LEARNING PROBABILISTIC AIRSPACE DEMAND FROM MICROSCOPIC AIRCRAFT INTERACTION GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Air traffic flow prediction is essential for proactive air traffic management, yet existing approaches predominantly forecast future regional flows from aggregated time-series histories. Such aggregation can obscure the states and interactions of individual aircraft that directly shape near-term traffic. We investigate whether the current airspace snapshot can instead support probabilistic demand forecasting without historical flow sequences. We introduce AEROGRAPH, which represents the variable aircraft population as a dynamic multi-relational graph with physically typed interactions and regional task nodes, and predicts discrete distributions over 15-minute aircraft counts. Across two real-world airports, microscopic snapshot-based models consistently outperform time-series models. Ablation studies identify kinematic interactions and regional task structure as the most consistent contributors. Under bidirectional cross-airport zero-shot transfer, AEROGRAPH retains its advantage under spatial-temporal domain shift. These results show that airspace snapshot modeling provides a strong alternative to aggregated time-series histories for near-term airspace demand forecasting.

1 INTRODUCTION

Short-term air traffic demand in terminal airspace is determined by individual aircraft moving through the airspace. Their positions, motion, and interactions with regional boundaries directly affect how many aircraft will appear in a region in the next few minutes. However, existing methods usually predict future demand from historical sequences of aggregated regional counts (Chen et al., 2016; Lin et al., 2019; Yan et al., 2023; Du et al., 2024). Such time series can capture recurring temporal patterns, but aggregation removes much of the aircraft-level information available at the current time. This motivates a simple representation question: *can a current airspace snapshot be used to predict the distribution of near-future regional demand without relying on historical demand time series?*

We study this question for two nested three-dimensional regions at each airport: an airspace approach region (AP) and its surrounding airspace control region (AR). At each one-minute forecast origin t, the target is the number of distinct aircraft observed in each region during $(t, t + 15\,\mathrm{min}]$. Modeling this target from a current snapshot raises two main challenges. First, the number of observed aircraft varies over time, and aircraft may interact through spatial proximity, relative motion, common intent, and flow order. These interactions provide different information about future regional demand. Second, some aircraft contributing to future demand may not yet be observed at the forecast origin. The future count is therefore uncertain, making a predictive distribution more informative than a single point prediction.

To address these challenges, we propose AEROGRAPH. At each forecast origin, AEROGRAPH represents the current aircraft population as a dynamic multi-relational graph. Aircraft are connected through four physically motivated relation types that describe spatial, kinematic, intent, and flow-order interactions. Aircraft are also connected to AP and AR task nodes to model their relations with the two prediction regions. A relation-aware graph encoder learns from these interactions, while region-conditioned demand-sensitive readouts aggregate aircraft information without removing population-size information. The model then predicts a discrete probability distribution over the 15-minute aircraft count for each region.

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

We evaluate AEROGRAPH on two real-world airports. Our experiments focus on three questions: whether aircraft-level current-state representations improve over aggregated time-series histories, whether physically typed interactions provide additional gains over set and generic graph encoders, and whether these advantages transfer across airports and observation periods. We further use ablation studies to examine the contributions of the main relation types and regional task structure.

The main contributions are as follows:

- We formulate probabilistic airspace demand forecasting from the current microscopic airspace snapshot, using aircraft-level states to predict identity-deduplicated future interval counts without relying on historical aggregate-flow sequences.
- We propose AEROGRAPH, which represents the variable aircraft population as a dynamic multi-relational graph with physically typed interactions and regional task nodes, and predicts discrete probability distributions over future regional aircraft counts.
- Across two real-world airports, we show that microscopic snapshot-based models consistently outperform aggregate time-series models, while AEROGRAPH provides further gains over strong set and generic graph encoders. These advantages persist under bidirectional zero-shot transfer across airports and observation periods.

2 RELATED WORK

Traffic flow prediction. Traffic flow forecasting typically models aggregate regional histories, sometimes augmented with graph structure, weather, or operational context (Chen et al., 2016; Lin et al., 2019; Yan et al., 2022; 2023; Du et al., 2024; Zeng et al., 2024). General-purpose forecasters such as TimesNet and iTransformer provide stronger sequence encoders (Wu et al., 2023; Liu et al., 2024). These approaches can exploit periodicity and longer histories, but address a different input-representation problem from ours. We instead ask whether the current aircraft population contains sufficient near-term evidence to forecast future regional demand directly. We focus on this representation question in short-horizon terminal-airspace forecasting.

Structural modeling. DeepSet theory (Zaheer et al., 2017; Lee et al., 2019) and GNNs provide alternatives for modeling variable-cardinality observations and structured interactions (Veličković et al., 2018; Hamilton et al., 2017; Schlichtkrull et al., 2018). Traffic graph forecasting typically assumes persistent sensor or location nodes (Li et al., 2018; Yu et al., 2018; Wu et al., 2019; Bai et al., 2020), while recent work has explored physically guided graph construction and observation-level graphs in environmental prediction and irregular time series (Hettige et al., 2024; Liu et al., 2026). Our setting differs in the prediction unit: transient aircraft are nodes, their relations are recomputed from the current state, and two regional task nodes collect evidence for interval counts. The physical relation labels serve as designed inductive biases for structuring aircraft interactions.

Predictive distribution learning. Probabilistic forecasting spans quantile, parametric, spline-based, and generative formulations (Koenker & Bassett, 1978; Salinas et al., 2020; Gasthaus et al., 2019; Li et al., 2024; Kollovieh et al., 2025). Joint-flow models can explicitly address consistency across marginals (Yalavarthi et al., 2026). We use a categorical count PMF because the target is discrete and its cumulative distribution function directly provides medians and prediction intervals. We therefore evaluate the marginal AP/AR PMFs with proper distributional scores and empirical interval diagnostics (Gneiting & Raftery, 2007; Gneiting & Katzfuss, 2014).

3 PROBLEM SETTING

Current airspace snapshot. The studied terminal airspace contains a lower approach region AP and a larger surrounding control region AR. At each one-minute forecast origin t, we construct the latest valid state of each retained aircraft, yielding the variable-cardinality unordered snapshot $S_t = \{s_i(t)\}_{i=1}^{N_t}$. Each state describes three-dimensional position, region membership, kinematics, motion relative to AP/AR boundaries, control context, local traffic, and validity indicators. Exact alignment and missing-field rules are given in Appendix A.1.

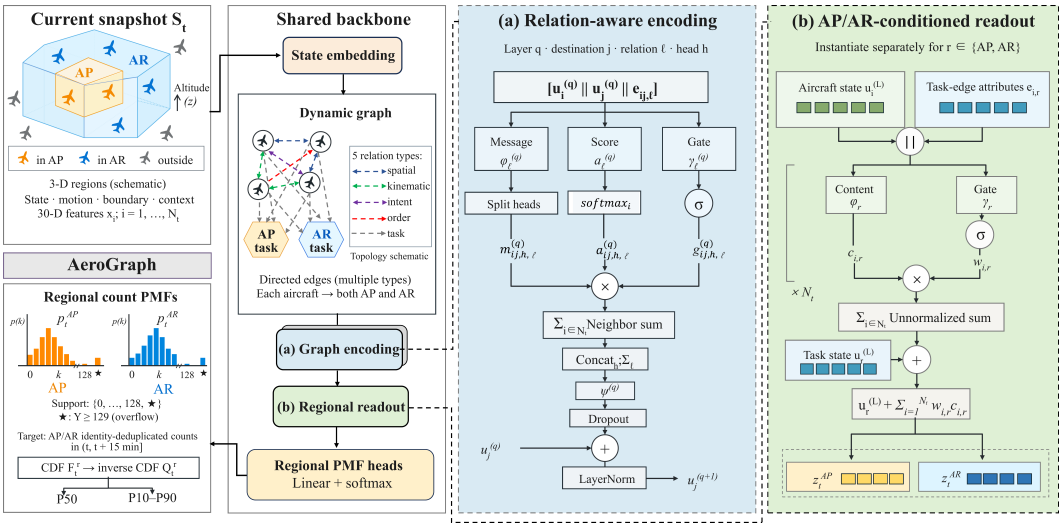

Figure 1: Overview of AEROGRAPH. A current airspace snapshot is represented as a dynamic graph with four aircraft–aircraft relations and two aircraft-to-region task edges. Relation-aware propagation produces contextualized aircraft and task states, which AP/AR-conditioned gated sums map to regional representations and discrete 15-minute demand PMFs.

AP extends from the surface to 6,000 ft and AR to 15,000 ft in the studied setting; region membership therefore depends on altitude as well as horizontal position. The snapshot retains aircraft in AP, AR, and relevant surrounding monitored airspace according to a fixed current-state rule. An aircraft outside both volumes can still carry evidence about future regional demand. Figure 3 illustrates the three-dimensional setting.

Traffic demand. Let $\Delta = 15\,\text{min}$ and let $\Omega_r \subset \mathbb{R}^3$ be region $r \in \{\text{AP}, \text{AR}\}$. For aircraft i, let $\mathcal{T}_{i,t}^+ \subset (t, t+\Delta]$ be the valid observation times used to construct its future label, and let $\mathcal{A}_t^+ = \{i : \mathcal{T}_{i,t}^+ \neq \varnothing\}$. With $\boldsymbol{\xi}_i(\tau)$ denoting the recorded position at such a time, the traffic demand is defined as

$$Y_t^r = \left|\left\{ i \in \mathcal{A}_t^+ \;\middle|\; \exists \tau \in \mathcal{T}_{i,t}^+ : \boldsymbol{\xi}_i(\tau) \in \Omega_r \right\}\right|. \tag{1}$$

Each identity contributes at most once per region, including after repeated records or re-entry. The label is defined by available future observations, not by inferred presence between records. Future observations construct labels only and never enter S_t or the prediction path. Thus, Y_t^r measures identity-deduplicated presence over the interval rather than boundary crossings or endpoint occupancy. The same aircraft may contribute once to both AP and AR, so the regional targets are not mutually exclusive.

Note that $\mathcal{A}_t^+$ need not equal the aircraft identities in S_t: an aircraft may first enter the monitored area after the forecast origin. Consequently, the target cannot be recovered by deterministically summing future-presence indicators over the current snapshot alone. Conditional uncertainty includes both the evolution of observed aircraft and demand from aircraft not yet observable. We learn regional predictive distributions $P(Y_t^r \mid S_t)$ over $\mathcal{K} = \{0, 1, \ldots, 128, \star\}$, where $\star$ denotes counts ≥ 129.

4 AEROGRAPH

Figure 1 summarizes AEROGRAPH. At each forecast origin, the model maps the airspace snapshot to two regional traffic demand distributions. A shared encoder represents aircraft states and regional task nodes, a dynamic multi-relational graph propagates typed interaction evidence, and region-conditioned gated readouts produce demand-sensitive representations for AP and AR. Regional heads then parameterize discrete probability mass functions (PMFs) over future 15-minute aircraft demand.

Table 1: Semantic construction of the five directed relation types. Exact edge-membership rules, attributes, and numerical validity conditions are fixed a priori and reported in Appendix A.3.

