeepInverse makes the observation process explicit through forward operators for image reconstruction (Tachella et al., 2025), while PhySense studies sensor placement together with reconstruction from sparse measurements (Ma et al., 2025). SparseWake focuses on the state of a selected source within a mixture. Because every source contributes to the local observation, the evaluation must account for both the target rule and interference. Retaining the individual source contributions makes it possible to supplement accuracy comparisons with paired interventions that change the measurement without changing the target.

Temporal representations. A history of flow measurements provides repeated observations as well as their chronological order. Those are distinct sources of information. Temporal convolutions, recurrent networks, and attention offer different ways to use ordered observations (Bai et al., 2018; Cho et al., 2014; Vaswani et al., 2017). Symmetric pooling, following the permutation-invariant

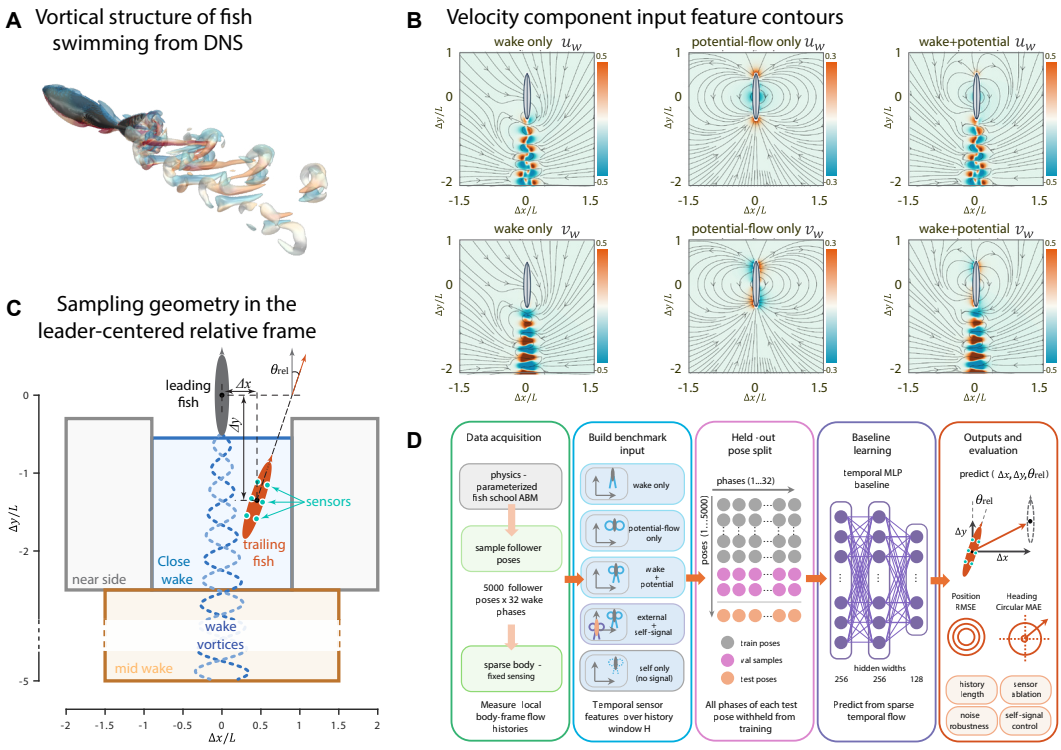

Figure 1: Physical and learning interface. (A) Vortical-wake visualization. (B) Surge and lateral velocity components of wake, body potential flow, and their vector combination. A–B are adapted from Zhou et al. (2025) to show the generator's physical basis. (C) Canonical sampling regions, relative pose, and six body-fixed probes. (D) Component-resolved acquisition, temporal features, held-out poses, and state estimation. In multi-source tasks, the same probes receive summed source velocities. Table 1 specifies each task's sampling and target rule.

construction of Zaheer et al. (2017), instead provides a reference that uses multiple frames without their order. SparseWake compares these representations and includes matched ordered-versus-shuffled training in the single-neighbor task to examine the contribution of chronology more directly. The pooled elements are observed frames, each of which can contain several source contributions. Predicting a set of latent sources would introduce a different output representation; comparisons that additionally supply decomposed inputs or source-level supervision would also need to account for that extra information.

3 Benchmark and Evaluation

3.1 Observation Interface

The observation model describes the flow that neighboring sources produce at six probes on an observer's body. The probes form left–right pairs at the anterior, midbody, and posterior locations. Each source contributes both body potential flow and a wake. We represent the former with four source–sink pairs and the latter with the physics-informed dynamic wake model of Zhou et al. (2025), parameterized using high-fidelity three-dimensional direct numerical simulations. Body length and nominal source speed are $L = 1$ and $U = 1$.

Table 1: Three task families and their independent evaluation units. Counts include the observer. All geometries have 32 phase endpoints; canonical split counts are per seed. Expanded controlled-mixture tests and the second common-prior test use separate scene populations; they are not extra seeds of one test.

Task	Scene and target definition	Train / validation / test
Canonical	One source; close-wake, near-side, and mid-wake regions. Observer pose in leader coordinates.	3,400 / 600 / 1,000 poses in grouped controls; original validation splits samples within 4,000 non-test poses.
Controlled mixtures	Target sampled first; farther sources added in nested scenes. Nearest source in observer coordinates.	4,000 / 1,000 / 1,000 base scenes; 16k extension; separate 2,000-scene tests.
Common prior	All sources proposed symmetrically; nearest selected after joint rejection. Same observer-frame target.	16,000 / 2,000 / 4,000 scenes; second test 4,000. Each split balances 3 and 5 fish.

At each probe, we add the source velocities and express the result in the observer's body frame. For an observer at position and heading $(\mathbf{x}_f, \theta_f)$ and a probe with body-frame offset $\mathbf{s}_j$, the external velocity is

$$\mathbf{z}_{j,t} = R(-\theta_f) \sum_{k=1}^{K} \left[\mathbf{u}_k^p(\mathbf{x}_f + R(\theta_f)\mathbf{s}_j; q_k) + \mathbf{u}_k^w(\mathbf{x}_f + R(\theta_f)\mathbf{s}_j; q_k, \phi_{k,t}) \right]. \tag{1}$$

Here q_k is the pose of source k and $\phi_{k,t}$ its wake phase; superscripts p, w denote the body potential and wake components, respectively. We store the two velocity components and compute magnitude *after* summation. The magnitude channel therefore includes cancellation between source contributions.

Temporal histories follow the changing wake phase at a given relative geometry. We sample the dynamic model under steady swimming conditions after an initial transient, retaining 32 phases over one tailbeat period, $T = 0.375$. A 24-frame input spans 23 intervals, or about 0.270 model time units. Relative source–observer positions and headings remain constant within a sequence, while different sources have independent initial phases. The observations are local fluid velocities; calibration to a physical sensor lies outside the observation model (Section 6).

3.2 Task distinctions

The canonical task establishes a reference for inference from one neighboring source. Its 5,000 poses cover close-wake, near-side, and mid-wake regions in proportions of 70%, 20%, and 10%, with relative headings within $\pm 30^\circ$. Component, sensor, and history sweeps predict position. Full-state experiments also predict heading.

Controlled mixtures add interference while retaining the same target. We first sample a designated nearest source at radius 0.8–$2.5L$, then add farther sources with broader angular and radial support. Training includes 2, 3, and 5 fish; 10 fish are reserved for extrapolation. This construction allows us to examine how additional sources affect estimation of the original target.

The common-prior task removes the distinction between how target and non-target sources are proposed. All sources share radius $[0.8, 5.5]L$, bearing $\pm 70^\circ$, and heading $\pm 30^\circ$ proposals. We reject scenes with overlap or without a source within $2.5L$, then select the nearest source as the target. Source indices enter symmetrically, and no minimum distance gap is imposed. Every retained scene has a deterministic label. Near a tie, a small geometric perturbation can change which source is nearest, but the label for the stored scene remains unambiguous. Because the task families differ in their scene and target priors, differences in their errors reflect more than fish count alone.

3.3 Inputs, estimators, and statistics

The input representation retains both absolute signal scale and variation relative to that scale. We concatenate the raw channels with copies divided by the six-probe mean speed, then append the mean itself. Training histories alone determine normalization. Neural models receive no source identities, states, phases, or fish-count labels. Common-prior targets encode position and sine–cosine heading;

canonical targets retain position and angle. Regression and retrieval use the same flattened history. A separate constant reference predicts the count-conditional training-target mean and is the only estimator given fish count.

Architecture comparisons use 24-frame inputs and three seeds, with the same observation access and optimizer presets for every model. We train with AdamW at learning rate 10^{-3} and weight decay 10^{-4}, using batches of 512. Training has a 180-epoch ceiling and early-stopping patience of 25. The Transformer includes temporal position encoding. Validation selects the checkpoints, which are frozen before test predictions are examined. The protocol includes neither architecture-specific tuning nor equal-FLOP matching.

For the main canonical architecture comparison, training and validation contain disjoint poses. The original protocol also held test poses apart, but allowed different samples from the same pose to appear in training and validation. A matched control measures sensitivity to that change in validation grouping. Supplements S1–S3 give the full generation, split, and model protocols.

Position RMSE pools squared Euclidean errors, making it sensitive to large localization failures. We evaluate heading with wrapped-angle MAE. Confidence intervals resample complete poses or scenes, keeping all phases and available seed predictions together. The models remain fixed during resampling, so these intervals describe variation across evaluation geometries. Seed dispersion is reported separately. In the canonical task, seeds change both model initialization and the held-out pose split.

The failure diagnostics are exploratory follow-ups: aggregate results from both common-prior test populations were available before those analyses were designed.

Usable benchmark artifacts. The benchmark data and accompanying code are publicly available through an anonymous repository.[1] The core package provides scene-major HDF5 data for all three task families with explicit train/validation/test assignments. A trainable baseline and a standalone evaluator support training new estimators and evaluating the supplied reference checkpoints, respectively. The accompanying schema distinguishes model inputs from privileged diagnostic fields, and a data card documents the dataset. Supplement S7 details package coverage, the released version, download instructions, and checks for both training and reference-checkpoint evaluation.

4 What local histories support

4.1 Reference accuracy depends on the target and evaluation protocol

Table 2: Reference performance with 24-frame histories. Position is pooled RMSE/L; heading is MAE in degrees. Canonical uses pose-disjoint validation. Every common-prior row uses the same second test population (2,000 scenes per count). Neural results pool three seeds; constant, ridge, and retrieval are deterministic. The constant alone receives fish count. Dashes denote unevaluated conditions; intervals are in Supplement S2.

	Canonical		Common: 3 fish		Common: 5 fish	
Model	Pos.	Head.	Pos.	Head.	Pos.	Head.
Count-informed constant	–	–	1.149	14.96	1.083	14.69
Ridge	–	–	0.805	12.64	0.956	13.40
History retrieval	–	–	0.649	9.21	0.947	12.32
MLP	0.063	0.60	0.472	7.50	0.802	11.18
TCN	0.071	0.47	0.492	7.81	0.793	11.37
GRU	0.091	0.55	0.502	8.32	0.807	11.63
Transformer + positions	0.074	0.57	0.500	8.01	0.814	11.59

The architecture comparison gives different answers for single-neighbor and multi-source inference (Table 2). In the canonical task, the preferred model depends on the metric: under the fixed presets, MLP has the lowest position error and TCN the lowest heading error. The paired MLP–TCN position interval nevertheless includes zero.