Relation	Semantic construction rule
Spatial	Local 3-D proximity and fixed nearest-neighbor evidence.
Kinematic	Closing motion or near-term closest-approach evidence.
Intent	Compatible region-oriented motion or similar estimated transition times.
Order	Heading-compatible directed predecessor–successor evidence along a flow.
Task	Aircraft-to-AP/AR edges carrying region-specific boundary and transition evidence.

4.1 Aircraft and Regional Task Nodes

For each retained aircraft i, we construct a 30-dimensional feature vector $\mathbf{x}_i$ from its current state $s_i(t)$. The features are grouped by physical role: (i) three-dimensional position, altitude, and region membership; (ii) ground speed, vertical rate, heading, and available control attributes; (iii) signed distance, motion trend, and current-state transition-time estimates relative to AP and AR boundaries; (iv) local traffic density and cyclic time; and (v) validity indicators. All derived quantities use information available no later than forecast origin t; causal reconstruction and missing-field handling are detailed in Appendix A.1.

Boundary features connect current geometry to the interval-demand target. Signed distance distinguishes aircraft inside a region from nearby aircraft outside it, while motion trend and transition-time estimates distinguish approaching, remaining, and departing traffic. These quantities extrapolate only the current state and do not use realized future trajectories.

A shared two-layer MLP with BatchNorm, GELU, and dropout maps each aircraft to

$$\mathbf{u}_i^{(0)} = \mathrm{MLP}(\mathbf{x}_i) \in \mathbb{R}^d, \qquad d = 128.$$

Applying the same encoder to every aircraft preserves permutation equivariance. Two learned embeddings $\mathbf{u}_{\mathrm{AP}}^{(0)}$ and $\mathbf{u}_{\mathrm{AR}}^{(0)}$ serve as regional task nodes. They are modeling anchors rather than observations, allowing a shared aircraft graph to accumulate different evidence for the two regional targets.

4.2 Dynamic Multi-Relational Aircraft Graph

At each origin, we construct

$$\mathcal{G}_t = (\mathcal{V}_t^{\mathrm{a}} \cup \{v_{\mathrm{AP}}, v_{\mathrm{AR}}\}, \{\mathcal{E}_{t,\ell}\}_{\ell \in \mathcal{R}}), \tag{2}$$

where $\mathcal{V}_t^{\mathrm{a}}$ contains the N_t aircraft nodes and each directed edge $(i, j) \in \mathcal{E}_{t,\ell}$ carries relation-specific attributes $\mathbf{e}_{ij,\ell}$. The relation set $\mathcal{R}$ contains four aircraft–aircraft relations and one aircraft-to-task relation, defined as $\mathcal{R} = \{\mathrm{spatial}, \mathrm{kinematic}, \mathrm{intent}, \mathrm{order}, \mathrm{task}\}$. Table 1 summarizes their physical semantics; exact membership thresholds, typed edge attributes, and numerical validity rules are provided in Appendix A.3.

An aircraft pair may satisfy multiple relations because the relations encode complementary evidence. Each retained aircraft has one outgoing task edge to AP and one to AR, with no reverse task edges. Transition times used in relation construction are current-state estimates rather than observed future arrivals. The graph is dynamic because its nodes, edges, and attributes are reconstructed from the current snapshot at every forecast origin, rather than learned as persistent topology.

4.3 Relation-Aware Message Passing

For propagation layer q, relation ℓ, and directed edge (i, j), we form the edge context

$$\mathbf{c}_{ij,\ell}^{(q)} = \left[\mathbf{u}_i^{(q)} \| \mathbf{u}_j^{(q)} \| \mathbf{e}_{ij,\ell}\right]. \tag{3}$$

Relation-specific networks produce a message, an attention score, and a continuous relevance gate:

$$\mathbf{m}_{ij,\ell}^{(q)} = \phi_\ell^{(q)}(\mathbf{c}_{ij,\ell}^{(q)}), \quad s_{ij,\ell}^{(q)} = a_\ell^{(q)}(\mathbf{c}_{ij,\ell}^{(q)}), \quad \mathbf{g}_{ij,\ell}^{(q)} = \sigma\left(\gamma_\ell^{(q)}(\mathbf{c}_{ij,\ell}^{(q)})\right). \tag{4}$$

Here $\mathbf{m}_{ij,\ell}^{(q)} \in \mathbb{R}^d$, while $\mathbf{s}_{ij,\ell}^{(q)}, \mathbf{g}_{ij,\ell}^{(q)} \in \mathbb{R}^H$ contain one score and gate per head. Let $\mathcal{N}_{t,\ell}(j) = \{i : (i,j) \in \mathcal{E}_{t,\ell}\}$. The message is split into H heads, and attention is normalized independently for each relation, destination, and head. For a nonempty incoming neighborhood,

$$\alpha_{ij,h,\ell}^{(q)} = \frac{\exp s_{ij,h,\ell}^{(q)}}{\sum_{k \in \mathcal{N}_{t,\ell}(j)} \exp s_{kj,h,\ell}^{(q)}}, \qquad \mathbf{M}_{j,h,\ell}^{(q)} = \sum_{i \in \mathcal{N}_{t,\ell}(j)} \alpha_{ij,h,\ell}^{(q)} g_{ij,h,\ell}^{(q)} \mathbf{m}_{ij,h,\ell}^{(q)}.$$

The destination state is updated as

$$\mathbf{u}_j^{(q+1)} = \mathrm{LayerNorm}\left[\mathbf{u}_j^{(q)} + \mathrm{Dropout}\left(\psi^{(q)}\left(\sum_{\ell \in \mathcal{R}} \mathrm{Concat}_h \, \mathbf{M}_{j,h,\ell}^{(q)}\right)\right)\right]. \tag{5}$$

A missing incoming relation contributes zero: if $\mathcal{N}_{t,\ell}(j) = \varnothing$, then $\mathbf{M}_{j,h,\ell}^{(q)} = \mathbf{0}$. Softmax attention expresses relative importance among neighbors of the same relation, whereas the independent sigmoid gate can attenuate weak evidence without transferring its weight to another neighbor. Relations are normalized separately so that edges with different physical meanings do not compete within a single softmax. The reported model uses $L = 2$ propagation layers and $H = 4$ heads, yielding 32-dimensional head messages.

4.4 Demand-Sensitive Regional Readout

After L propagation layers, each region aggregates aircraft without normalizing away their number. For $r \in \{\mathrm{AP}, \mathrm{AR}\}$, let $\mathbf{e}_{i,r}$ denote the corresponding aircraft-to-task edge attributes. We compute

$$\mathbf{c}_{i,r} = \phi_r\left([\mathbf{u}_i^{(L)} \| \mathbf{e}_{i,r}]\right), \qquad w_{i,r} = \sigma\left(\gamma_r\left([\mathbf{u}_i^{(L)} \| \mathbf{e}_{i,r}]\right)\right),$$

and form

$$\mathbf{z}_t^r = \mathbf{u}_r^{(L)} + \sum_{i=1}^{N_t} w_{i,r} \mathbf{c}_{i,r}. \tag{6}$$

Here $\mathbf{c}_{i,r} \in \mathbb{R}^{128}$ is a region-conditioned content vector and $w_{i,r} \in (0,1)$ is a scalar relevance weight. Unlike mean or softmax pooling, the weights are not constrained to sum to one, so the representation can preserve accumulated evidence from multiple relevant aircraft. AP and AR use separate content and gate functions, allowing the same aircraft to contribute differently to the two regional targets.

The task state $\mathbf{u}_r^{(L)}$ contains evidence accumulated through graph propagation, while the gated sum provides a direct region-conditioned path from all final aircraft states. This construction preserves demand-sensitive accumulated evidence while leaving the final predictor unconstrained.

4.5 Discrete Demand Distribution Learning

A regional head maps $\mathbf{z}_t^r$ to logits $\mathbf{o}_t^r$ over

$$\mathcal{K} = \{0, 1, \ldots, 128, \star\},$$

where $\star$ denotes counts of at least 129. The regional PMF is

$$p_{t,k}^r = \frac{\exp(o_{t,k}^r)}{\sum_{j \in \mathcal{K}} \exp(o_{t,j}^r)}, \qquad k \in \mathcal{K}. \tag{7}$$

For quantile extraction, index $\star$ by 129, define $F_t^r(k) = \sum_{j=0}^{k} p_{t,j}^r$ for $k \le 128$, and set $F_t^r(129) = 1$. The quantile query

$$Q_t^r(\nu) = \min\{k \in \{0, \ldots, 129\} : F_t^r(k) \ge \nu\}$$

gives the predictive median $Q_t^r(0.5)$ and, for example, the central 80% interval $[Q_t^r(0.1), Q_t^r(0.9)]$. The overflow class preserves total probability mass but does not distinguish demand above 128; a returned value of 129 therefore denotes the event $Y_t^r \ge 129$. The AP and AR heads define separate regional marginals over a shared encoder.

Training minimizes

$$\mathcal{L} = \mathcal{L}_{\mathrm{nRPS}} + \lambda_{\mathrm{N}}\mathcal{L}_{\mathrm{NLL}} + \lambda_{\mathrm{G}}\mathcal{L}_{\mathrm{gate}}, \qquad \lambda_{\mathrm{N}} = 0.05, \quad \lambda_{\mathrm{G}} = 10^{-4}. \tag{8}$$

The normalized ranked probability score (nRPS) rewards ordinal distribution quality, NLL rewards probability assigned to the observed demand class, and $\mathcal{L}_{\mathrm{gate}}$ regularizes relation-message gate activations. AP and AR receive equal weight in the distributional losses. Exact class mapping, loss definitions, and macro averaging are given in Appendix A.2.

5 EXPERIMENTS

We evaluate AEROGRAPH through three complementary settings: within-airport comparisons across observation representations and encoders, matched component ablations, and bidirectional zero-shot transfer across airport and observation period.

5.1 EXPERIMENTAL SETUP

Data and protocol. We evaluate two anonymized airports in China observed during non-overlapping periods: Airport A from May–October 2025 and Airport B from November 2025–June 2026. We use the 15-minute identity-deduplicated targets defined in Section 3 and chronological train/validation/test partitions separated by a 15-minute embargo. To ensure identical evaluation origins across model families, an origin is retained only when both its current airspace snapshot and the complete 24-hour history required by temporal baselines are available. This yields 181,556 training and 22,685 test origins for Airport A, and 276,026 training and 34,489 test origins for Airport B. Because airport and observation period change jointly, the transfer experiments evaluate a compound spatial–temporal domain shift. Figure 2 summarizes the target distributions, observation windows, training-set sizes, and common-origin coverage.

Baselines and controlled comparison. TimesNet and iTransformer take as input a two-channel 96-step AP/AR count history spanning the preceding 24 h(Wu et al., 2023; Liu et al., 2024); $\mathrm{MLP}_{Current}$ uses instantaneous AP/AR occupancy, and MLP_{Hist} concatenates the flattened 96-step AP/AR count history with the current AP/AR occupancies. DeepSets and Set Transformer operate on the same aircraft features as AEROGRAPH without explicit aircraft–aircraft interactions (Zaheer et al., 2017; Lee et al., 2019). GIN and GATv2 operate on their deduplicated union; and R-GCN preserves relation identities through a generic relational architecture (Xu et al., 2019; Brody et al., 2022; Schlichtkrull et al., 2018). $\mathrm{AEROGRAPH}_{homo}$ is a homogeneous variant of AEROGRAPH, which collapses the four aircraft–aircraft relation types into one.