[1] https://huggingface.co/datasets/SparseWake/sparsewake

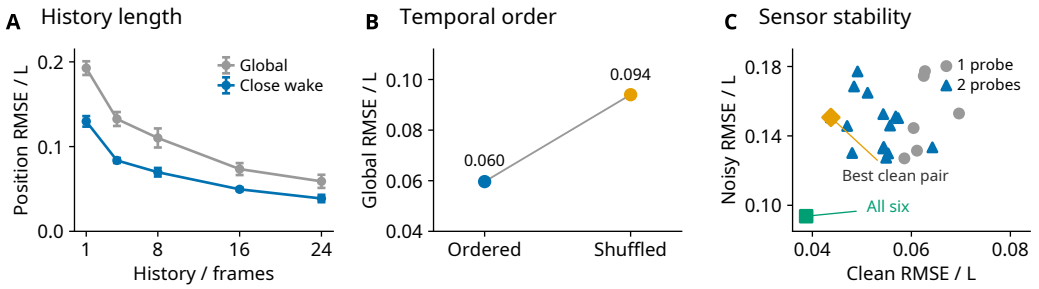

Figure 2: History and sensor layout control different aspects of canonical inference. (A) Position-only history sweep; error bars show three-seed standard deviations. (B) Matched full-state MLP retraining with ordered or shuffled histories. The interval reported in the text describes the paired RMSE penalty, not either marginal score. (C) Clean versus noisy close-wake RMSE for six single probes, fifteen pairs, and the full array. Noise uses the sensor study's shared reference at $\alpha = 0.01$. The highlighted pair is selected by clean error. A and B use different target protocols.

On the second common-prior test, the four architectures reach closely grouped five-fish position errors, with RMSE spanning 0.793–$0.814L$. All models are evaluated on the same population, and the learned nonlinear estimators outperform the regression and retrieval references. Paired comparisons favor MLP over each sequence alternative for three fish; for five fish, the MLP–TCN and MLP–GRU intervals include zero. The five-fish models also share a spatial failure pattern, examined in Section 5. These comparisons establish performance under the shared presets; a general ranking of model families would require broader evaluation.

Canonical accuracy also depends on how validation poses are grouped. In matched MLP runs, replacing sample validation with pose-disjoint validation raises pooled position RMSE from 0.0575 to $0.0628L$, approximately 9%. The paired increase is $0.0053L$, with a 95% interval of $[0.0022, 0.0089]L$. Within each seed, both runs use the same held-out test poses and endpoint count. Grouping changes both training-pose diversity and validation selection, so the measured increase includes both effects.

4.2 Repeated observations and chronological order give different benefits

Longer histories improve canonical position-only estimation. Close-wake RMSE falls from $0.130L$ with one frame to $0.039L$ with 24 (Figure 2A). A longer history, however, supplies both more observations and their temporal order. The history-length sweep alone cannot separate those contributions.

To examine order directly, we retrain the same full-state MLP with its preceding frames shuffled. Whole-pose RMSE is $0.094L$, compared with $0.060L$ for ordered training (Figure 2B). The paired increase is $0.034L$, with interval $[0.030, 0.039]L$. Retraining both models makes chronology a more direct comparison than shuffling only at test time, which also changes the input distribution relative to training.

In controlled mixtures, much of the temporal benefit comes from repeated observations (Table S5). Unordered histories improve on the current-frame baseline. Ordered models add a smaller benefit that depends on the condition: TCN improves over unordered pooling for three and five fish, with positive paired scene intervals, but performs worse for two fish. Here the architecture changes along with access to order. The matched canonical retraining holds the architecture constant and provides the more direct test of chronology.

4.3 Clean accuracy does not determine measurement stability

Noise tests distinguish accurate clean estimates from estimates that remain stable under corrupted measurements. We first compare the body potential and wake components using noise scaled to each component, accounting for the smaller absolute magnitude of potential-flow signals. At $\alpha = 0.01$, close-wake errors are $0.301L$ for potential flow, $0.084L$ for wake, and $0.094L$ for their combination

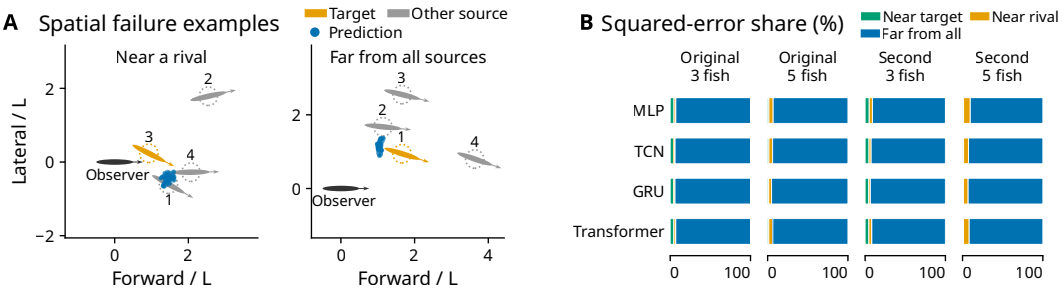

Figure 3: Far-from-all predictions dominate squared localization error. (A) Original-test five-fish MLP examples, selected by median scene error among scenes with at least half their endpoints in the stated class. Points show all phases and seeds; circles have radius $0.25L$. The two scenes use the same spatial scale. Ellipses mark bodies and arrows their headings. (B) Squared-error shares for all four architectures, both fish counts, and both test populations. Classes use the closest source and the same $0.25L$ radius; they partition total squared target error. Table S13 reports exact error shares and wrong-cell frequencies.

(Figure S2C). A separate experiment applies a shared absolute noise level to test sensitivity to a common noise floor (Supplement S4).

Sensor layout produces a similar distinction. The pair selected for lowest clean error achieves RMSE of $0.044L$, close to the full array's $0.039L$. Under the sensor-study 1% perturbation, however, those errors rise to $0.151L$ and $0.094L$, respectively. The full array is more stable under this perturbation despite the pair's similar clean accuracy. Figure 2C shows all 22 evaluated layouts. Choosing a layout therefore requires its clean error and noise response as well as its sensor count.

Robustness training improves noisy-input accuracy on the common-prior task. The training procedure combines perturbed inputs with noise-aware validation selection. For three-fish TCN on the second test, 1%-noise RMSE decreases from 0.753 to $0.546L$, with paired improvement interval $[0.190, 0.223]L$. All clean-condition improvement intervals include zero. The measured gain belongs to the combined training and selection procedure. Testing at 5% adds a further stress beyond the augmentation range.

5 What multi-source failures mean

5.1 Large errors come from predictions far from every source

An estimate can be closer to a rival than to the target without being close to either. Assigning that estimate to the rival's Voronoi region tells us which source is nearest to the prediction, but not whether the rival has been accurately localized. We therefore examine both source assignment and distance to the assigned source.

We assign each predicted point to its closest source, then apply a proximity threshold of $r = 0.25L$. The resulting classes distinguish predictions near the target, predictions near an assigned rival, and predictions farther than r from every source. Every prediction belongs to exactly one class. Target identity remains fixed throughout the phase sequence; the classification describes the output's location, not an internal source-selection decision. We also check radii $0.1L$, $0.5L$, and a separation-relative scale.

Predictions far from every source account for most of the squared localization error. For the original MLP, their shares are 93.0% with three fish and 93.1% with five fish (Figure 3B). Yet this class contains only 36.9% and 58.7% of endpoints, respectively. Large errors give these predictions a disproportionate share of the loss. Reporting both class frequency and error share makes that distinction explicit.

Accurately localizing a rival contributes much less to the total error. Predictions within $0.25L$ of an assigned rival account for 2.2% and 5.4% of squared target error in the three- and five-fish cases. Wrong-cell assignments occur at rates of 4.0% and 14.6%, respectively. Figure 3A illustrates the

distinction with median-error scenes selected within each class: one estimate lies near a rival, while the other lies away from every source. The examples illustrate the geometry; the classification of all predictions determines the error shares.

The dominant error pattern persists across model families and test populations. At $r = 0.25L$, far-from-all predictions contribute 90.4–95.1% of squared target error across all four architectures, both fish counts, and both tests (Figure 3B; Table S13). For MLP/TCN on the original and second tests, near-rival contributions remain about 1–8% at $0.25L$ and 5–19% at $0.5L$. The shared result is a geometric pattern of failure. Its mechanism is unresolved: conditional averaging could produce such estimates, as could inaccurate regression or weak observability.

5.2 Geometry and signal strength describe different aspects of competition

Two sources at similar distances from the observer can still be far apart from each other. Radial gap alone therefore misses part of the geometry relevant to localization error. We jointly bin scenes by the nearest–next-nearest distance gap and their bearing separation. Some bins with small gaps and large angular separations have higher absolute RMSE. Their ordering changes after errors are normalized by the distance between the candidate sources (Supplement S5). Part of the absolute-error trend thus reflects how far apart the candidates lie; higher RMSE in those bins alone does not demonstrate stronger ambiguity or poorer identifiability.

Selecting the nearest source does not ensure that its flow dominates the observation. To measure its relative contribution, we reconstruct each source's velocity contribution from its stored pose and phase using the unchanged forward kernels. Their sum reproduces the saved velocities with a maximum component discrepancy of 4.77×10^{-7}. We define the target-to-interference ratio as

$$\rho = \frac{\mathrm{RMS}(\mathbf{u}_\star)}{\mathrm{RMS}(\sum_{k \neq \star} \mathbf{u}_k)}, \tag{2}$$

where $\star$ denotes the nearest source. RMS is calculated over all 32 phases, six probes, and both velocity components. The denominator uses the sum of the non-target velocities, so cancellation between those contributions is maintained. Both the reconstructed contributions and the ratio are derived from evaluation metadata and are never supplied to the estimators.

The nearest source often contributes less flow than its competitors. Across the two five-fish test populations, it is the strongest individual contributor in about 52–53% of scenes. Its RMS exceeds that of the vector-summed interference in about 39%. Lower target-to-interference ratios coincide with larger errors in both populations (Supplement S5). This association motivates a paired test: change a source's contribution while holding every source position and the target label fixed.

5.3 Paired source interventions separate competition from overall amplitude

Amplifying a source can change both its relative contribution and the overall amplitude of the measurement. Our interventions separate these effects. We multiply one reconstructed source contribution by a gain while preserving scene geometry, phase, and the nearest-target label. For source contribution $\mathbf{u}_j$ and original mixture $\mathbf{z}$, the edited observation is

$$\mathbf{z}^{(\alpha)} = \mathbf{z} + (\alpha - 1)\mathbf{u}_j, \qquad \alpha \in \{0.5, 0.75, 1, 1.5, 2\}. \tag{3}$$

Each scene receives a target edit and, separately, an edit to its strongest original non-target source. We then construct two controls. The RMS-matched input rescales the edited mixture to the original scene RMS, preserving overall amplitude while changing the relative source contributions. The global-amplitude control instead scales the entire original mixture to the edited RMS. It changes overall amplitude while preserving the source ratios.

After each edit, we recompute velocity magnitude and apply the frozen training normalization. Evaluation uses MLP/TCN with all three seeds, both fish counts, and both existing test populations. The protocol was fixed before the intervention summaries were examined. These edits change observation components while keeping the physical scene labels unchanged, extending beyond the training generator's amplitude rule.