All microscopic encoders share the same aircraft set and features, AP/AR task nodes, aircraft-to-task edges, task-conditioned regional readout, categorical PMF head, and evaluation origins. Their comparison therefore primarily varies the treatment of aircraft–aircraft interactions. All primary probabilistic models use the categorical support and nRPS+NLL objective defined in Section 4.5; AEROGRAPH additionally retains its relation-gate regularizer. Models are trained from scratch using five prespecified seeds, and PMF checkpoints are selected by validation macro nWIS. TimesNet, iTransformer, and DeepSets are additionally evaluated with native point heads to test whether the representation-level ordering depends on the PMF output. Optimization and tuning details are provided in Appendix A.7.

Metrics. Macro normalized weighted interval score (nWIS) is the primary distributional metric, with AP and AR normalized by their corresponding target-airport training-label means and weighted equally. We additionally report nRPS, NLL, and regional P50 MAE/RMSE. Raw interval coverage and mean prediction-interval width (MPIW) at 50%, 80%, and 90% are reported in Appendix A.6; no post-hoc calibration is applied.

5.2 WITHIN-AIRPORT FORECASTING

Microscopic state versus aggregate history. Table 2 shows a clear representation-level separation: aircraft-level current-state models substantially outperform aggregate-history and count-only baselines at both airports. For example, AEROGRAPH reduces macro nWIS from 0.05608 to

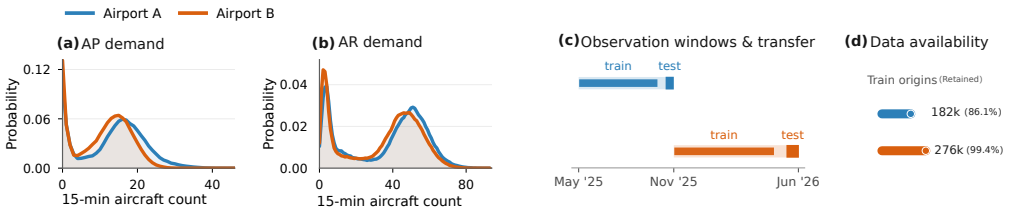

Figure 2: Two-airport experimental context. (a–b) Empirical AP and AR target distributions. (c) Non-overlapping observation windows used for within-airport evaluation and bidirectional zero-shot transfer. (d) Training-set sizes and common-origin coverage.

Table 2: Within-airport forecasting (five-seed mean ± s.d.). Reported nRPS and nWIS values are multiplied by 10^3 and 10^2, respectively. Lower is better; bold marks the best value within each airport and head family.

	Point Forecasting				Probabilistic Forecasting		
Model	$\mathrm{MAE}_{\mathrm{AP}}$	$\mathrm{RMSE}_{\mathrm{AP}}$	$\mathrm{MAE}_{\mathrm{AR}}$	$\mathrm{RMSE}_{\mathrm{AR}}$	nRPS	NLL	nWIS
(a) Airport A							
Models with point forecast heads							
TimesNet	1.962 ± 0.052	2.602 ± 0.064	3.401 ± 0.158	4.541 ± 0.214	—	—	—
iTransformer	1.884 ± 0.039	2.510 ± 0.055	3.475 ± 0.135	4.623 ± 0.216	—	—	—
DeepSets	1.304 ± 0.028	1.852 ± 0.046	1.675 ± 0.126	2.278 ± 0.138	—	—	—
Models with probabilistic forecast heads							
TimesNet	1.859 ± 0.016	2.538 ± 0.020	3.210 ± 0.081	4.315 ± 0.116	13.923 ± 0.265	2.428 ± 0.019	6.384 ± 0.095
iTransformer	1.824 ± 0.027	2.501 ± 0.039	3.404 ± 0.094	4.575 ± 0.145	14.409 ± 0.301	2.462 ± 0.020	6.504 ± 0.109
$\mathrm{MLP}_{Current}$	1.897 ± 0.063	2.592 ± 0.087	3.629 ± 0.154	4.856 ± 0.182	15.238 ± 0.407	2.481 ± 0.016	6.891 ± 0.178
MLP_{Hist}	1.640 ± 0.031	2.262 ± 0.046	2.732 ± 0.106	3.673 ± 0.172	12.053 ± 0.456	2.310 ± 0.044	5.608 ± 0.207
DeepSets	1.257 ± 0.013	1.779 ± 0.014	1.477 ± 0.121	2.032 ± 0.154	7.558 ± 0.348	1.859 ± 0.036	3.824 ± 0.114
Set Transformer	1.346 ± 0.037	1.903 ± 0.053	1.462 ± 0.046	2.010 ± 0.050	7.742 ± 0.096	1.878 ± 0.014	3.993 ± 0.064
GIN	1.429 ± 0.060	2.011 ± 0.093	1.582 ± 0.026	2.175 ± 0.036	8.300 ± 0.128	1.953 ± 0.009	4.256 ± 0.113
GATv2	1.306 ± 0.030	1.853 ± 0.041	1.419 ± 0.048	1.955 ± 0.064	7.517 ± 0.198	1.856 ± 0.024	3.870 ± 0.097
R-GCN	1.263 ± 0.021	1.795 ± 0.028	1.398 ± 0.052	1.932 ± 0.064	7.331 ± 0.170	1.823 ± 0.021	3.761 ± 0.075
$\mathrm{AeroGraph}_{homo}$	1.263 ± 0.037	1.798 ± 0.053	1.373 ± 0.014	1.903 ± 0.015	7.234 ± 0.135	1.816 ± 0.020	3.736 ± 0.089
AeroGraph (Ours)	**1.224 ± 0.005**	**1.746 ± 0.006**	**1.364 ± 0.024**	**1.889 ± 0.026**	**7.125 ± 0.050**	**1.807 ± 0.005**	**3.649 ± 0.016**
(b) Airport B							
Models with point forecast heads							
TimesNet	1.958 ± 0.045	2.617 ± 0.060	3.148 ± 0.036	4.203 ± 0.038	—	—	—
iTransformer	1.933 ± 0.029	2.604 ± 0.048	3.252 ± 0.073	4.308 ± 0.095	—	—	—
DeepSets	1.520 ± 0.021	2.062 ± 0.030	1.693 ± 0.078	2.165 ± 0.072	—	—	—
Models with probabilistic forecast heads							
TimesNet	1.890 ± 0.009	2.609 ± 0.019	3.200 ± 0.043	4.302 ± 0.062	14.048 ± 0.113	2.457 ± 0.010	7.706 ± 0.033
iTransformer	1.910 ± 0.024	2.627 ± 0.024	3.309 ± 0.050	4.411 ± 0.042	14.455 ± 0.239	2.514 ± 0.047	7.902 ± 0.151
$\mathrm{MLP}_{Current}$	1.849 ± 0.023	2.555 ± 0.034	2.841 ± 0.066	3.858 ± 0.099	12.851 ± 0.238	2.342 ± 0.016	7.176 ± 0.113
MLP_{Hist}	1.689 ± 0.023	2.323 ± 0.034	2.590 ± 0.038	3.473 ± 0.052	11.788 ± 0.089	2.304 ± 0.010	6.636 ± 0.073
DeepSets	1.369 ± 0.013	1.935 ± 0.018	1.453 ± 0.023	2.027 ± 0.035	7.772 ± 0.092	1.893 ± 0.012	4.823 ± 0.046
Set Transformer	1.330 ± 0.016	1.879 ± 0.023	1.409 ± 0.014	1.971 ± 0.021	7.551 ± 0.049	1.862 ± 0.009	4.699 ± 0.041
GIN	1.429 ± 0.021	2.019 ± 0.030	1.588 ± 0.039	2.228 ± 0.055	8.374 ± 0.135	1.965 ± 0.020	5.176 ± 0.070
GATv2	1.336 ± 0.015	1.885 ± 0.021	1.446 ± 0.026	2.007 ± 0.036	7.645 ± 0.090	1.874 ± 0.013	4.715 ± 0.047
R-GCN	1.378 ± 0.017	1.944 ± 0.024	1.451 ± 0.030	2.049 ± 0.042	7.807 ± 0.105	1.895 ± 0.015	4.853 ± 0.054
$\mathrm{AeroGraph}_{homo}$	1.321 ± 0.007	1.871 ± 0.010	1.407 ± 0.022	1.971 ± 0.031	7.529 ± 0.070	1.861 ± 0.012	4.687 ± 0.035
AeroGraph (Ours)	**1.302 ± 0.007**	**1.844 ± 0.009**	**1.376 ± 0.020**	**1.926 ± 0.027**	**7.380 ± 0.060**	**1.844 ± 0.008**	**4.597 ± 0.028**

0.03649 relative to MLP_{Hist} at Airport A, and from 0.06636 to 0.04597 at Airport B. $\mathrm{MLP}_{Current}$ remains well behind the microscopic encoders, indicating that instantaneous regional occupancy alone does not account for the gain.

The same ordering persists under native regression heads: DeepSets outperforms both temporal encoders at both airports, showing that the microscopic-state advantage is not specific to the categorical PMF parameterization.

Typed graph refinement. Within the microscopic family, AeroGraph further improves over set and generic graph encoders. $\mathrm{AeroGraph}_{homo}$ obtains macro nWIS of 0.03736 and 0.04687 at Airports A and B, compared with 0.03649 and 0.04597 for AeroGraph, with the same broad

ordering in nRPS and NLL. Because the aircraft representation, regional task interface, PMF head, objective, and evaluation origins are shared, these results attribute an additional gain to the treatment of aircraft interactions rather than to the output formulation.

Probabilistic forecasting performance. Lower proper scores are not obtained simply by widening prediction intervals. At Airport A, for example, AEROGRAPH's 80% AP interval is substantially narrower than TimesNet's (3.689 vs. 5.704) at similar coverage (0.869 vs. 0.860). At similar coverage, AEROGRAPH produces substantially sharper intervals together with better proper scores; complete coverage and width diagnostics are reported in Appendix A.6.

5.3 ABLATION STUDY ACROSS AIRPORTS

Table 3: Core relation and graph/task-structure ablations across airports (five-seed mean ± s.d.). Reported nRPS and nWIS values are multiplied by 10^3 and 10^2, respectively. Lower is better; bold marks the best value within each airport panel.