Amplifying the target improves localization even when total RMS is unchanged. On the second five-fish test, doubling its contribution reduces MLP RMSE from 0.802 to $0.596L$ and TCN RMSE

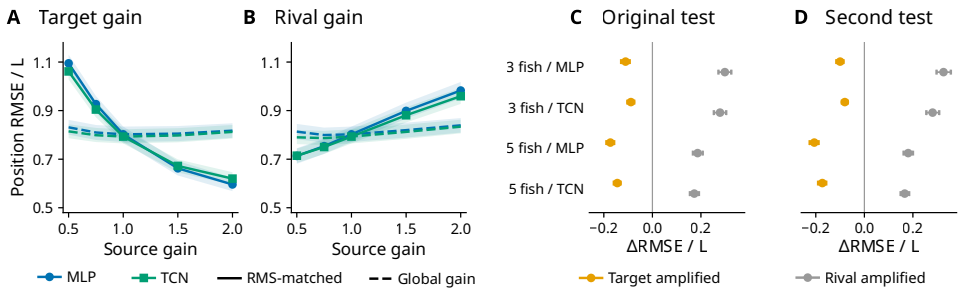

Figure 4: Relative source contribution changes error at fixed total RMS. (A–B) Second-test five-fish target and rival gain curves. Solid curves preserve the original total RMS; dashed controls change only global amplitude to match the unnormalized edit's RMS. Bands are scene-bootstrap 95% intervals. (C–D) Paired RMSE changes at gain 2 under RMS matching, for both models and both fish counts in each test population. Negative values mean improvement; bars are paired scene-bootstrap 95% intervals. Rival edits amplify the strongest original non-target source. Geometry, labels, and checkpoints stay fixed; the edits extend beyond the training amplitude rule.

from 0.793 to $0.620L$ (Figure 4A). The paired improvement intervals are $[0.189, 0.225]L$ for MLP and $[0.157, 0.190]L$ for TCN. The corresponding global-amplitude controls yield errors of 0.817 and $0.812L$, respectively. Changing overall velocity scale alone does not reproduce the improvement.

Amplifying the rival produces the opposite response. At unchanged total RMS, doubling the strongest rival raises five-fish RMSE to $0.983L$ for MLP and $0.959L$ for TCN (Figure 4B). The predictions also shift toward that rival. The mean rivalward displacement magnitude along the target–rival axis is $0.193L$ for MLP and $0.191L$ for TCN, with paired intervals of $[0.179, 0.207]L$ and $[0.179, 0.204]L$, respectively (Figure S6F). Supplement S5.5 gives the full gain curves, raw-amplitude interventions, and endpoint summaries.

The direction of the error change is consistent across every evaluated combination of MLP/TCN model, fish count, and test population (Figure 4C–D). At $\alpha = 2$ with total RMS unchanged, target amplification lowers RMSE and amplification of the strongest original rival raises it. The paired comparisons and global-amplitude controls establish that relative source contribution affects the fitted estimators beyond the effect of overall signal scale.

6 Discussion & Conclusion

SparseWake provides a controlled setting for evaluating neighbor localization before integrating estimates into navigation or coordination policies. It separates two questions in local flow sensing: what sparse observations reveal about a neighbor, and how inference changes when other sources contribute to the same measurement. In the single-neighbor controls, chronological order improves estimation beyond repeated observations, while similar clean accuracy can conceal different noise sensitivity. In mixed flows, the nearest source need not dominate the signal, and changing relative source strength changes the fitted estimators' errors even at fixed total amplitude.

Four architectures share the spatial failure pattern and reach closely grouped five-fish errors on the second test. Their fixed presets cannot separate capacity from optimization, input representation, or available information; greater capacity could still offer a partial remedy. The interventions likewise do not identify the mechanism behind far-from-all predictions. Conditional averaging, inaccurate or spatially biased regression, and weak observability remain possible explanations. Neither analysis establishes a physical identifiability limit.

The physical setting holds relative source–observer geometry constant within each sequence and superposes individually generated fields. It omits mutual wake evolution, propulsion responses to neighbors, and jointly enforced exact multibody boundary conditions. Without physical sensor calibration, numerical noise multipliers cannot be assigned instrument-noise levels, and transfer error to real swimming is not quantified. Sensor-layout results concern the tested measurement protocol rather than a universal minimum sensor count. Moving configurations and other sensing modalities offer possible extensions.

REPRODUCIBILITY STATEMENT

The SparseWake datasets and accompanying code are publicly available without registration or access requests on Hugging Face, SparseWake, release `iclr2027-v1.0rc2`. The release includes the processed benchmark datasets, small sample datasets for integration checks, reference checkpoints, training and evaluation code, and data-generation code operating on a frozen wake library. The accompanying documentation specifies data schemas, split protocols, dependencies, and usage instructions. Archive-level SHA-256 checksums and per-file manifests are provided for integrity verification. The linked repository revision fixes the released artifact version.

AI USE STATEMENT

In preparing this submission, we used a large language model only for language polishing to correct grammar and improve the clarity and readability of the text. It was not used for research conceptualization, experimental design, methodology, data analysis, or the drafting of initial manuscript. All edits were reviewed and verified by the authors, who confirmed that no meaning, claim, or result was altered or introduced.

REFERENCES

Ahmad T. Abdulsadda and Xiaobo Tan. Nonlinear estimation-based dipole source localization for artificial lateral line systems. *Bioinspiration & Biomimetics*, 8(2):026005, 2013. doi: 10.1088/1748-3182/8/2/026005.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. URL `https://arxiv.org/abs/1803.01271`.

Florent Bonnet, Jocelyn Mazari, Paola Cinnella, and Patrick Gallinari. AirfRANS: High fidelity computational fluid dynamics dataset for approximating reynolds-averaged navier–stokes solutions. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23463–23478, 2022. doi: 10.52202/068431-1705. URL `https://proceedings.neurips.cc/paper_files/paper/2022/hash/94ab7b23a345f93333eac8748a66c763-Abstract-Datasets_and_Benchmarks.html`.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1179. URL `https://aclanthology.org/D14-1179/`.

Branislava Ćurčić-Blake and Sietse M. van Netten. Source location encoding in the fish lateral line canal. *Journal of Experimental Biology*, 209(8):1548–1559, 2006. doi: 10.1242/jeb.02140.

S. Dijkgraaf. The functioning and significance of the lateral-line organs. *Biological Reviews*, 38(1): 51–105, 1963. doi: 10.1111/j.1469-185X.1963.tb00654.x.

Gen Li, Dmitry Kolomenskiy, Hao Liu, Benjamin Thiria, and Ramiro Godoy-Diana. Hydrodynamical fingerprint of a neighbour in a fish lateral line. *Frontiers in Robotics and AI*, 9:825889, 2022. doi: 10.3389/frobt.2022.825889.

James C. Liao. The role of the lateral line and vision on body kinematics and hydrodynamic preference of rainbow trout in turbulent flow. *Journal of Experimental Biology*, 209(20):4077–4090, 2006. doi: 10.1242/jeb.02487.

Yuezhou Ma, Haixu Wu, Hang Zhou, Huikun Weng, Jianmin Wang, and Mingsheng Long. PhySense: Sensor placement optimization for accurate physics sensing. In *Advances in Neural Information Processing Systems*, volume 38, pp. 35697–35725, 2025. doi: 10.52202/085713-1200. URL `https://proceedings.neurips.cc/paper_files/paper/2025/hash/332b4fbe322e11a71fa39d91c664d8fa-Abstract-Conference.html`.

Prasong J. Mekdara, Fazila Nasimi, Margot A. B. Schwalbe, and Eric D. Tytell. Tail beat synchronization during schooling requires a functional posterior lateral line system in giant dan-

ios, Devario aequipinnatus. *Integrative and Comparative Biology*, 61(2):427–441, 2021. doi: 10.1093/icb/icab071.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam H. Parker, Miles Cranmer, and Shirley Ho. The well: a large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 44989–45037, 2024. doi: 10.52202/079017-1430.

Yanyun Pu, Zheyi Hang, Gaoang Wang, and Huan Hu. Bionic artificial lateral line underwater localization based on the neural network method. *Applied Sciences*, 12(14):7241, 2022. doi: 10.3390/app12147241.

Julián Tachella, Matthieu Terris, Samuel Hurault, Andrew Wang, Leo Davy, Jérémy Scanvic, Victor Sechaud, Romain Vo, Thomas Moreau, Thomas Davies, Dongdong Chen, Nils Laurent, Brayan Monroy, Jonathan Dong, Zhiyuan Hu, Minh-Hai Nguyen, Florian Sarron, Pierre Weiss, Paul Escande, Mathurin Massias, Thibaut Modrzyk, Brett Levac, Tobías I. Liaudat, Maxime Song, Johannes Hertrich, Sebastian Neumayer, and Georg Schramm. Deepinverse: A python package for solving imaging inverse problems with deep learning. *Journal of Open Source Software*, 10(115): 8923, 2025. doi: 10.21105/joss.08923. URL https://doi.org/10.21105/joss.08923.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1596–1611, 2022. doi: 10.52202/068431-0117. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0a9747136d411fb83f0cf81820d44afb-Abstract-Datasets_and_Benchmarks.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. URL https://arxiv.org/abs/1706.03762.

Ben J. Wolf, Jos van de Wolfshaar, and Sietse M. van Netten. Three-dimensional multi-source localization of underwater objects using convolutional neural networks for artificial lateral lines. *Journal of the Royal Society Interface*, 17(162):20190616, 2020. doi: 10.1098/rsif.2019.0616.

Yingchen Yang, Jack Chen, Jonathan Engel, Saunvit Pandya, Nannan Chen, Craig Tucker, Sheryl Coombs, Douglas L. Jones, and Chang Liu. Distant touch hydrodynamic imaging with an artificial lateral line. *Proceedings of the National Academy of Sciences*, 103(50):18891–18895, 2006. doi: 10.1073/pnas.0609274103.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Yufan Zhai, Xingwen Zheng, and Guangming Xie. Fish lateral line inspired flow sensors and flow-aided control: A review. *Journal of Bionic Engineering*, 18:264–291, 2021. doi: 10.1007/s42235-021-0034-y.

Ji Zhou, Jung-Hee Seo, and Rajat Mittal. Effect of hydrodynamic wakes in dynamical models of large-scale fish schools. *Physics of Fluids*, 37(1):011912, 2025. doi: 10.1063/5.0250013.

S1 PHYSICAL GENERATOR, COORDINATES, AND SCENE CONSTRUCTION

S1.1 FORWARD MODEL AND OBSERVATION CHANNELS

The flow generator combines body potential flow, represented by four source–sink pairs, with the physics-informed dynamic wake model of Zhou et al. (2025). The wake model is parameterized using high-fidelity three-dimensional direct numerical simulations of a carangiform swimmer.

The benchmark samples a stored wake library after an initial transient of approximately 30 periods, retaining 32 phases of a nominal $T = 0.375$ tailbeat. To evaluate a scene, we translate and rotate each source field using its position and heading, then query the observer probes. The resulting velocities are expressed in the observer's body-fixed coordinates.

Each scene stores a $32 \times 6 \times 3$ raw-observation array: two observer-frame velocity components and their magnitude at every probe and phase. Component and source velocities are summed before computing magnitude, so the speed channel is the magnitude of the total velocity, not the sum of source magnitudes. Common-prior data also retain wake and potential velocities separately.

Data checks reconstruct the first two raw channels from wake plus potential and the third from their Euclidean norm. The maximum stored-precision component residual in the canonical 160,000-sample dataset is approximately 1.2×10^{-7}.

Separate evaluation of source contributions allows us to retain potential and wake fields for controlled component comparisons before summing them at the focal fish's probes. Common-prior estimators receive the combined observations. The retained decomposition belongs to the simulator, with its physical scope specified in Section 6.

S1.2 FRAMES, PROBES, AND TEMPORAL WINDOWS

Target coordinates differ between the single-source and multi-source tasks. In the canonical task, the leader points along world $+y$, and the target is the observer's leader-centered position and relative heading. Multi-source targets instead locate the nearest source in the focal observer's body frame:

$$\mathbf{y} = \big[R(-\theta_f)(\mathbf{x}_\star - \mathbf{x}_f),\ \sin(\theta_\star - \theta_f),\ \cos(\theta_\star - \theta_f)\big]. \tag{4}$$

Multi-source heading is decoded with atan2. Canonical models retain their three continuous position-and-angle outputs. Both target conventions use wrapped heading error for evaluation.

Six probes form anterior, middle, and posterior left–right pairs. With the body-frame first axis pointing forward, their offsets are $(0.227464, \pm 0.066790)L$, $(0, \pm 0.075)L$, and $(-0.227464, \pm 0.066790)L$, respectively. Probe numbers in Figure S2A follow this ordering. Sensing-width experiments scale only the lateral coordinates, leaving the source-body model unchanged.

Each 24-frame window ends at the designated endpoint and reaches back 23 phase intervals, wrapping cyclically through the library. Its duration is therefore $23T/32 \simeq 0.270$ model time units. Wake phase supplies the time variation; relative positions and headings remain constant over the window.

S1.3 CANONICAL AND MULTI-SOURCE SAMPLERS

Canonical sampling allocates 70% of poses to the close wake, 20% to the near side, and 10% to the mid wake. Before overlap rejection, close-wake bounds are $\Delta x/L \in [-0.9, 0.9]$ and $\Delta y/L \in [-2.5, -0.55]$. Near-side bounds are $|\Delta x|/L \in [0.9, 2.0]$ and $\Delta y/L \in [-2.5, -0.3]$; mid-wake bounds are $\Delta x/L \in [-1.5, 1.5]$ and $\Delta y/L \in [-5, -2.5]$. Relative heading spans $\pm 30°$. The canonical overlap rule uses an ellipse with semiaxes $0.5L$ and $0.075L$.

Controlled mixtures place the target first, at radius 0.8–$2.5L$, bearing $\pm 70°$, and relative heading $\pm 30°$. Additional sources are at least $0.05L$ farther, with an additional radial range up to $3L$, full-circle bearings, and headings within $\pm 60°$. Conservative body-separation checks reject overlap. Adding sources creates nested configurations with the same target. A shared base-scene identifier links the fish-count conditions for paired evaluation. Because the target is placed first under a different prior, the construction is deliberately not exchangeable source sampling.

The common-prior sampler proposes every source from radius $[0.8, 5.5]L$, bearing $\pm 70°$, and heading $\pm 30°$. It jointly rejects overlaps and scenes with no source within $2.5L$. The rejection rule is invariant

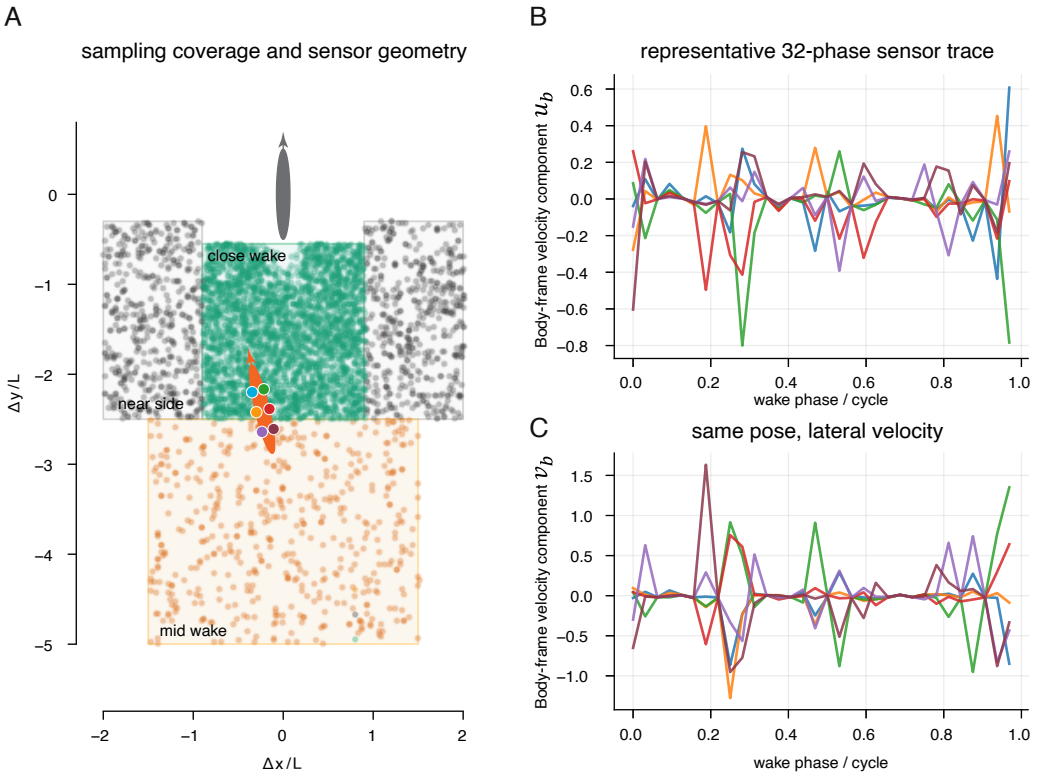

Figure S1: Canonical spatial coverage and phase-resolved observations. Sampling spans close-wake, near-side, and mid-wake regions. Traces show both velocity components at the six probes over the 32 saved phases. Relative geometry stays constant during each trace while wake phase advances.

to source ordering; nearest identity is computed afterward. Three- and five-fish scenes contain two and four sources respectively, always excluding the observer from source count. Each split balances the two fish counts. Initial source phases are independent. The smallest retained nearest-distance gap is about $8.7 \times 10^{-5} L$; no gap threshold was imposed.

The records retain source poses, active-source masks, phase offsets, nearest-source index, and nearest-to-next-nearest gap for diagnostic queries. Only diagnostic queries use these metadata fields; they are excluded from estimator inputs. Audits verify nearest labels and nonoverlap across the original 22,000 common-prior scenes. Independent direct queries check 288 probe vectors across audited temporal and common-prior conditions. The audit checks consistency with the implemented model. It contains no physical sensor validation.

S2 Evaluation populations, grouping, and uncertainty

S2.1 Splits and test-population map

Canonical splits keep complete test poses out of training and validation. The original protocol holds out 1,000 of 5,000 poses per seed, then splits phase endpoints from the remaining 4,000 poses between training and validation.

Endpoint budgets are matched in the grouping comparison: both arms contain 108,800 training, 19,200 validation, and 32,000 test endpoints. The grouped arm uses 3,400, 600, and 1,000 disjoint poses; the sample arm shares pose support between training and validation. Test poses are identical between arms within each seed. Grouping reduces training-pose coverage at the same time as it changes validation selection, so the comparison measures their combined effect.

The original controlled-mixture suite uses 4,000 training, 1,000 validation, and 1,000 test base scenes. The scale extension keeps the original training bank and adds 12,000 training scenes, retains validation, and evaluates old and new models on a separate 2,000-scene test. The subsequent temporal controls reuse the 16k training bank but use another 2,000-scene test. Scores from these distinct tests are never treated as paired measurements of one scene population.

The original common-prior bank contains 16,000 training, 2,000 validation, and 4,000 test scenes. A second, geometrically disjoint test contains 4,000 scenes and was generated with seed 2026092508 after regression, retrieval, and augmentation choices were fixed. Every model in main Table 2 is evaluated on this same second population. The GRU/Transformer follow-up also evaluates frozen checkpoints on both existing tests.

The mechanism analysis was designed after aggregate outcomes from both populations had been inspected. Agreement across these populations provides an exploratory consistency check, not confirmation on an untouched test.

S2.2 Metrics and uncertainty units

For S fitted seeds, N scenes, and $P = 32$ endpoints, pooled position error is

$$E = \sqrt{\frac{1}{SNP} \sum_{s,i,t} \|\widehat{\mathbf{x}}_{s,i,t} - \mathbf{x}_i\|_2^2}. \tag{5}$$

Scene RMSE pools seeds and endpoints within a scene before taking the square root. It is used for distributions and deterministic example selection. Heading MAE averages absolute wrapped angular error. A mean Euclidean-error spatial map is explicitly distinguished from RMSE when retained from earlier diagnostics.

Main-reference intervals and paired contrasts use 2,000 bootstrap draws of complete scene or pose clusters. When canonical test sets overlap across seeds, all available predictions for a given pose stay together. Paired contrasts resample the same identifiers for both models. Augmentation intervals retain the earlier 4,000-scene-bootstrap calculation.

Bootstrap intervals hold the fitted models fixed and use no multiple-comparison correction. Run dispersion is reported separately as the standard deviation over three seeds. Those seeds change both initialization and held-out pose assignment in the canonical task, while common-prior runs share scene splits. A seed standard deviation consequently describes variation between runs, not a confidence interval over new geometries.

On the second common-prior test, position-RMSE seed standard deviations for MLP, TCN, GRU, and Transformer are 0.0067, 0.0063, 0.0137, $0.0237L$ (three fish) and 0.0045, 0.0110, 0.0060, $0.0056L$ (five fish), respectively. These describe three fixed training runs, separately from the scene intervals.

S2.3 Original test and grouping sensitivity

The original common-prior results are reported separately from the second test in Table S3, alongside canonical validation sensitivity. The generator is the same for both populations, yet their five-fish errors differ. Architecture comparisons use a common test population so that a model contrast is not mixed with this population difference. Grouped canonical validation increases MLP position error by $0.005309L$ (95% interval $[0.002207, 0.008939]L$) and heading error by 0.0627° ($[0.0498, 0.0758]^\circ$). Under grouped validation, MLP minus TCN is $-0.00805L$ for position ($[-0.02532, 0.00473]L$) and 0.1277° for heading ($[0.1118, 0.1434]^\circ$). The position interval includes zero, while the heading comparison favors TCN.

S3 Temporal controls, reference models, and resources

S3.1 Representations and architectures

Each phase contributes 37 features: 18 raw measurements from the six probes, 18 copies divided by mean probe speed, and the mean speed itself. Across 24 frames, flattening produces the 888-feature vector supplied to the MLP. Sequence models retain the same values in a 24×37 array.

Table S1: Uncertainty for Table 2. Brackets give 95% intervals from 2,000 complete-pose/scene bootstrap draws, conditional on fitted models. They describe variation over held-out geometries; retraining dispersion is reported separately. Canonical clusters retain all available seed predictions for each pose.

Population	Model	Position / L	Heading / degrees
Second, 3 fish	Constant	1.149 [1.129, 1.169]	14.96 [14.59, 15.33]
Second, 5 fish	Constant	1.083 [1.063, 1.103]	14.69 [14.31, 15.06]
Second, 3 fish	Ridge	0.805 [0.783, 0.827]	12.64 [12.27, 13.00]
Second, 5 fish	Ridge	0.956 [0.932, 0.980]	13.40 [13.01, 13.77]
Second, 3 fish	Retrieval	0.649 [0.620, 0.677]	9.21 [8.87, 9.50]
Second, 5 fish	Retrieval	0.947 [0.917, 0.975]	12.32 [11.92, 12.72]
Second, 3 fish	MLP	0.472 [0.445, 0.497]	7.50 [7.23, 7.76]
Second, 5 fish	MLP	0.802 [0.771, 0.834]	11.18 [10.84, 11.54]
Second, 3 fish	TCN	0.492 [0.467, 0.518]	7.81 [7.55, 8.08]
Second, 5 fish	TCN	0.793 [0.765, 0.822]	11.37 [11.03, 11.72]
Second, 3 fish	GRU	0.502 [0.476, 0.526]	8.32 [8.05, 8.60]
Second, 5 fish	GRU	0.807 [0.778, 0.836]	11.63 [11.29, 11.98]
Second, 3 fish	Transformer	0.500 [0.475, 0.525]	8.01 [7.74, 8.29]
Second, 5 fish	Transformer	0.814 [0.786, 0.844]	11.59 [11.25, 11.94]
Canonical	MLP	0.063 [0.057, 0.069]	0.60 [0.58, 0.61]
Canonical	TCN	0.071 [0.056, 0.089]	0.47 [0.45, 0.49]
Canonical	GRU	0.091 [0.066, 0.117]	0.55 [0.52, 0.59]
Canonical	Transformer	0.074 [0.062, 0.086]	0.57 [0.54, 0.59]

Table S2: Paired position differences on the second common-prior test. Negative MLP-minus-model differences favor MLP. Each comparison resamples the same complete scenes for both fitted models; intervals are exploratory and uncorrected.