	Point Forecasting				Distribution Quality		
Model	$\mathrm{MAE}_{\mathrm{AP}}$	$\mathrm{RMSE}_{\mathrm{AP}}$	$\mathrm{MAE}_{\mathrm{AR}}$	$\mathrm{RMSE}_{\mathrm{AR}}$	nRPS	NLL	nWIS
(a) Airport A							
Reference							
AEROGRAPH (Ours)	1.224 ± 0.005	1.746 ± 0.006	**1.364 ± 0.024**	**1.889 ± 0.026**	**7.125 ± 0.050**	**1.807 ± 0.005**	3.649 ± 0.016
Relation structure							
w/o spatial relation	**1.217 ± 0.012**	**1.731 ± 0.019**	1.378 ± 0.009	1.903 ± 0.022	7.151 ± 0.058	1.812 ± 0.013	**3.642 ± 0.037**
w/o kinematic relation	1.257 ± 0.013	1.788 ± 0.030	1.397 ± 0.021	1.930 ± 0.029	7.288 ± 0.062	1.830 ± 0.009	3.736 ± 0.029
w/o intent relation	1.241 ± 0.012	1.764 ± 0.019	1.365 ± 0.018	1.909 ± 0.025	7.205 ± 0.056	1.823 ± 0.007	3.685 ± 0.025
w/o order relation	1.233 ± 0.014	1.757 ± 0.018	1.402 ± 0.040	1.940 ± 0.048	7.233 ± 0.053	1.825 ± 0.003	3.680 ± 0.012
w/o task relation	1.247 ± 0.014	1.776 ± 0.018	1.392 ± 0.015	1.929 ± 0.016	7.277 ± 0.093	1.823 ± 0.016	3.725 ± 0.046
Graph with task-only relation	1.249 ± 0.023	1.777 ± 0.031	1.396 ± 0.026	1.935 ± 0.026	7.256 ± 0.058	1.834 ± 0.006	3.698 ± 0.041
(b) Airport B							
Reference							
AEROGRAPH (Ours)	**1.302 ± 0.007**	**1.844 ± 0.009**	**1.376 ± 0.020**	**1.926 ± 0.027**	**7.380 ± 0.060**	**1.844 ± 0.008**	**4.597 ± 0.028**
Graph with multi-relation structure							
w/o spatial relation	1.331 ± 0.014	1.878 ± 0.016	1.398 ± 0.007	1.946 ± 0.015	7.520 ± 0.065	1.863 ± 0.013	4.691 ± 0.042
w/o kinematic relation	1.342 ± 0.015	1.889 ± 0.027	1.405 ± 0.021	1.972 ± 0.031	7.555 ± 0.068	1.868 ± 0.009	4.695 ± 0.033
w/o intent relation	1.337 ± 0.014	1.932 ± 0.017	1.397 ± 0.018	1.952 ± 0.027	7.495 ± 0.062	1.862 ± 0.007	4.668 ± 0.029
w/o order relation	1.341 ± 0.015	1.885 ± 0.016	1.408 ± 0.040	1.977 ± 0.052	7.525 ± 0.058	1.864 ± 0.003	4.675 ± 0.014
w/o task relation	1.324 ± 0.019	1.869 ± 0.019	1.397 ± 0.015	1.947 ± 0.018	7.502 ± 0.104	1.860 ± 0.017	4.671 ± 0.058
Graph with task-only relation	1.325 ± 0.020	1.872 ± 0.020	1.403 ± 0.027	2.012 ± 0.028	7.543 ± 0.081	1.870 ± 0.012	4.691 ± 0.049

Table 3 identifies kinematic interaction as the most consistent relation-level contributor across both airports: removing it degrades both point and distributional scores. Spatial, intent, and order relations show more airport-dependent contributions.

Regional task structure also contributes significantly: neither the task-only graph nor removal of aircraft-to-task edges matches the full model across both airports. Additional feature and regularization ablations supporting these choices are reported in Appendix A.4; complete interval diagnostics for all variants are reported in Appendix A.6.

5.4 BIDIRECTIONAL ZERO-SHOT TRANSFER

Transfer protocol. For A→B, all model parameters, normalization statistics, hyperparameters, and checkpoint choices are determined on Airport A and frozen before evaluation on Airport B; B→A reverses the procedure. Target-airport geometry may define AP/AR task features, but target labels are never used for training, fine-tuning, calibration, model selection, or post-processing.

Cross-domain performance. AEROGRAPH achieves the lowest mean across all seven reported metrics in both zero-shot directions. Its macro nWIS is 0.04804 on A→B versus 0.04938 for AEROGRAPH_{homo}, and 0.04493 on B→A versus 0.04587, with the same advantage extending to AP/AR point errors, nRPS, and NLL. Thus, the benefit of microscopic relational modeling persists when both airport and observation period change, without target-label adaptation. Because nWIS is normalized by target-airport scales, raw values across targets are not directly comparable; relative to the corresponding same-target references, the transfer cost is direction-dependent.

Table 4: Representative-family bidirectional zero-shot transfer (five-seed mean ± s.d.). Reported nRPS and nWIS values are multiplied by 10^3 and 10^2, respectively. Lower is better; bold marks the best value within each direction.

	Point Forecasting				Distribution Quality		
Model	$\mathrm{MAE}_{\mathrm{AP}}$	$\mathrm{RMSE}_{\mathrm{AP}}$	$\mathrm{MAE}_{\mathrm{AR}}$	$\mathrm{RMSE}_{\mathrm{AR}}$	nRPS	NLL	nWIS
(a) Airport A (train) → Airport B (test)							
TimesNet	1.935 ± 0.020	2.640 ± 0.023	3.284 ± 0.044	4.402 ± 0.060	14.327 ± 0.173	2.461 ± 0.012	7.798 ± 0.089
MLP_{Hist}	1.779 ± 0.016	2.413 ± 0.024	2.785 ± 0.079	3.732 ± 0.104	12.591 ± 0.264	2.364 ± 0.024	7.037 ± 0.120
DeepSets	1.512 ± 0.044	2.110 ± 0.059	1.710 ± 0.162	2.318 ± 0.204	8.863 ± 0.515	1.985 ± 0.048	5.430 ± 0.226
Set Transformer	1.591 ± 0.089	2.210 ± 0.117	1.653 ± 0.113	2.255 ± 0.142	8.922 ± 0.311	1.995 ± 0.032	5.582 ± 0.228
R-GCN	1.425 ± 0.035	1.999 ± 0.045	1.528 ± 0.070	2.104 ± 0.085	8.113 ± 0.220	1.906 ± 0.022	5.020 ± 0.111
$\mathrm{AEROGRAPH}_{homo}$	1.409 ± 0.034	1.976 ± 0.038	1.473 ± 0.043	2.036 ± 0.047	7.922 ± 0.187	1.891 ± 0.024	4.938 ± 0.111
AEROGRAPH (Ours)	**1.356 ± 0.010**	**1.911 ± 0.009**	**1.452 ± 0.005**	**2.008 ± 0.006**	**7.743 ± 0.034**	**1.885 ± 0.005**	**4.804 ± 0.023**
(b) Airport B (train) → Airport A (test)							
TimesNet	1.925 ± 0.038	2.653 ± 0.059	3.270 ± 0.023	4.370 ± 0.045	23.795 ± 0.129	2.521 ± 0.009	7.365 ± 0.082
MLP_{Hist}	1.708 ± 0.026	2.335 ± 0.041	2.662 ± 0.050	3.549 ± 0.065	21.477 ± 0.111	2.359 ± 0.010	6.381 ± 0.052
DeepSets	1.355 ± 0.015	1.899 ± 0.020	1.453 ± 0.038	2.004 ± 0.060	8.193 ± 0.147	1.918 ± 0.019	4.699 ± 0.062
Set Transformer	1.320 ± 0.008	1.848 ± 0.015	1.429 ± 0.030	1.982 ± 0.054	8.023 ± 0.064	1.886 ± 0.009	4.608 ± 0.022
R-GCN	1.373 ± 0.018	1.922 ± 0.018	1.484 ± 0.026	2.057 ± 0.037	8.317 ± 0.124	1.925 ± 0.014	4.756 ± 0.056
$\mathrm{AEROGRAPH}_{homo}$	1.312 ± 0.010	1.850 ± 0.002	1.412 ± 0.016	1.964 ± 0.030	7.966 ± 0.047	1.872 ± 0.006	4.587 ± 0.014
AEROGRAPH (Ours)	**1.282 ± 0.025**	**1.813 ± 0.036**	**1.370 ± 0.009**	**1.898 ± 0.015**	**7.773 ± 0.075**	**1.849 ± 0.009**	**4.493 ± 0.032**

Directional asymmetry. Airport B provides 276,026 training origins, about 52% more than Airport A, and retains 99.43% of common origins versus 86.13%. Its training traffic state is also closer to the Airport-A test regime, especially in AR mean (33.98 vs. 33.84), and Airport-B training begins immediately after the Airport-A test period, whereas A-trained models are evaluated on Airport B roughly eight months later. These differences in source size, traffic-state similarity, and temporal proximity provide a plausible context for the asymmetry. Most importantly, AEROGRAPH preserves its relative advantage in both directions under simultaneous geographic and temporal shift. Additional PMF comparators and native temporal-head results are reported in Appendix A.5; complete interval diagnostics are reported in Appendix A.6.

6 Discussion and Conclusion

Across both airports, the largest empirical separation among the tested models is between observation representations: aircraft-level current state consistently outperforms aggregate-history and count-only alternatives. Within this feature- and head-matched microscopic setting, typed graph structure provides further gains, with particularly consistent evidence for kinematic interactions and regional task structure. These results support physically typed interaction as a useful inductive bias.

The current formulation models AP and AR as separate marginal distributions, motivating joint region-aware forecasting as a natural extension. The zero-shot setting combines airport and temporal shift, providing a demanding test of cross-domain robustness. Future-unseen aircraft constitute an irreducible component of predictive uncertainty.

Overall, microscopic current state provides the dominant near-term predictive signal, while physically typed graph structure adds complementary gains. Future work should disentangle geographic from temporal transfer using temporally aligned multi-airport data, extend the marginal regional outputs toward joint region-aware probabilistic forecasting, and develop explicitly domain-robust training procedures.

AI USE STATEMENT

Generative AI tools were used during this work in three supporting capacities. First, they were used for language editing and manuscript organization, including improving clarity, concision, and consistency of technical writing. Second, they were used to assist literature retrieval and discovery by suggesting potentially relevant research directions, keywords, and prior work; all cited papers were subsequently located, read, and verified by the authors. Third, generative AI was used as an interactive research assistant during parts of research ideation and execution, including brainstorming alternative modeling and experimental formulations, discussing implementation choices, and assisting with code drafting, debugging, and review.Figure 3 was generated using a generative AI image tool and was reviewed by the authors for scientific accuracy and consistency with the manuscript. The authors independently determined the final problem formulation, model design, experimental protocol, and interpretation of the results, and executed and verified the reported experiments. AI-generated suggestions, text, code, and references were not treated as authoritative and were reviewed and, where necessary, modified or discarded by the authors. The authors take full responsibility for the technical content, results, claims, citations, and final manuscript.

ETHICS STATEMENT

This work studies probabilistic forecasting in a safety-sensitive air traffic setting. The proposed method is intended for research on short-horizon regional demand forecasting and does not produce certified capacity limits, separation instructions, trajectory commands, or autonomous control actions. Human operators therefore retain full decision authority. The operational data used in this study are anonymized and subject to owner-controlled access and release restrictions. We do not claim that improvements in forecasting metrics directly establish operational safety, efficiency, or controller-benefit guarantees.