Fish	Compared with MLP	Difference / L	95% interval
3	TCN	-0.0204	[-0.03115, -0.00960]
3	GRU	-0.0300	[-0.04090, -0.01862]
3	Transformer + positions	-0.0287	[-0.04170, -0.01611]
5	TCN	0.0091	[-0.00264, 0.02013]
5	GRU	-0.0052	[-0.01611, 0.00592]
5	Transformer + positions	-0.0120	[-0.02414, -0.00005]

Table S3: Primary formal comparisons with 95% grouped-pose/scene intervals and seed dispersion. All values are in body lengths. Regional canonical results and individual training records accompany the artifact.

Task	Validation	Model	RMSE	95% interval	Seed SD
Canonical	sample	mlp	0.0575	[0.0516, 0.0638]	0.0061
Canonical	pose	mlp	0.0628	[0.0566, 0.0692]	0.0070
Canonical	pose	tcn	0.0708	[0.0563, 0.0888]	0.0212
Canonical	pose	gru	0.0912	[0.0662, 0.1168]	0.0190
Canonical	pose	transformer	0.0738	[0.0622, 0.0857]	0.0118
Fish 3	scene	mlp	0.4915	[0.4632, 0.5208]	0.0033
Fish 3	scene	tcn	0.5055	[0.4780, 0.5328]	0.0030
Fish 3	scene	gru	0.5131	[0.4859, 0.5397]	0.0172
Fish 3	scene	transformer	0.5105	[0.4840, 0.5384]	0.0255
Fish 5	scene	mlp	0.7437	[0.7152, 0.7735]	0.0018
Fish 5	scene	tcn	0.7434	[0.7166, 0.7726]	0.0013
Fish 5	scene	gru	0.7683	[0.7409, 0.7966]	0.0060
Fish 5	scene	transformer	0.7716	[0.7442, 0.8010]	0.0104

Normalizers are fitted only on training histories, including the repeated use of phases in cyclic windows. The current-frame baseline omits preceding observations. The unordered encoder pools the history symmetrically, retaining multiple observations without their chronology.

The MLP has hidden widths 256, 256, and 128, with layer normalization, GELU, and dropout 0.05. The TCN embeds phases into 96 channels and applies residual convolution blocks with kernel size 3 and dilations 1, 2, and 4. Its prediction head combines the last and mean temporal features. The GRU has hidden size 96 and a 128-unit prediction head.

The Transformer uses 64-dimensional embeddings, four attention heads, two encoder layers, and 128-dimensional feedforward blocks. It applies dropout 0.1, sinusoidal positions over 24 frames, and mean pooling; encoder layers are initialized separately. An earlier pilot omitted positional encoding and was permutation invariant in evaluation mode. Only the formal, position-encoded model is used here to compare order-aware attention. Engineering checks verify the permutation behavior of both implementations.

Formal runs share AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 512, and gradient clipping at 5. Training stops at a maximum of 180 epochs or validation patience 25. No hyperparameter search is performed. The shared protocol controls observation access and optimizer presets, while model capacity and elapsed compute differ.

Earlier canonical experiments use separate budgets. Original full-state training used batch size 1,024; both grouping arms are retrained under the common follow-up implementation. The position-only size study uses 100 epochs and patience 15. Its results remain separate from the 180-epoch reference.

Table S4: Resources for the 21 new runs. Selected epochs span the three seeds; minutes are mean wall time per run. Frozen common-prior MLP/TCN timings are excluded from this hardware-specific summary.

Task	Validation	Model	Parameters	Selected epochs	Minutes
canonical	sample	mlp	327,939	112–179	2.9
canonical	pose	mlp	327,939	88–179	2.5
canonical	pose	tcn	140,483	36–88	2.6
canonical	pose	gru	51,683	53–179	2.8
canonical	pose	transformer	78,083	116–155	6.3
common	scene	gru	51,812	4–6	2.2
common	scene	transformer	78,212	2–3	5.0

Resources are measured on one NVIDIA GeForce RTX 3060 Ti with PyTorch 2.11.0 and TF32 matrix multiplication enabled. Reported minutes include validation and checkpoint overhead. Tasks and models select substantially different stopping epochs. A common epoch ceiling therefore leaves compute unequal. Independent replay reproduces all 21 follow-up runs' saved predictions on the original GPU backend; CPU floating-point results can differ slightly.

S3.2 Regression, retrieval, and prior reference

Ridge regression uses flattened standardized histories, with validation selecting regularization 10^{-4} from $\{10^{-4}, 10^{-2}, 1, 100\}$. Retrieval projects histories onto the largest 64 training-covariance eigenvectors, retaining 95.14% of feature variance, and indexes all 512,000 training windows. Validation selects 15 neighbors from $\{1, 5, 15\}$ using squared loss on standardized position and sine–cosine heading. Retrieved target vectors are averaged before heading decoding.

Regression and retrieval have no access to fish count. Only the constant reference receives that metadata, allowing it to predict the count-conditional training-target mean and provide a stronger prior baseline than an unconditional mean.

S3.3 Complete temporal controls

Matched retraining tests the contribution of chronological order in the canonical full-state MLP. Whole-pose pooled RMSE is $0.05965L$ with ordered inputs and $0.09400L$ with preceding frames shuffled, a paired increase of $0.03436L$ ($[0.03000, 0.03879]L$). Test-only shuffling produces $0.76269L$. The large test-only penalty also includes an input-distribution shift. Evidence about chronological order therefore rests primarily on the matched retraining comparison.

Table S5: All controlled-mixture temporal conditions. Global position RMSE/L pools three seeds on 2,000 base scenes per condition. Synchronized sources share initial phase; all sources use the same nominal frequency. Ten fish are absent from training.

Condition	Current frame	Unordered	MLP	TCN
2 fish	0.1697	0.1219	0.1077	0.1314
3 fish	0.5026	0.3606	0.3264	0.3350
3 synchronized	0.5218	0.3668	0.3377	0.3467
5 fish	0.8125	0.6592	0.6485	0.6212
10 fish	1.0969	1.0418	1.0323	1.0012

The benefit of an ordered architecture depends on the condition. TCN performs worse than unordered pooling with two fish (Table S5). For three fish, its heading MAE is 5.548° with independent initial phase offsets and 5.561° with synchronized offsets. Initial phase is the variable in this synchronization control. Swimming frequencies remain equal throughout the library, so heterogeneous frequencies are outside the test.

Table S6: Matched canonical full-state MLP temporal controls. Global position RMSE/L pools held-out poses and seeds. Differences use paired whole-pose 95% intervals. Test-only shuffling also changes the input distribution.

Condition	RMSE	Increase over ordered [95% CI]
Ordered training	0.060	—
Shuffled retraining	0.094	0.034 [0.030, 0.039]
Test-only shuffle	0.763	0.703 [0.678, 0.727]

Table S7: Canonical position-only history and training-size sweeps, mean $\pm$ standard deviation over three seeds. The history sweep uses the main training protocol; the size sweep uses a separate 100-epoch budget and patience 15. Their 24-frame, 4,000-pose results must not be combined as one setting.

History / frames	Global RMSE/L	Close-wake RMSE/L
1	0.1926 ± 0.0081	0.1298 ± 0.0062
4	0.1326 ± 0.0082	0.0837 ± 0.0029
8	0.1102 ± 0.0112	0.0698 ± 0.0050
16	0.0735 ± 0.0070	0.0496 ± 0.0020
24	0.0590 ± 0.0077	0.0387 ± 0.0043
Training poses (separate budget)		Close-wake RMSE/L
500		0.1765 ± 0.0105
1000		0.0965 ± 0.0122
2500		0.0531 ± 0.0015
4000		0.0451 ± 0.0008

S4 Components, sensing layouts, and measurement corruption

S4.1 Noise protocols answer different questions

Perturbations enter through velocity components. We then recompute magnitude and features using the unchanged clean-training normalizers. The noise reference differs between studies, and equal numerical multipliers do not represent a shared physical instrument-noise level.

A shared absolute noise floor perturbs potential-flow signals more strongly in relative terms because their raw magnitude is smaller. The component-scaled experiment accounts for this difference by scaling perturbations to each component. A common absolute floor and a component-scaled perturbation impose different noise conditions. Neither protocol requires combined inputs to outperform every isolated component at every noise level.

Table S8: Noise-reference map. The scale is always computed from training data. These definitions govern within-protocol comparisons.

Study	Reference	Interpretation
Shared component noise	Shared absolute velocity scale	Fixed instrument corruption of signals with different raw magnitudes.
Relative component noise	Each component's training velocity-channel scale	Sensitivity to relative corruption; matched standard-normal draws across comparisons.
Probe-subset study	Common sensor-study reference	Sensor redundancy under one shared perturbation scale.
Self-signal controls	Total-input training-channel reference	Corruption of prescribed external-plus-self observations.
Common-prior robustness	Scalar RMS 0.213386 across training scenes, probes, phases, and velocity components	Paired trace noise across all compared models and seeds.

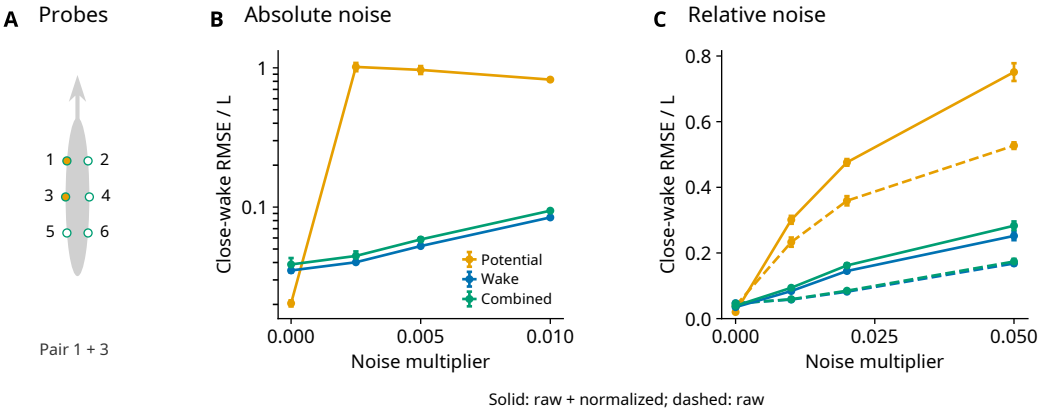

Figure S2: Probe geometry and component perturbation trends. (A) Numbered probes; the highlighted pair has the lowest clean error in Figure 2C. (B) Shared absolute noise floor, with a log error axis. (C) Component-scaled perturbations with raw and raw-plus-normalized features. Error bars show standard deviations over three seeds. Noise multipliers cannot be equated between B and C.

Table S9: Shared absolute-noise component comparison. Close-wake RMSE/L, mean $\pm$ seed standard deviation. Potential-only nonmonotonicity occurs after severe degradation.

Component	0	0.0025	0.005	0.01
Potential	0.020 ± 0.001	1.02 ± 0.07	0.97 ± 0.06	0.82 ± 0.02
Wake	0.035 ± 0.001	0.040 ± 0.001	0.053 ± 0.000	0.084 ± 0.001
Combined	0.039 ± 0.004	0.045 ± 0.004	0.059 ± 0.002	0.094 ± 0.001

S4.2 Layout redundancy and self-signal controls

The layout study evaluates six single probes, fifteen pairs, and the full six-probe array. The pair selected for lowest clean error stays fixed throughout the noise sweep. At multipliers $0, 0.0025, 0.005, 0.01$, its close-wake RMSE is $0.044, 0.058, 0.088, 0.151L$; the full array gives $0.039, 0.045, 0.058, 0.094L$. The numbers of pairs below $0.1L$ are 15, 15, 14, and 0, respectively.