REPRODUCIBILITY STATEMENT

Appendix A provides the main details required to reproduce the methodology, including chronological data splits and embargo rules, causal state reconstruction, feature and label definitions, graph construction, model objectives, evaluation metrics, optimization settings, and the zero-shot transfer protocol. All reported trainable-model results use five prespecified random seeds, and the experimental software and hardware environment is documented in the appendix. We additionally provide an anonymous reference implementation at `https://anonymous.4open.science/r/aerograph-uq-2027/`, containing the main graph-construction, model, loss, metric, preprocessing, training, configuration, and testing components used in this work. The raw operational trajectories, exact airspace geometry, and site-specific data-processing components cannot be publicly redistributed because of data-owner restrictions. Therefore, the released materials support inspection and reproduction of the proposed methodology, while exact numerical reproduction of the reported results requires access to the restricted operational data.

REFERENCES

Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in Neural Information Processing Systems*, 33:17804–17815, 2020.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=F72ximsx7C1`.

Dan Chen, Minghua Hu, Yuanyuan Ma, and Jianan Yin. A network-based dynamic air traffic flow model for short-term en route traffic prediction. *Journal of Advanced Transportation*, 50(8):2174–2192, 2016. doi: 10.1002/atr.1453.

Wenbo Du, Shenwen Chen, Zhishuai Li, Xianbin Cao, and Yisheng Lv. A spatial–temporal approach for multi-airport traffic flow prediction through causality graphs. *IEEE Transactions on Intelligent Transportation Systems*, 25(1):532–544, 2024. doi: 10.1109/TITS.2023.3308903.

Jan Gasthaus, Konstantinos Benidis, Yuyang Wang, Syama Sundar Rangapuram, David Salinas, Valentin Flunkert, and Tim Januschowski. Probabilistic forecasting with spline quantile function RNNs. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1901–1910, Naha, Japan, 2019. PMLR.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014. doi: 10.1146/annurev-statistics-062713-085831.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1024–1034, Long Beach, CA, USA, 2017. Curran Associates, Inc.

Kethmi Hirushini Hettige, Jiahao Ji, Shili Xiang, Cheng Long, Gao Cong, and Jingyuan Wang. AirPhyNet: Harnessing physics-guided neural networks for air quality prediction. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c52cfe4455c441c691663a2ff5b0b2b1-Abstract-Conference.html.

Roger Koenker and Gilbert Bassett, Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi: 10.2307/1913643.

Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow matching with gaussian process priors for probabilistic time series forecasting. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ee1a1ecc92f35702b5c29dad3dc909ea-Abstract-Conference.html.

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

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, Vancouver, BC, Canada, 2018. OpenReview.net. URL https://openreview.net/forum?id=SJiHXGWAZ.

Yuxin Li, Wenchao Chen, Xinyue Hu, Bo Chen, Baolin Sun, and Mingyuan Zhou. Transformer-modulated diffusion models for probabilistic multivariate time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/516a9317af9d89e9f2251bd7fde49b8f-Abstract-Conference.html.

Yi Lin, Jian-Wei Zhang, and Hong Liu. Deep learning based short-term air traffic flow prediction considering temporal–spatial correlation. *Aerospace Science and Technology*, 93:105113, 2019. doi: 10.1016/j.ast.2019.04.021.

Xvyuan Liu, Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Chenjuan Guo, Bin Yang, and Jilin Hu. ASTGI: Adaptive spatio-temporal graph interactions for irregular multivariate time series forecasting. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/28d38c036365420f61ce03300418e44a-Abstract-Conference.html.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, Vienna, Austria, 2024. OpenReview.net. URL https://openreview.net/forum?id=JePfAI8fah.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020. doi: 10.1016/j.ijforecast.2019.07.001.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web*, volume 10843 of *Lecture Notes in Computer Science*, pp. 593–607, Cham, Switzerland, 2018. Springer. doi: 10.1007/978-3-319-93417-4_38.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, Vancouver, BC, Canada, 2018. OpenReview.net. URL https://openreview.net/forum?id=rJXMpikCZ.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, Kigali, Rwanda, 2023. OpenReview.net. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph WaveNet for deep spatial–temporal graph modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 1907–1913, Macao, China, 2019. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2019/264.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Vijaya Krishna Yalavarthi, Randolf Scholz, Christian Klötergens, Kiran Madhusudhanan, Stefan Born, and Lars Schmidt-Thieme. Reliable probabilistic forecasting of irregular time series through marginalization-consistent flows. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/899c6a43b9976e1077522fe5a39cafa3-Abstract-Conference.html.

Zhen Yan, Hongyu Yang, Fan Li, and Yi Lin. A deep learning approach for short-term airport traffic flow prediction. *Aerospace*, 9(1):11, 2022. doi: 10.3390/aerospace9010011.

Zhen Yan, Hongyu Yang, Yuankai Wu, and Yi Lin. A multi-view attention-based spatial–temporal network for airport arrival flow prediction. *Transportation Research Part E: Logistics and Transportation Review*, 170: 102997, 2023. doi: 10.1016/j.tre.2022.102997.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 3634–3640, Stockholm, Sweden, 2018. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2018/505.

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, pp. 3391–3401, Long Beach, CA, USA, 2017. Curran Associates, Inc.

Yang Zeng, Minghua Hu, Haiyan Chen, Ligang Yuan, Sameer Alam, and Dabin Xue. Improved air traffic flow prediction in terminal areas using a multimodal spatial–temporal network for weather-aware (MST-WA) model. *Advanced Engineering Informatics*, 62:102935, 2024. doi: 10.1016/j.aei.2024.102935.

A Reproducibility Details

A.1 Airspace, Data, and Causal State Reconstruction

The approach airspace (AP) is the lower terminal volume surrounding the airport, extending from the surface to 6,000 ft in the studied setting. The airspace control region (AR) is the larger surrounding volume extending to 15,000 ft. "Outside" denotes monitored airspace outside both 3-D region boundaries. Figure 3 is a schematic rather than a map to scale.

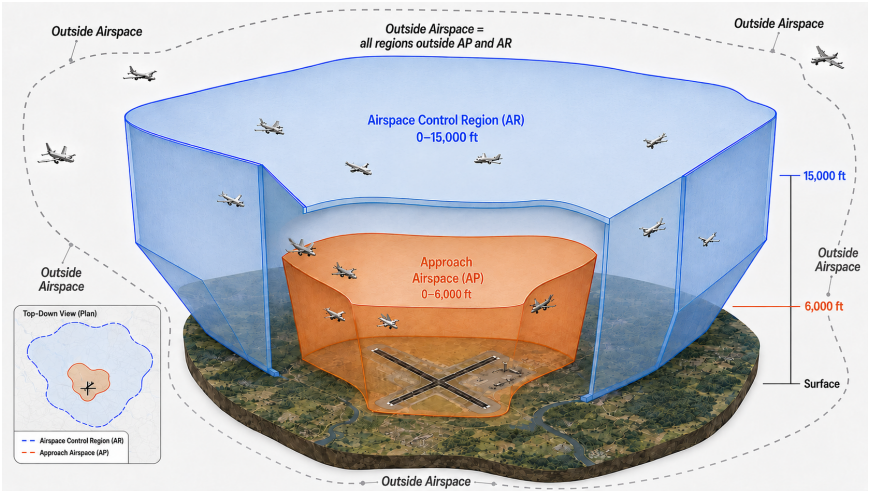

Figure 3: Studied terminal-airspace setting. The snapshot contains aircraft in AP, AR, and relevant surrounding monitored airspace at the forecast origin.

At alignment time u, a continuous state x_i is interpolated only between valid same-aircraft observations available no later than forecast origin t:

$$\widetilde{x}_i(u) = (1-\lambda)x_i(\tau_1) + \lambda x_i(\tau_2), \qquad \lambda = \frac{u-\tau_1}{\tau_2-\tau_1}, \quad \tau_1 \le u \le \tau_2 \le t. \tag{9}$$

Heading is interpolated on the unit circle. Categorical, region-membership, and control states use last observation carried forward. If the latest state precedes the origin, position and altitude are propagated for at most eight seconds with the latest ground velocity and vertical rate. Longer gaps and origins without a valid reconstruction are rejected. Commanded altitude is imputed from observed height when unavailable; missing vertical rate is first estimated from finite differences of same-aircraft height. Every reconstructed or imputed field retains a validity indicator.

Table 5: Event-time reconstruction and missing-field handling.

Field group	Missing rate	Alignment / imputation	Max. gap	Indicator
Position / height	$\le .073\%$	Causal interpolation; short propagation	8 s	Yes
Ground speed	$.000033\%$	Causal interpolation	8 s	Yes
Heading	$.022\%$	Unit-circle interpolation	8 s	Yes
Vertical rate	66.42%	Height difference; otherwise zero	8 s	Yes
Commanded altitude	37.84–71.74%	LOCF; otherwise height	8 s	Yes
Categorical / control	Field dependent	LOCF	8 s	Yes
Mach / commanded speed	71.56% / 89.08%	Excluded	—	—

A negative signed boundary distance denotes a state inside the corresponding 3-D region. Motion trend and transition-time features extrapolate the current state relative to that boundary; they never use a realized future trajectory. Local-density features count neighbors at fixed spatial scales and do not retrieve aggregate-flow history. The aircraft encoder applies BatchNorm over aircraft nodes

in the training mini-batch and uses running statistics at inference; learned regional task embeddings bypass this normalization. This network normalization is distinct from the training-only standardization of continuous input features.

Feature standardization is fitted on training data only. The causal reconstruction has no future-dependent fitted parameter and is identical in training, validation, testing, and serving.

A.2 PMF Support, Losses, and Metric Definitions

Let $b(y) = y$ for $y \le 128$ and $b(y) = \star$ otherwise. No held-out target fell in the overflow class; the class is retained to define a valid distribution under larger counts. For a mini-batch $\mathcal{B}$, the per-origin scores and equal-weight regional macro losses are

$$\ell^r_{\mathrm{nRPS},t} = \frac{1}{129}\sum_{k=0}^{128}\left(F^r_t(k) - \mathbb{I}[Y^r_t \le k]\right)^2, \qquad \ell^r_{\mathrm{NLL},t} = -\log p^r_{t,b(Y^r_t)},$$
$$\mathcal{L}_x = \frac{1}{2|\mathcal{B}|}\sum_{t\in\mathcal{B}}\sum_{r\in\{\mathrm{AP},\mathrm{AR}\}} \ell^r_{x,t}, \qquad x \in \{\mathrm{nRPS},\mathrm{NLL}\}. \tag{10}$$

NLL distinguishes mass assigned to the observed class, whereas nRPS also penalizes ordinal displacement through the CDF. Their positive weighted sum is a proper scoring objective for the modeled, possibly coarsened count classes (Gneiting & Raftery, 2007; Gneiting & Katzfuss, 2014). Equal AP/AR weighting does not make either regional score scale invariant.

For each mini-batch, let $\mathcal{A}_\mathcal{B}$ index every constructed relation-message gate over origins, layers, edges, and heads. The regularizer is

$$\mathcal{L}_{\mathrm{gate}} = \frac{1}{|\mathcal{A}_\mathcal{B}|}\sum_{a\in\mathcal{A}_\mathcal{B}} g_a. \tag{11}$$

It excludes the regional readout weights $w_{i,r}$ and is an activation regularizer, not a distributional scoring rule. Together with Equation 8, these definitions fully specify the training loss.

Let $\mathcal{Q} = \{.05, .10, .25, .50, .75, .90, .95\}$ and $\rho_\nu(u) = u(\nu - \mathbb{I}[u < 0])$ denote the pinball loss. The composite weighted interval score used for model selection and evaluation is

$$\mathrm{WIS}_r = \frac{1}{n}\sum_t \frac{2}{|\mathcal{Q}|}\sum_{\nu\in\mathcal{Q}} \rho_\nu(Y^r_t - Q^r_t(\nu)), \qquad \text{macro nWIS} = \frac{1}{2}\sum_{r\in\{\mathrm{AP},\mathrm{AR}\}} \frac{\mathrm{WIS}_r}{s_r}, \tag{12}$$

where s_r is the corresponding training-label mean of the evaluation airport. For zero-shot comparisons this target-derived quantity is used only as a common reporting scale and never enters training, model selection, or post-processing.

For central level $c \in \{.5, .8, .9\}$, define $\underline{y}^r_{t,c} = Q^r_t((1-c)/2)$ and $\overline{y}^r_{t,c} = Q^r_t(1-(1-c)/2)$. We report

$$\mathrm{Coverage}_{r,c} = \frac{1}{n}\sum_t \mathbb{I}[\underline{y}^r_{t,c} \le Y^r_t \le \overline{y}^r_{t,c}], \tag{13}$$

$$\mathrm{MPIW}_{r,c} = \frac{1}{n}\sum_t (\overline{y}^r_{t,c} - \underline{y}^r_{t,c}). \tag{14}$$

Coverage and width are descriptive diagnostics and must be interpreted together. No post-hoc calibration is applied to the reported distributions.

A.3 Edge Construction

Spatial attributes encode relative displacement and vertical separation; kinematic attributes encode closing speed, heading difference, and time and distance of closest approach (TCPA/DCPA). Intent attributes encode common region-oriented motion and entry-time agreement, order attributes encode directed predecessor–successor evidence, and task attributes encode signed boundary distance, motion trend, transition time, and regional membership. An aircraft pair may satisfy multiple relations,

Table 6: Operational specification of the fixed directed edge types. All quantities are derived from the current snapshot; dimensions refer to typed edge attributes.

Relation	Membership rule	Attr. dim.
Spatial	Horizontal distance within 30 NM or among the eight nearest aircraft; self-edges excluded.	5
Kinematic	Positive closing speed within 30 NM, or $0 < \mathrm{TCPA} < 10$ min and $\mathrm{DCPA} <$ 10 NM.	7
Intent	Common positive AP/AR approach tendency, or entry times within 5 min when both are below 30 min.	4
Order	Compatible headings ($\cos \Delta\psi > 0.8$) and directed transition order $0 \leq \tau_{i,r} < \tau_{j,r} < 30$ min for the same region r.	6
Task	Every aircraft has one outgoing edge to AP and one to AR; no reverse task edge.	7

but self-edges are excluded. Each aircraft has one outgoing edge to each task node, including while outside the corresponding region, and no reverse task edge is constructed.

In the order rule, $\tau_{i,r}$ is aircraft i's transition time toward region r as estimated from the current state, not an observed future arrival. Intent membership includes outside-to-AR approach. Transition-time features are capped at 30 minutes; speeds below 10^{-4} km/s are treated as invalid when deriving them, and edge attributes are clipped to the configured numerical bounds. Spatial nearest-neighbor construction is fixed rather than learned sparsification.

A.4 Additional Feature and Objective Ablations

Table 7 reports the secondary design ablations omitted from the core relation and task-structure table in the main paper.

Table 7: Additional feature and objective ablations (five-seed mean ± sample s.d.). AeroGraph is repeated as the reference, and all variants use the shared PMF head. Reported nRPS and nWIS values are multiplied by 10^3 and 10^2, respectively. Lower is better.

	Point Forecasting				Distribution Quality		
Model	$\mathrm{MAE}_{\mathrm{AP}}$	$\mathrm{RMSE}_{\mathrm{AP}}$	$\mathrm{MAE}_{\mathrm{AR}}$	$\mathrm{RMSE}_{\mathrm{AR}}$	nRPS	NLL	nWIS
(a) Airport A							
AeroGraph (Ours)	1.224 ± 0.005	1.746 ± 0.006	1.364 ± 0.024	1.889 ± 0.026	7.125 ± 0.050	1.807 ± 0.005	3.649 ± 0.016
$\lambda_{\mathrm{G}} = 0$	1.239 ± 0.015	1.763 ± 0.021	1.432 ± 0.030	1.979 ± 0.039	7.321 ± 0.078	1.833 ± 0.012	3.711 ± 0.036
unsigned boundary distance	1.239 ± 0.011	1.753 ± 0.016	1.369 ± 0.016	1.906 ± 0.023	7.200 ± 0.055	1.825 ± 0.007	3.686 ± 0.023
w/o exit-time + validity	1.246 ± 0.013	1.774 ± 0.019	1.380 ± 0.020	1.910 ± 0.027	7.189 ± 0.061	1.816 ± 0.008	3.688 ± 0.028
(b) Airport B							
AeroGraph (Ours)	1.302 ± 0.007	1.844 ± 0.009	1.376 ± 0.020	1.926 ± 0.027	7.380 ± 0.060	1.844 ± 0.008	4.597 ± 0.028
$\lambda_{\mathrm{G}} = 0$	1.325 ± 0.017	1.854 ± 0.019	1.422 ± 0.030	1.997 ± 0.042	7.575 ± 0.086	1.871 ± 0.012	4.688 ± 0.041
unsigned boundary distance	1.338 ± 0.012	1.862 ± 0.014	1.389 ± 0.016	1.949 ± 0.024	7.480 ± 0.061	1.863 ± 0.007	4.663 ± 0.027
w/o exit-time + validity	1.318 ± 0.014	1.857 ± 0.017	1.395 ± 0.020	1.952 ± 0.029	7.485 ± 0.067	1.857 ± 0.008	4.667 ± 0.032

A.5 Additional Transfer Comparisons

Table 8 reports the PMF comparators and native temporal heads omitted from the representative-family transfer table. Native point heads do not define distributional scores.

Table 8: Additional bidirectional zero-shot transfer comparisons omitted from the representative-family main table (five-seed mean ± sample s.d.). PMF point forecasts use P50; native heads use direct regression outputs, for which distributional scores are undefined. Reported nRPS and nWIS values are multiplied by 10^3 and 10^2, respectively.

	Point Forecasting				Distribution Quality		
Model	$\mathrm{MAE}_{\mathrm{AP}}$	$\mathrm{RMSE}_{\mathrm{AP}}$	$\mathrm{MAE}_{\mathrm{AR}}$	$\mathrm{RMSE}_{\mathrm{AR}}$	nRPS	NLL	nWIS
(a) Airport A → Airport B							
Native heads							
TimesNet	2.004 ± 0.043	2.658 ± 0.038	3.369 ± 0.051	4.475 ± 0.072	—	—	—
iTransformer	1.972 ± 0.032	2.629 ± 0.035	3.404 ± 0.110	4.519 ± 0.161	—	—	—
Additional PMF heads							
iTransformer	1.952 ± 0.025	2.668 ± 0.033	3.376 ± 0.035	4.510 ± 0.041	14.645 ± 0.122	2.490 ± 0.009	7.952 ± 0.064
$\mathrm{MLP}_{Current}$	2.295 ± 0.062	3.076 ± 0.090	3.458 ± 0.135	4.588 ± 0.179	15.862 ± 0.398	2.550 ± 0.019	9.029 ± 0.203
GIN	1.545 ± 0.016	2.159 ± 0.025	1.643 ± 0.056	2.257 ± 0.058	8.795 ± 0.116	2.007 ± 0.012	5.471 ± 0.027
GATv2	1.470 ± 0.072	2.052 ± 0.098	1.565 ± 0.077	2.140 ± 0.096	8.339 ± 0.412	1.941 ± 0.043	5.171 ± 0.263
(b) Airport B → Airport A							
Native heads							
TimesNet	1.864 ± 0.048	2.513 ± 0.073	3.158 ± 0.031	4.210 ± 0.037	—	—	—
iTransformer	1.835 ± 0.029	2.493 ± 0.067	3.190 ± 0.031	4.252 ± 0.032	—	—	—
Additional PMF heads							
iTransformer	1.976 ± 0.033	2.708 ± 0.039	3.316 ± 0.083	4.427 ± 0.095	24.103 ± 0.328	2.566 ± 0.041	7.526 ± 0.151
$\mathrm{MLP}_{Current}$	1.834 ± 0.042	2.519 ± 0.058	2.827 ± 0.067	3.817 ± 0.095	15.218 ± 0.289	2.382 ± 0.022	6.720 ± 0.130
GIN	1.435 ± 0.054	2.009 ± 0.028	1.654 ± 0.031	2.284 ± 0.034	8.982 ± 0.246	1.995 ± 0.011	5.044 ± 0.046
GATv2	1.323 ± 0.034	1.847 ± 0.032	1.475 ± 0.034	2.036 ± 0.047	8.168 ± 0.150	1.911 ± 0.017	4.648 ± 0.075

A.6 Complete Interval Results

The following tables report the raw regional coverage and MPIW omitted from the compact main-paper tables. Every entry is computed independently for each seed before aggregation. Width is not ranked without its corresponding coverage, and native point heads are excluded because they do not define predictive intervals.

A.6.1 Within-airport comparison

Table 9: Within-airport PMF results: AP interval coverage and mean prediction-interval width. Values are five-seed mean ± sample standard deviation. Coverage and width must be interpreted jointly; no post-hoc calibration is applied.