An earlier subset check gives global/close-wake errors of $0.063/0.040L$ for anterior-plus-middle pairs, $0.082/0.045L$ for a side triplet, and $0.081/0.051L$ for the posterior pair. The earlier subset checks provide context for the exhaustive layout scatter, which includes every single probe and pair.

Matched observer poses isolate the effect of adding the observer's self-signal. In the aligned-orientation position task, global errors for external, self-only, and summed measurements are $0.033, 1.219, 0.038L$, respectively; close-wake errors are $0.020, 0.784, 0.026L$. The modeled self-

signal alone gives poor neighboring-position estimates. In the randomized-orientation full-state task below, summed inputs are also harder than external-only inputs.

Table S10: Randomized-orientation self-signal control, mean ± three-seed standard deviation. Position is RMSE/L and heading is close-wake MAE in degrees. This paired task is separate from the canonical component and common-prior studies.

Input	Global position	Close position	Close heading
External	0.090 ± 0.016	0.058 ± 0.004	0.71 ± 0.01
External plus self	0.132 ± 0.019	0.087 ± 0.006	1.08 ± 0.06
Self only	1.216 ± 0.009	0.783 ± 0.008	10.5 ± 3.7

S4.3 Common-prior robustness training

Common-prior noise stress tests use independent Gaussian perturbations at $\alpha \in \{0, 0.01, 0.05\}$ times the training velocity RMS, 0.213386. Each complete 32-phase trace is perturbed before overlapping windows are constructed. The same realization is shared across models and seeds. A 32-scene smoke test verifies the zero-noise construction.

Original-test 5%-noise errors are $0.945/1.068L$ for three-fish MLP/TCN and $0.929/0.992L$ for five fish. The original-test summaries average seed RMSE arithmetically. Second-test summaries below instead pool errors within the second population; the aggregation rules differ.

Augmentation draws a strength uniformly from $[0, 0.01]$ times the same training RMS for each window and redraws perturbations on each visit. Architecture, optimizer, batch size, epoch ceiling, and patience match the clean references.

Checkpoint selection also changes: validation averages clean loss and fixed-1%-noise loss, whereas clean references use clean-only validation. The validation-noise seed is 2026092507 and the test-noise seed is 2026092509. The 5% test exceeds the augmentation range and measures extrapolation under stronger corruption.

Table S11: Fresh common-prior test after the augmentation and baseline protocols were fixed. Position RMSE in body lengths on 2,000 new scenes per count. Neural rows pool three seeds. Regression and retrieval are deterministic fits. Noise standard deviation is α times the training velocity RMS.

	3 fish			5 fish		
Model	$\alpha = 0$	0.01	0.05	$\alpha = 0$	0.01	0.05
Ridge	0.805			0.956		
History retrieval	0.649			0.947		
MLP	0.472	0.549	0.941	0.802	0.833	0.978
MLP + noise training	0.468	0.509	0.870	0.803	0.825	0.955
TCN	0.492	0.753	1.064	0.793	0.854	1.012
TCN + noise training	0.498	0.546	0.862	0.794	0.813	0.923

Noisy-input accuracy improves under the combined training-and-selection procedure, while every clean-input improvement interval includes zero. Both perturbations and checkpoint selection change in this comparison; augmentation alone is not the isolated cause of the gain. Position and heading results are reported together because their responses differ.

S5 Complete failure diagnostics and sensitivity checks

S5.1 Spatial classes and error contributions

For each predicted point, let a be the closest valid source center and $\star$ the true nearest source. Padded source slots are excluded. A wrong-cell prediction satisfies $a \neq \star$; ties within $10^{-12}L^2$ in squared distance do not count as wrong-cell assignments. At radius r, near-rival predictions additionally lie within r of source a. Far-all predictions lie farther than r from every source. The remaining

Table S12: Fresh-test heading MAE and augmentation effects. Heading columns are in degrees for clean-trained and noise-trained models. Position gain is clean-trained RMSE minus noise-trained RMSE in body lengths, with a paired 95% scene interval. The first four rows are deterministic clean-input baselines.

Fish	Model	Noise	Heading	+Noise train	Pos. gain	95% interval
3	Ridge	0	12.64			
3	Retrieval	0	9.21			
5	Ridge	0	13.40			
5	Retrieval	0	12.32			
3	MLP	0	7.50	7.57	0.0040	[-0.0018, 0.0098]
3	MLP	0.01	8.51	8.19	0.0391	[0.0311, 0.0475]
3	MLP	0.05	12.22	11.67	0.0712	[0.0625, 0.0802]
3	TCN	0	7.81	7.92	-0.0065	[-0.0134, 0.0005]
3	TCN	0.01	9.74	8.73	0.2066	[0.1903, 0.2231]
3	TCN	0.05	12.52	11.63	0.2026	[0.1888, 0.2165]
5	MLP	0	11.18	11.17	-0.0009	[-0.0072, 0.0056]
5	MLP	0.01	11.34	11.26	0.0083	[0.0018, 0.0148]
5	MLP	0.05	12.70	12.46	0.0231	[0.0172, 0.0297]
5	TCN	0	11.37	11.46	-0.0010	[-0.0078, 0.0056]
5	TCN	0.01	11.74	11.58	0.0404	[0.0317, 0.0494]
5	TCN	0.05	13.14	12.62	0.0884	[0.0776, 0.0993]

points are near target. With target-assigned ties, these three radius classes are mutually exclusive and exhaustive.

Frequency counts how often a class occurs; squared-error share measures how much of the loss it carries. We obtain the latter by dividing the class's total squared target error by the overall total. Neither quantity identifies the cause of error. Assignment to a rival's cell also remains a geometric description, without implying an internal source choice or accurate localization of that rival.

Table S13: Spatial failure diagnostics across four model families and both common-prior tests. Wrong cell is an endpoint percentage. Near rival and far all report shares of total squared target error at $r = 0.25L$. Adjusted r_s correlates signal-ratio and error ranks after residualizing both against gap, bearing separation, and target radius.

Population	Fish	Model	Wrong cell	Near rival	Far all	Adj. r_s
Original	3	MLP	4.0	2.2	93.0	-0.249
Original	5	MLP	14.6	5.4	93.1	-0.603
Original	3	TCN	4.2	1.4	93.9	-0.157
Original	5	TCN	14.4	4.6	93.8	-0.591
Original	3	GRU	3.8	1.6	94.0	-0.177
Original	5	GRU	15.0	3.5	95.1	-0.604
Original	3	Transformer	4.4	2.3	93.1	-0.205
Original	5	Transformer	16.0	5.0	93.5	-0.612
Second	3	MLP	3.6	4.0	90.8	-0.281
Second	5	MLP	17.7	8.4	90.4	-0.636
Second	3	TCN	3.9	2.7	92.5	-0.209
Second	5	TCN	17.0	6.1	92.6	-0.632
Second	3	GRU	3.6	2.0	93.4	-0.219
Second	5	GRU	17.5	5.8	93.0	-0.631
Second	3	Transformer	4.0	3.6	91.7	-0.238
Second	5	Transformer	18.2	7.1	91.7	-0.624

The radius $r = 0.25L$ defines a diagnostic neighborhood, not a physical resolution limit. Figure S3 varies it from zero to $0.75L$ in $0.025L$ increments, marking $0.25L$ for reference. A separate check sets the radius to one quarter of the true-to-assigned-source separation; complete per-scene records are retained in the diagnostic artifact.

Across phase endpoints, the predicted point may switch its assigned source even though geometry and the true nearest target stay fixed. Such switches measure variation across phases; no geometric perturbation is applied in this diagnostic.

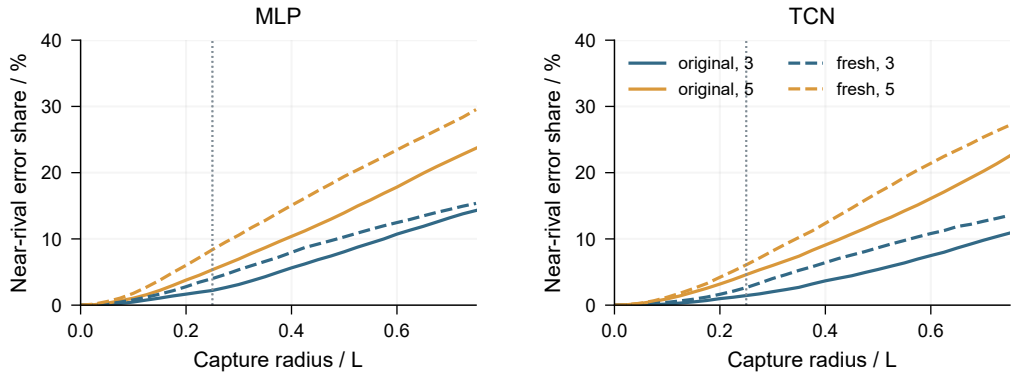

Figure S3: Radius sensitivity of near-rival squared-error share. Solid and dashed curves distinguish original and second common-prior tests; colors distinguish fish counts. The trend is expected as the capture neighborhood grows, but accurate proximity to a rival remains a minority of total squared error over these radii. All endpoints of three frozen seeds are included.

S5.2 Joint geometry bins and candidate-separation normalization

The geometry analysis crosses radial gap with bearing separation between the observer-to-nearest and observer-to-next-nearest vectors. Bearing separation is determined by source positions and is distinct from relative body heading. Radial-gap bins are $[0, 0.1)$, $[0.1, 0.5)$, and $[0.5, \infty)L$; bearing bins are $[0, 30)$, $[30, 60)$, and $[60, 180]$ degrees. Some small-gap cells contain only about ten independent scenes. Cell counts are therefore reported alongside errors: additional phases do not increase the number of independent geometries.

Table S14: Separation-normalized MLP error, with independent scene counts in parentheses. Squared endpoint errors are divided by squared nearest–next-nearest separation before aggregation and square root. This dimensionless diagnostic complements the absolute RMSE in Figure S4, which remains the main metric.

Population	Fish	Gap / L	$< 30°$	30–60°	$\geq 60°$
Original	3	< 0.1	1.00 (10)	0.63 (13)	0.49 (22)
Original	3	0.1–0.5	0.67 (49)	0.44 (63)	0.38 (76)
Original	3	≥ 0.5	0.20 (683)	0.20 (500)	0.14 (584)
Original	5	< 0.1	1.31 (19)	0.69 (48)	0.57 (48)
Original	5	0.1–0.5	1.01 (105)	0.61 (162)	0.47 (199)
Original	5	≥ 0.5	0.42 (534)	0.35 (377)	0.27 (508)
Second	3	< 0.1	0.73 (11)	0.70 (12)	0.62 (11)
Second	3	0.1–0.5	0.56 (44)	0.48 (57)	0.39 (60)
Second	3	≥ 0.5	0.20 (632)	0.19 (560)	0.13 (613)
Second	5	< 0.1	1.04 (13)	0.66 (36)	0.70 (53)
Second	5	0.1–0.5	0.92 (117)	0.60 (157)	0.54 (199)
Second	5	≥ 0.5	0.44 (517)	0.36 (429)	0.28 (479)

Normalizing each squared target error by squared nearest–next-nearest separation changes the angle ordering. Part of the absolute-error trend follows candidate separation. Normalized and absolute errors consequently supply complementary views, without designating any angle bin as intrinsically easy. Full cell-level intervals and wrong-cell statistics are retained in the analysis table.