Model	50% interval		80% interval		90% interval	
	Coverage	MPIW	Coverage	MPIW	Coverage	MPIW
(a) Airport A						
TimesNet	0.636±0.015	2.992±0.112	0.860±0.011	5.704±0.200	0.933±0.008	7.356±0.247
iTransformer	0.637±0.010	2.926±0.119	0.863±0.011	5.620±0.241	0.933±0.009	7.356±0.370
$\mathrm{MLP}_{Current}$	0.680±0.014	3.504±0.150	0.886±0.012	6.737±0.317	0.944±0.009	8.758±0.477
MLP_{Hist}	0.646±0.018	2.558±0.084	0.858±0.019	4.911±0.165	0.924±0.016	6.346±0.215
DeepSets	0.705±0.007	2.178±0.065	0.893±0.006	4.250±0.128	0.952±0.005	5.577±0.192
Set Transformer	0.684±0.015	2.244±0.050	0.880±0.012	4.363±0.101	0.943±0.009	5.722±0.128
AeroGraph$_{homo}$	0.695±0.012	2.111±0.079	0.880±0.010	4.075±0.156	0.943±0.008	5.334±0.199
GIN	0.677±0.028	2.371±0.182	0.878±0.025	4.599±0.356	0.942±0.019	6.010±0.467
GATv2	0.693±0.009	2.203±0.067	0.884±0.007	4.267±0.137	0.945±0.005	5.580±0.188
R-GCN	0.695±0.009	2.134±0.100	0.884±0.008	4.128±0.193	0.944±0.006	5.403±0.255
AeroGraph (Ours)	0.682±0.004	1.914±0.043	0.869±0.006	3.689±0.093	0.931±0.002	4.810±0.112
(b) Airport B						
TimesNet	0.605±0.007	2.747±0.058	0.820±0.009	5.224±0.104	0.900±0.006	6.736±0.127
iTransformer	0.593±0.018	2.695±0.081	0.809±0.017	5.147±0.140	0.889±0.015	6.674±0.175
$\mathrm{MLP}_{Current}$	0.636±0.002	2.903±0.017	0.856±0.004	5.585±0.033	0.934±0.001	7.203±0.055
MLP_{Hist}	0.609±0.018	2.414±0.117	0.824±0.018	4.629±0.223	0.905±0.016	6.007±0.290
DeepSets	0.669±0.006	2.123±0.046	0.866±0.007	4.084±0.103	0.931±0.005	5.293±0.119
Set Transformer	0.659±0.004	1.982±0.038	0.854±0.004	3.806±0.076	0.922±0.002	4.944±0.105
AeroGraph$_{homo}$	0.652±0.008	1.916±0.058	0.846±0.009	3.659±0.108	0.914±0.007	4.739±0.147
GIN	0.629±0.011	1.943±0.085	0.822±0.009	3.716±0.205	0.897±0.010	4.817±0.160
GATv2	0.673±0.007	2.105±0.060	0.871±0.006	4.021±0.135	0.934±0.008	5.206±0.115
R-GCN	0.655±0.008	2.058±0.070	0.851±0.007	3.921±0.165	0.922±0.007	5.090±0.130
AeroGraph (Ours)	0.658±0.002	1.901±0.032	0.849±0.004	3.627±0.070	0.917±0.001	4.685±0.084

Table 10: Within-airport PMF results: AR interval coverage and mean prediction-interval width. Values are five-seed mean ± sample standard deviation. Coverage and width must be interpreted jointly; no post-hoc calibration is applied.

Model	50% interval		80% interval		90% interval	
	Coverage	MPIW	Coverage	MPIW	Coverage	MPIW
(a) Airport A						
TimesNet	0.571±0.011	5.106±0.125	0.827±0.013	9.796±0.260	0.914±0.010	12.732±0.361
iTransformer	0.541±0.019	4.997±0.168	0.798±0.020	9.610±0.327	0.892±0.018	12.497±0.443
MLP$_{Current}$	0.546±0.023	5.560±0.125	0.808±0.019	10.682±0.257	0.904±0.012	13.900±0.326
MLP$_{Hist}$	0.571±0.027	4.133±0.178	0.819±0.027	7.948±0.347	0.903±0.024	10.334±0.456
DeepSets	0.696±0.028	2.652±0.081	0.897±0.019	5.107±0.163	0.954±0.012	6.701±0.226
Set Transformer	0.701±0.011	2.648±0.094	0.906±0.006	5.110±0.168	0.961±0.004	6.718±0.224
AeroGraph$_{homo}$	0.696±0.013	2.372±0.122	0.899±0.009	4.576±0.224	0.956±0.006	5.997±0.305
GIN	0.699±0.024	2.939±0.216	0.907±0.019	5.645±0.408	0.962±0.012	7.382±0.538
GATv2	0.698±0.016	2.549±0.074	0.906±0.010	4.952±0.170	0.959±0.006	6.493±0.234
R-GCN	0.675±0.012	2.308±0.056	0.886±0.006	4.490±0.124	0.950±0.004	5.887±0.174
AeroGraph (Ours)	0.680±0.013	2.256±0.051	0.887±0.012	4.345±0.123	0.948±0.007	5.667±0.169
(b) Airport B						
TimesNet	0.538±0.011	4.667±0.122	0.796±0.012	8.970±0.229	0.892±0.009	11.674±0.300
iTransformer	0.515±0.025	4.633±0.270	0.771±0.029	8.903±0.518	0.871±0.027	11.578±0.683
MLP$_{Current}$	0.611±0.006	4.768±0.046	0.857±0.003	9.069±0.125	0.931±0.002	11.689±0.156
MLP$_{Hist}$	0.586±0.023	4.076±0.231	0.833±0.025	7.834±0.459	0.916±0.019	10.187±0.599
DeepSets	0.672±0.014	2.385±0.077	0.881±0.008	4.648±0.117	0.940±0.006	6.071±0.172
Set Transformer	0.664±0.007	2.219±0.048	0.874±0.008	4.313±0.100	0.936±0.006	5.610±0.129
AeroGraph$_{homo}$	0.652±0.007	2.110±0.014	0.860±0.008	4.114±0.038	0.926±0.005	5.349±0.049
GIN	0.633±0.015	2.303±0.120	0.841±0.014	4.460±0.345	0.912±0.014	5.812±0.220
GATv2	0.669±0.011	2.288±0.085	0.872±0.007	4.398±0.160	0.937±0.010	5.778±0.150
R-GCN	0.667±0.012	2.277±0.095	0.867±0.008	4.370±0.225	0.930±0.011	5.709±0.175
AeroGraph (Ours)	0.651±0.013	2.069±0.064	0.861±0.013	4.028±0.160	0.927±0.008	5.233±0.220

A.6.2 Two-airport ablations

Table 11: Ablation results: AP interval coverage and mean prediction-interval width. Values are five-seed mean ± sample standard deviation. Coverage and width must be interpreted jointly; no post-hoc calibration is applied.

Model	50% interval		80% interval		90% interval	
	Coverage	MPIW	Coverage	MPIW	Coverage	MPIW
(a) Airport A						
AeroGraph (Ours)	0.682±0.004	1.914±0.043	0.869±0.006	3.689±0.093	0.931±0.002	4.810±0.112
w/o spatial relation	0.702±0.008	2.068±0.075	0.890±0.007	3.994±0.134	0.947±0.005	5.206±0.181
w/o kinematic relation	0.671±0.010	1.897±0.070	0.861±0.010	3.663±0.145	0.925±0.006	4.777±0.185
w/o intent relation	0.706±0.009	2.133±0.060	0.898±0.009	4.145±0.125	0.956±0.006	5.374±0.165
w/o order relation	0.711±0.008	2.136±0.021	0.895±0.008	4.136±0.055	0.953±0.004	5.407±0.051
Task-only graph	0.705±0.012	2.109±0.047	0.889±0.011	4.094±0.103	0.951±0.008	5.387±0.146
w/o task edges	0.697±0.016	2.095±0.164	0.886±0.018	4.051±0.340	0.947±0.013	5.285±0.441
$\lambda_{\mathrm{G}} = 0$	0.704±0.012	2.063±0.085	0.887±0.012	3.990±0.175	0.948±0.008	5.197±0.230
unsigned boundary distance	0.699±0.008	2.125±0.055	0.890±0.008	4.094±0.115	0.950±0.005	5.330±0.150
w/o exit-time + validity	0.689±0.009	2.010±0.065	0.877±0.009	3.873±0.135	0.936±0.006	5.044±0.175
(b) Airport B						
AeroGraph (Ours)	0.658±0.002	1.901±0.032	0.849±0.004	3.627±0.070	0.917±0.001	4.685±0.084
w/o spatial relation	0.653±0.004	1.951±0.063	0.848±0.005	3.739±0.121	0.917±0.004	4.829±0.166
w/o kinematic relation	0.655±0.006	1.896±0.053	0.846±0.007	3.619±0.109	0.915±0.004	4.674±0.139
w/o intent relation	0.654±0.005	1.962±0.045	0.849±0.006	3.773±0.094	0.918±0.004	4.865±0.124
w/o order relation	0.653±0.005	1.963±0.016	0.848±0.005	3.770±0.063	0.918±0.003	4.876±0.038
Task-only graph	0.653±0.005	1.940±0.032	0.849±0.005	3.715±0.067	0.919±0.004	4.799±0.093
w/o task edges	0.656±0.008	1.952±0.106	0.850±0.009	3.745±0.215	0.921±0.007	4.837±0.277
$\lambda_{\mathrm{G}} = 0$	0.655±0.007	1.943±0.064	0.849±0.008	3.724±0.131	0.919±0.006	4.809±0.173
unsigned boundary distance	0.656±0.005	1.960±0.041	0.850±0.005	3.757±0.086	0.918±0.004	4.851±0.113
w/o exit-time + validity	0.657±0.005	1.928±0.049	0.849±0.006	3.686±0.101	0.918±0.004	4.760±0.131

Table 12: Ablation results: AR interval coverage and mean prediction-interval width. Values are five-seed mean ± sample standard deviation. Coverage and width must be interpreted jointly; no post-hoc calibration is applied.

Model	50% interval		80% interval		90% interval	
	Coverage	MPIW	Coverage	MPIW	Coverage	MPIW
(a) Airport A						
AEROGRAPH (Ours)	0.680 ± 0.013	2.256 ± 0.051	0.887 ± 0.012	4.345 ± 0.123	0.948 ± 0.007	5.667 ± 0.169
w/o spatial relation	0.698 ± 0.010	2.431 ± 0.066	0.901 ± 0.007	4.683 ± 0.119	0.956 ± 0.004	6.106 ± 0.157
w/o kinematic relation	0.687 ± 0.014	2.345 ± 0.090	0.891 ± 0.010	4.519 ± 0.175	0.950 ± 0.007	5.914 ± 0.220
w/o intent relation	0.727 ± 0.012	2.610 ± 0.080	0.916 ± 0.009	5.005 ± 0.150	0.965 ± 0.006	6.512 ± 0.190
w/o order relation	0.699 ± 0.013	2.478 ± 0.049	0.901 ± 0.008	4.784 ± 0.067	0.955 ± 0.005	6.242 ± 0.067
Task-only graph	0.706 ± 0.010	2.554 ± 0.071	0.909 ± 0.007	4.958 ± 0.123	0.961 ± 0.003	6.482 ± 0.145
w/o task edges	0.702 ± 0.021	2.476 ± 0.230	0.898 ± 0.017	4.722 ± 0.415	0.953 ± 0.012	6.155 ± 0.544
$\lambda_{\mathrm{G}} = 0$	0.695 ± 0.016	2.499 ± 0.115	0.895 ± 0.013	4.844 ± 0.220	0.952 ± 0.009	6.316 ± 0.285
unsigned boundary distance	0.687 ± 0.011	2.350 ± 0.075	0.895 ± 0.009	4.595 ± 0.145	0.952 ± 0.006	6.010 ± 0.185
w/o exit-time + validity	0.692 ± 0.012	2.343 ± 0.085	0.896 ± 0.010	4.505 ± 0.165	0.953 ± 0.007	5.889 ± 0.210
(b) Airport B						
AEROGRAPH (Ours)	0.651 ± 0.013	2.069 ± 0.064	0.861 ± 0.013	4.028 ± 0.160	0.927 ± 0.008	5.233 ± 0.220
w/o spatial relation	0.659 ± 0.010	2.183 ± 0.106	0.870 ± 0.009	4.239 ± 0.205	0.933 ± 0.006	5.493 ± 0.267
w/o kinematic relation	0.654 ± 0.014	2.127 ± 0.113	0.863 ± 0.011	4.137 ± 0.228	0.928 ± 0.008	5.379 ± 0.286
w/o intent relation	0.670 ± 0.012	2.299 ± 0.100	0.878 ± 0.009	4.440 ± 0.195	0.940 ± 0.007	5.732 ± 0.247
w/o order relation	0.659 ± 0.013	2.214 ± 0.061	0.869 ± 0.008	4.302 ± 0.087	0.932 ± 0.006	5.572 ± 0.087
Task-only graph	0.657 ± 0.008	2.188 ± 0.093	0.869 ± 0.006	4.272 ± 0.170	0.934 ± 0.003	5.551 ± 0.206
w/o task edges	0.668 ± 0.021	2.226 ± 0.253	0.874 ± 0.018	4.303 ± 0.469	0.936 ± 0.014	5.571 ± 0.617
$\lambda_{\mathrm{G}} = 0$	0.657 ± 0.016	2.227 ± 0.144	0.866 ± 0.014	4.340 ± 0.286	0.930 ± 0.010	5.616 ± 0.371
unsigned boundary distance	0.654 ± 0.011	2.130 ± 0.094	0.866 ± 0.009	4.184 ± 0.189	0.930 ± 0.007	5.435 ± 0.241
w/o exit-time + validity	0.656 ± 0.012	2.126 ± 0.106	0.866 ± 0.011	4.128 ± 0.215	0.931 ± 0.008	5.364 ± 0.273