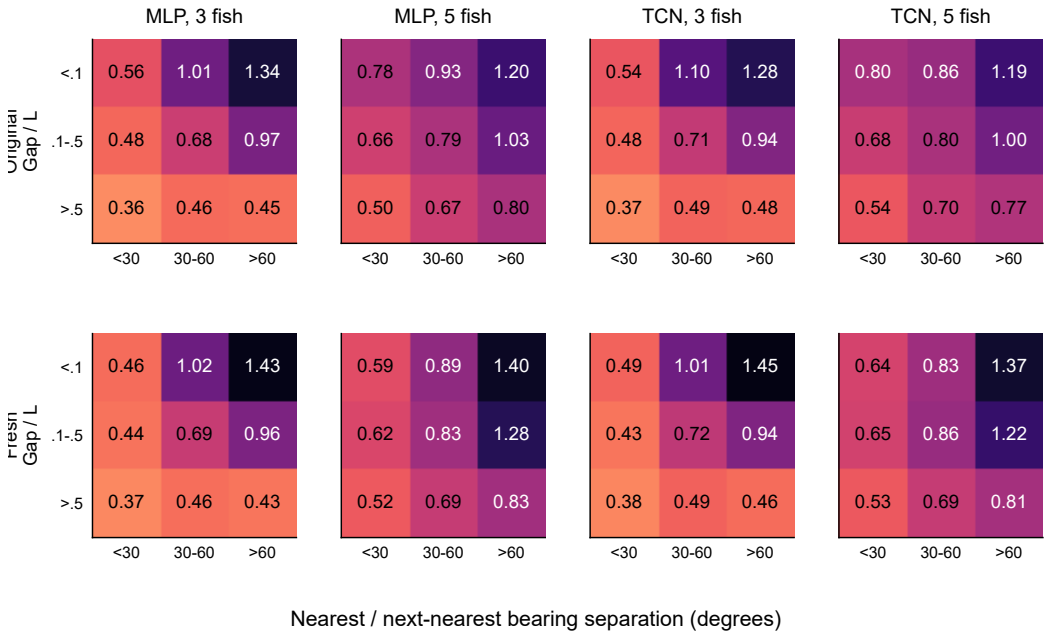

Figure S4: Absolute position RMSE/L across gap and bearing-separation bins, on both tests and for both original reference models. Every cell is retained with one shared color scale; numerical values are printed. Scene counts are shared across models within a population and appear in Table S14. The larger absolute error in some widely separated candidate configurations is not by itself an identifiability result.

S5.3 Signal reconstruction and adjusted associations

Source reconstruction uses the existing geometry and initial phase with the frozen library and kernels. Summed potential and wake contributions reproduce saved observations on both common-prior populations with maximum component discrepancy 4.77×10^{-7}. Source RMS averages squared velocity over 32 phases, six probes, and two components. Equation 2 takes the RMS of the vector-summed non-target velocity, preserving cancellation in the denominator. Neither the reconstructed source decomposition nor the ratio is supplied to the estimator.

Geometric proximity does not guarantee signal dominance. With five fish, the target is the strongest individual source in 53.45% of original scenes and about 52% of second-test scenes. Its RMS exceeds vector-summed interference in about 39% of both populations.

Ratio curves in Figure S5 use common thresholds across both reference models and populations. Each bin is placed at its population's median ratio, and its height is pooled RMSE rather than mean scene RMSE. Intervals resample complete scenes while holding fitted predictions fixed.

To adjust the association for measured geometry, we use radial gap and bearing separation together with target radius as covariates. Signal ratio and scene error are first rank-transformed. We remove their dependence on those covariates and correlate the residuals. Table S13 applies this procedure to all four architectures and both tests. Residual association can still reflect unmeasured factors; it is not an information-theoretic bound.

S5.4 Deterministic example selection and remaining alternatives

Examples in main Figure 3A are selected deterministically from original-test five-fish MLP scenes. A scene is eligible when at least half its seed-by-phase endpoints belong to the illustrated class. Eligible scenes are stably sorted by scene mean squared error, and the middle scene is selected.

The near-rival example is scene 19513: 33 scenes are eligible, 71.9% of its endpoints belong to the class, and its RMSE is $0.869L$. The far-all example is scene 18144: 1,195 scenes are eligible, 100% of its endpoints belong to the class, and its RMSE is $0.647L$. Eligibility counts describe this selection procedure, not endpoint prevalence. Plots show all three seeds and 32 phases.

The shared spatial failure pattern is the result established by the dominant far-all contribution. Explaining it is a further task: conditional averaging may be involved, but spatially biased regression and weakly informative observations remain alternatives. Source interventions test sensitivity to relative contribution without selecting a mechanism. Future source-set comparisons must match any additional label supervision; probabilistic outputs should use calibrated evaluation scores, with no best-of-many oracle selection.

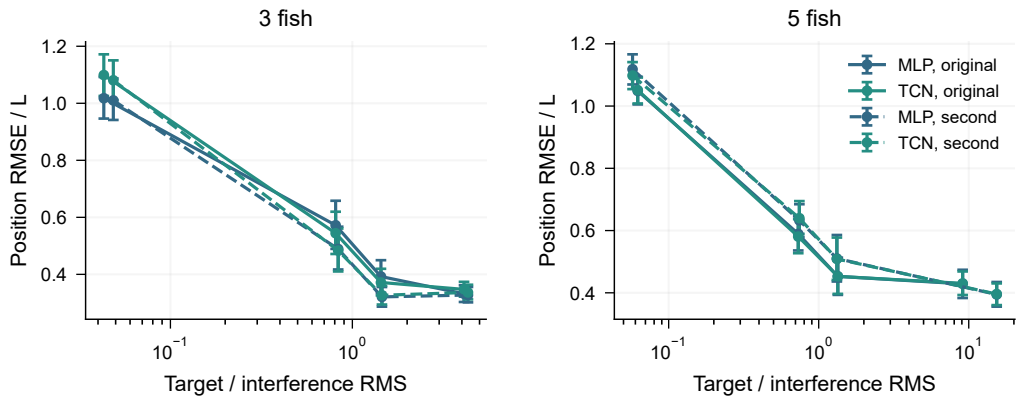

Figure S5: Target-to-interference ratio and error before interventions, for both counts, models, and existing test populations. Strata use the same ratio thresholds. Error bars resample complete scenes; x coordinates are the median ratio within each stratum. The stronger five-fish rank association cannot be explained simply by multiplying all errors by a larger constant, since ranks are scale-invariant. Different scene distributions and conditional relationships remain possible explanations.

S5.5 Paired source-contribution interventions

Before examining edited-input summaries, we recorded a protocol covering all 4,000 scenes in each previously observed population, with 2,000 scenes per count. MLP/TCN checkpoints and training normalizers remain frozen. We select the rival once, using the strongest individual non-target contribution over the original full trace, and retain it across gains. Model response plays no role in selecting scenes or sources.

For the edited mixture $\mathbf{z}^{(\alpha)}$ in Equation 3, define $g_\alpha = \mathrm{RMS}(\mathbf{z}^{(\alpha)})/\,\mathrm{RMS}(\mathbf{z})$. Each edit gives three inputs: the raw mixture $\mathbf{z}^{(\alpha)}$, the RMS-matched mixture $\mathbf{z}^{(\alpha)}/g_\alpha$, and the global-gain control $g_\alpha\mathbf{z}$. The matched input restores original total RMS; the global control preserves original source ratios.

RMS is computed over the complete velocity trace rather than speed channels or a single window. Speed is then recomputed from edited velocities before feature construction. Nearest-target position and relative-heading labels remain unchanged.

Five gains, two source choices, and three input constructions yield 25 distinct inputs per scene and model, because all identity conditions share one baseline. Each input is evaluated over all 32 phase endpoints and three seeds.

Unedited predictions reproduce the saved baselines exactly on the original GPU backend, providing a direct check before source edits are applied. Source reconstruction is within 4.77×10^{-7} of saved velocity components, and total RMS is checked for every matched edit. Independent verification reassembles selected edited inputs and checks saved predictions and aggregate differences.

Prediction movement measures displacement along the fixed rival–target axis. It is the mean of $(\widehat{\mathbf{x}}^{(\alpha)} - \widehat{\mathbf{x}}^{(1)}) \cdot \mathbf{d}$, where $\mathbf{d}$ points from the rival toward the target. Positive values indicate targetward movement in units of L; negative values indicate rivalward movement. Movement does not require reaching either source and is distinct from near-rival incidence.

Position differences and movement intervals use 2,000 paired complete-scene bootstrap draws, retaining all phases and seed predictions. As for the other exploratory diagnostics, no multiple-comparison correction is applied to these intervals.

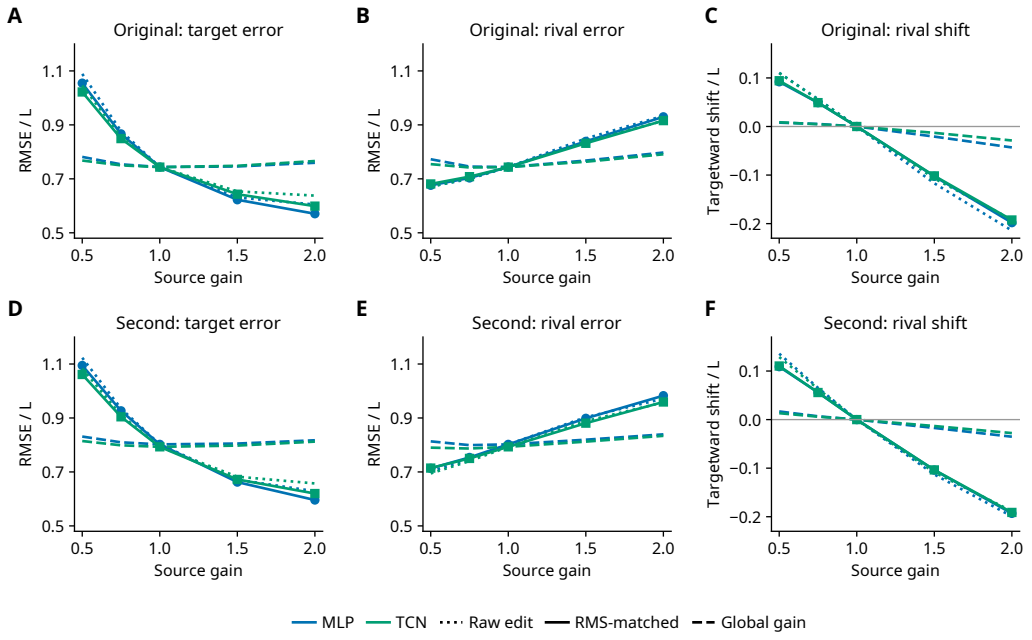

Figure S6: Five-fish source-gain responses on the original (A–C) and second (D–F) tests. Columns show target-edit RMSE, rival-edit RMSE, and rival-edit movement along the rival–target axis. Positive movement is toward the target. Colors distinguish MLP/TCN; line styles distinguish raw source edits, RMS-matched edits, and global-gain controls. Edits change modeled velocity contributions rather than simulating changes in physical propulsion. Three-fish curves and scene-cluster sufficient statistics accompany the analysis artifact. The endpoint table includes both counts.

Table S15: Change in position RMSE/L at source gain $\alpha = 2$, with total scene RMS matched to the unedited input. Negative changes improve accuracy. Brackets are paired 95% scene intervals, conditional on three fitted seeds. The target label remains nearest by geometry in every condition.

Population	Fish	Model	Target gain: ΔRMSE	Rival gain: ΔRMSE
Original	3	MLP	−0.111 [-0.129, -0.094]	0.298 [0.272, 0.325]
Original	3	TCN	−0.089 [-0.103, -0.075]	0.278 [0.253, 0.304]
Original	5	MLP	−0.173 [-0.189, -0.158]	0.186 [0.167, 0.207]
Original	5	TCN	−0.145 [-0.159, -0.131]	0.172 [0.155, 0.191]
Second	3	MLP	−0.100 [-0.115, -0.086]	0.326 [0.297, 0.356]
Second	3	TCN	−0.081 [-0.092, -0.069]	0.281 [0.255, 0.308]
Second	5	MLP	−0.206 [-0.225, -0.189]	0.180 [0.162, 0.200]
Second	5	TCN	−0.173 [-0.190, -0.157]	0.166 [0.149, 0.184]

At $\alpha = 2$ with total RMS matched, target amplification improves RMSE and rival amplification worsens it in every listed population/model/count combination. The global-gain comparisons attribute this response to relative source contribution beyond overall amplitude. The test concerns fitted estimators receiving inputs outside the original amplitude rule. It models no change in natural

swimming behavior and does not select conditional averaging as the explanation for far-all predictions; those interpretation limits also apply to the main intervention results.

S6 Distribution shift, coverage, and regional diagnostics

S6.1 Training-count coverage, scene scale, and sensing width

Table S16: Controlled multi-source sensing. Position RMSE in body lengths on 1,000 held-out scenes, pooled across three training seeds. Counts include the observer. Single trains on two fish. Mixed trains on 2/3/5 fish with an equal per-epoch window budget. Ten fish are test-only.

Fish	MLP single	MLP mixed	TCN single	TCN mixed
2	0.034	0.169	0.032	0.166
3	0.702	0.424	0.657	0.438
5	1.031	0.746	1.000	0.719
10	1.276	1.042	1.286	1.023

Mixed-count training improves performance with interference while sacrificing some two-fish accuracy. The result applies to the controlled-mixture sampler, whose target-first construction differs from common-prior nearest selection. Ten fish are absent from training, and all fish counts include the observer. The larger bank adds 12k scenes to the original 4k training scenes. Training data and

Table S17: Data-scale comparison on 2,000 fresh test scenes. Position RMSE in body lengths, pooled across three seeds. Both training sets contain mixed 2/3/5-fish data. The 16k set includes the original 4k training scenes. All comparisons use the same fresh test scenes. Data and per-epoch compute increase together.

Fish	MLP 4k	MLP 16k	TCN 4k	TCN 16k
2	0.149	0.100	0.158	0.134
3	0.398	0.343	0.400	0.358
5	0.688	0.628	0.664	0.608
10	1.049	1.009	1.022	0.984

per-epoch compute increase together, so this comparison measures their combined effect. Position improvements do not extend to every output: two-fish TCN heading MAE changes from 2.69° to 3.07°.

Table S18: Sensor-layout results on the original 1,000 test scenes. Position RMSE is in body lengths, pooled across three seeds. Mixed trains on counts 2/3/5 at width 1.0; +layout adds widths 0.7 and 1.3. Widths 0.85 and 1.15 are test-only interpolation conditions. Only probe positions change across widths; source hydrodynamics stay fixed.

Fish	Width	MLP mixed	MLP +layout	TCN mixed	TCN +layout
2	0.70	0.211	0.148	0.223	0.165
2	1.00	0.169	0.150	0.166	0.164
2	1.30	0.203	0.146	0.215	0.159
3	0.70	0.436	0.410	0.455	0.436
3	0.85	0.424	0.414	0.441	0.437
3	1.00	0.424	0.417	0.438	0.438
3	1.15	0.428	0.418	0.437	0.436
3	1.30	0.438	0.417	0.445	0.432
5	0.70	0.749	0.746	0.722	0.712
5	1.00	0.746	0.753	0.719	0.715
5	1.30	0.742	0.752	0.712	0.710
10	1.00	1.042	1.059	1.023	1.029

Layout training scales only lateral probe coordinates by 0.7, 1.0, and 1.3. Intermediate widths 0.85 and 1.15 are test-only interpolation conditions. Only sampling geometry changes; source hydrodynamics

remain the same. Gains depend on count and width; ten-fish layout augmentation does not improve the original-test position score.

S6.2 Unseen wake parameters

Wake expansion ratio V_s/U takes values 0.20, 0.25, and 0.30. Transfer compares training at the nominal ratio against training on two source ratios, evaluating an excluded ratio with equal labeled sample counts per condition. Neither target-domain test result selects the model. Close-wake position

Table S19: Unseen wake-ratio transfer. Errors are mean ± three-seed standard deviation. Two-domain training uses the listed ratios; position is global RMSE/L and heading is MAE in degrees.

Test	Train ratios	Base pos.	Two-domain pos.	Base head.	Two-domain head.
0.20	0.25+0.30	$0.340 \pm .001$	$0.139 \pm .003$	$6.26 \pm .04$	$1.43 \pm .14$
0.30	0.20+0.25	$0.276 \pm .003$	$0.156 \pm .006$	$4.77 \pm .14$	$2.03 \pm .18$

error falls from 0.257 to $0.099L$ at unseen ratio 0.20 and from 0.195 to $0.112L$ at unseen ratio 0.30. Training across wake ratios improves transfer along the tested generator parameter. The experiment covers that parameter shift only; testing other fluid simulators or physical fish would require a different evaluation.

S6.3 Canonical spatial and heading variation

The original full-state canonical MLP has global position RMSE $0.060 \pm 0.006L$ and heading MAE $0.58 \pm 0.04°$. Error varies by region: close-wake, near-side, and mid-wake position RMSE is 0.040 ± 0.002, 0.040 ± 0.003, and $0.147 \pm 0.016L$, respectively. Corresponding heading MAE is 0.53 ± 0.03, 0.54 ± 0.05, and $1.04 \pm 0.18°$. All regional summaries average seed-level metrics from the original sample-validation protocol.

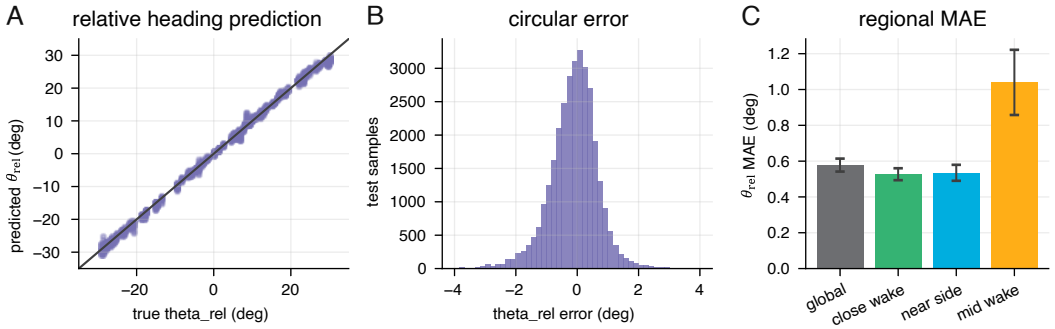

Figure S7: Canonical full-state heading diagnostics from saved temporal-MLP predictions. Scatter and wrapped residuals show relative-heading prediction; regional summaries use mean and standard deviation over the three seed-indexed evaluations. The target is body heading, not wake phase.

S7 Data access, usability, and reproduction

S7.1 Data access and package contents

The data and code are available on Hugging Face as release `iclr2027-v1.0rc2`. The link fixes the repository revision. The core package `SparseWake_core_1.0rc2.zip` contains training data for all three task families. Coverage includes 5,000 canonical poses; controlled 2/3/5-fish conditions with 4,000/1,000/1,000 base-scene splits and a 1,000-scene ten-fish test; and the common-prior 16,000/2,000/4,000-scene bank plus the separate 4,000-scene second test.

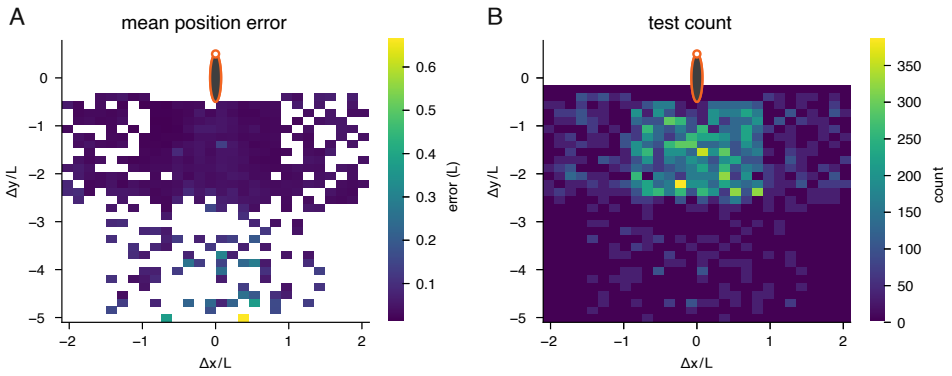

Figure S8: Spatial error and test coverage for one canonical full-state seed. The error map bins mean Euclidean position error, whereas aggregate benchmark tables report RMSE. The companion map reports held-out phase counts; empty bins have no observations. It prevents sparse coverage from being read as low error.

A shared scene-major HDF5 interface stores targets alongside split flags and geometry identifiers. Component fields are included where available, and diagnostic metadata are marked explicitly. Target conventions are preserved: canonical records use leader-centered coordinates and continuous angles; multi-source records use observer-frame position and sine–cosine heading.

The package includes a data card, schema, protocols, checksums, a trainable MLP, and an evaluator with six common-prior MLP/TCN checkpoints. Canonical splits follow the seed-1 grouped protocol. Controlled counts share base-scene identifiers and must preserve paired split assignments.

The companion reproduction archive provides generation and analysis code, the frozen wake library, auxiliary datasets, and saved experimental evidence. The release excludes the complete upstream MATLAB schooling application and DNS solver. Supplied checkpoints cover only part of the supplementary studies.

Licenses. The released data and result tables are licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). New original code is licensed under the MIT License. Previously released utilities and third-party components retain their respective licenses, as specified in the license files and third-party notices included in each archive.

S7.2 Usability checks and versioning

The portable utility reads each task, builds cyclic histories, fits training-only normalization, selects validation checkpoints, and evaluates held-out scenes. CPU smoke tests train models on all three task families with small integration-test budgets. A separate frozen-reference evaluator reproduces the main common-prior metrics. NumPy, h5py, and PyTorch are sufficient to run both checks independently of the training project and simulator.

Version `1.0-rc2` records file checksums. Changes to targets, splits, or preprocessing require a new version and an erratum identifying affected results. Error reports should include the version, checksum, command, and observed behavior. The anonymous authors maintain the release through the SparseWake Hugging Face account. Questions and error reports can be submitted through the repository's Discussions page.

S7.3 EVIDENCE-TO-ARTIFACT MAP

Table S20: Compact reproduction map. The typesetting archive and trainable core-data archive serve different purposes. Complete per-scene outputs remain analysis artifacts rather than printed appendix tables.

Evidence	Artifact and verification
Main reference scores	Core `common.h5`, `common_second_test.h5`, six checkpoints, and `evaluate_reference.py`.
New-model usability	`benchmark.py`, seven scene-major HDF5 files, schema, data card, and three task-specific CPU smoke tests.
History and sensing	Figure 2: history summaries, paired shuffle intervals, and all 22 layout records. Tables S6–S7 retain exact supporting values.
Spatial failure	Figure 3: all 16 architecture/count/test error decompositions and deterministic case identifiers; supplementary radius and source-strength records.
Source intervention	Figure 4 and Supplement S5.5: frozen protocol, source/checkpoint hashes, gain curves, scene-cluster statistics, and paired bootstrap summaries.
Formal architectures	Retained grouped-pose controls and GRU/Transformer checkpoints, separate from the earlier MLP/TCN inference audit.
Manuscript	One root `main.tex`; independently compiled archive, numerical evidence ledger, and rendered-page review.