A.6.3 BIDIRECTIONAL ZERO-SHOT TRANSFER

Table 13: Zero-shot transfer results: AP interval coverage and mean prediction-interval width. Values are five-seed mean ± sample standard deviation. Coverage and width must be interpreted jointly; no post-hoc calibration is applied.

Model	50% interval		80% interval		90% interval	
	Coverage	MPIW	Coverage	MPIW	Coverage	MPIW
(a) Airport A → Airport B						
TimesNet	0.630 ± 0.015	3.096 ± 0.104	0.854 ± 0.011	5.915 ± 0.183	0.929 ± 0.008	7.638 ± 0.231
iTransformer	0.619 ± 0.006	2.992 ± 0.117	0.844 ± 0.009	5.738 ± 0.217	0.919 ± 0.006	7.439 ± 0.297
MLP$_{Current}$	0.618 ± 0.019	3.855 ± 0.143	0.838 ± 0.017	7.438 ± 0.273	0.912 ± 0.012	9.776 ± 0.390
MLP$_{Hist}$	0.606 ± 0.010	2.579 ± 0.091	0.827 ± 0.013	4.945 ± 0.182	0.905 ± 0.011	6.396 ± 0.242
DeepSets	0.659 ± 0.013	2.438 ± 0.068	0.864 ± 0.011	4.757 ± 0.141	0.933 ± 0.008	6.257 ± 0.199
Set Transformer	0.633 ± 0.024	2.416 ± 0.058	0.845 ± 0.021	4.690 ± 0.112	0.919 ± 0.015	6.157 ± 0.143
AEROGRAPH$_{homo}$	0.658 ± 0.008	2.168 ± 0.079	0.858 ± 0.006	4.192 ± 0.158	0.927 ± 0.005	5.477 ± 0.206
GIN	0.648 ± 0.019	2.399 ± 0.180	0.854 ± 0.022	4.658 ± 0.352	0.925 ± 0.018	6.090 ± 0.471
GATv2	0.653 ± 0.019	2.285 ± 0.069	0.858 ± 0.014	4.418 ± 0.148	0.928 ± 0.009	5.780 ± 0.211
R-GCN	0.660 ± 0.009	2.232 ± 0.106	0.863 ± 0.008	4.322 ± 0.208	0.932 ± 0.006	5.657 ± 0.277
AEROGRAPH (Ours)	0.653 ± 0.006	2.016 ± 0.062	0.851 ± 0.006	3.890 ± 0.122	0.920 ± 0.004	5.059 ± 0.158
(b) Airport B → Airport A						
TimesNet	0.625 ± 0.009	2.693 ± 0.049	0.834 ± 0.007	5.139 ± 0.094	0.908 ± 0.005	6.653 ± 0.134
iTransformer	0.612 ± 0.019	2.703 ± 0.106	0.823 ± 0.018	5.141 ± 0.192	0.900 ± 0.013	6.682 ± 0.257
MLP$_{Current}$	0.655 ± 0.008	2.797 ± 0.024	0.877 ± 0.006	5.380 ± 0.037	0.944 ± 0.003	6.966 ± 0.055
MLP$_{Hist}$	0.637 ± 0.021	2.413 ± 0.099	0.846 ± 0.019	4.613 ± 0.192	0.921 ± 0.012	5.986 ± 0.259
DeepSets	0.702 ± 0.004	2.100 ± 0.049	0.888 ± 0.006	4.037 ± 0.111	0.947 ± 0.005	5.238 ± 0.123
Set Transformer	0.691 ± 0.005	1.949 ± 0.033	0.877 ± 0.005	3.741 ± 0.070	0.937 ± 0.004	4.862 ± 0.103
AEROGRAPH$_{homo}$	0.689 ± 0.007	1.900 ± 0.067	0.872 ± 0.008	3.630 ± 0.121	0.934 ± 0.007	4.707 ± 0.154
GIN	0.654 ± 0.007	1.922 ± 0.035	0.848 ± 0.004	3.685 ± 0.198	0.916 ± 0.008	4.774 ± 0.152
GATv2	0.713 ± 0.007	2.097 ± 0.054	0.893 ± 0.009	4.000 ± 0.210	0.950 ± 0.006	5.178 ± 0.147
R-GCN	0.687 ± 0.007	2.047 ± 0.067	0.876 ± 0.008	3.909 ± 0.173	0.936 ± 0.003	5.067 ± 0.213
AEROGRAPH (Ours)	0.695 ± 0.005	1.874 ± 0.054	0.876 ± 0.006	3.586 ± 0.095	0.935 ± 0.004	4.622 ± 0.125

Table 14: Zero-shot transfer results: AR interval coverage and mean prediction-interval width. Values are five-seed mean ± sample standard deviation. Coverage and width must be interpreted jointly; no post-hoc calibration is applied.

Model	50% interval		80% interval		90% interval	
	Coverage	MPIW	Coverage	MPIW	Coverage	MPIW
(a) Airport A → Airport B						
TimesNet	0.563±0.008	5.183±0.102	0.825±0.008	9.964±0.207	0.914±0.007	12.955±0.284
iTransformer	0.548±0.013	5.169±0.128	0.812±0.013	9.969±0.248	0.906±0.010	13.001±0.328
MLP$_{Current}$	0.588±0.020	5.989±0.076	0.855±0.013	11.508±0.173	0.937±0.008	15.040±0.218
MLP$_{Hist}$	0.559±0.026	4.141±0.204	0.807±0.029	7.968±0.399	0.897±0.023	10.361±0.519
DeepSets	0.650±0.034	2.793±0.081	0.865±0.026	5.372±0.171	0.936±0.017	7.053±0.233
Set Transformer	0.663±0.026	2.800±0.085	0.881±0.018	5.400±0.172	0.947±0.011	7.102±0.218
AeroGraph$_{homo}$	0.676±0.006	2.479±0.145	0.884±0.008	4.775±0.273	0.947±0.006	6.256±0.362
GIN	0.684±0.016	2.952±0.224	0.895±0.014	5.685±0.424	0.954±0.010	7.439±0.558
GATv2	0.672±0.018	2.694±0.086	0.886±0.012	5.212±0.191	0.949±0.008	6.836±0.274
R-GCN	0.647±0.020	2.412±0.064	0.865±0.013	4.671±0.137	0.936±0.009	6.127±0.195
AeroGraph (Ours)	0.663±0.008	2.348±0.073	0.872±0.011	4.522±0.175	0.937±0.008	5.894±0.238
(b) Airport B → Airport A						
TimesNet	0.545±0.008	4.631±0.107	0.800±0.009	8.911±0.210	0.895±0.008	11.592±0.281
iTransformer	0.534±0.027	4.594±0.233	0.784±0.030	8.818±0.468	0.878±0.024	11.463±0.613
MLP$_{Current}$	0.631±0.005	4.709±0.069	0.867±0.002	8.954±0.150	0.939±0.002	11.572±0.192
MLP$_{Hist}$	0.591±0.023	4.038±0.219	0.836±0.022	7.762±0.429	0.917±0.015	10.087±0.572
DeepSets	0.699±0.013	2.335±0.076	0.900±0.006	4.585±0.106	0.953±0.005	5.984±0.151
Set Transformer	0.686±0.012	2.161±0.050	0.887±0.010	4.238±0.101	0.945±0.007	5.504±0.132
AeroGraph$_{homo}$	0.676±0.006	2.038±0.023	0.878±0.007	4.018±0.038	0.939±0.004	5.219±0.055
GIN	0.644±0.012	2.293±0.164	0.850±0.009	4.442±0.315	0.922±0.007	5.785±0.465
GATv2	0.684±0.012	2.239±0.065	0.880±0.009	4.311±0.145	0.941±0.005	5.664±0.205
R-GCN	0.685±0.013	2.249±0.050	0.880±0.008	4.328±0.125	0.942±0.006	5.644±0.154
AeroGraph (Ours)	0.684±0.003	2.034±0.050	0.884±0.008	3.982±0.120	0.943±0.006	5.180±0.136

A.7 Optimization and Compute

For message passing, a relation with no incoming neighbor contributes zero; head outputs are concatenated before relation outputs are summed. The residual update preserves the previous destination state. Sigmoid gates modulate every constructed message but do not remove edges or reduce computation. With $M_t = \sum_\ell |\mathcal{E}_{t,\ell}|$ and fixed network width, graph storage is $\mathcal{O}(N_t + M_t)$ and edge-wise propagation is linear in LM_t; a naive pairwise graph-construction pass is $\mathcal{O}(N_t^2)$.

The aircraft MLP and graph encoder use width 128 and dropout. The graph has two propagation layers and four 32-dimensional attention heads. All methods in the shared-head comparison are trained from scratch; the earlier deterministic-backbone warm start is not used. Optimization uses Adam with $(\beta_1, \beta_2, \epsilon) = (.9, .999, 10^{-8})$, weight decay 10^{-5}, batches of 32 origins, and global gradient-norm clipping at 1.0. Learning-rate candidates and tuning budgets are matched across model families. Training is capped at 100 epochs. ReduceLROnPlateau uses factor .5 and patience 5, and early stopping uses patience 20. The scheduler, stopping rule, and saved PMF checkpoint are all driven by validation macro nWIS rather than test data.

The five prespecified seeds are used for every result in the main tables and interval appendix. Source-airport statistics are used for input standardization in zero-shot transfer; no target-airport labels influence training or checkpoint selection. Experiments ran on NVIDIA GeForce RTX 4090 GPUs, an AMD EPYC 7302 CPU, and 125 GiB RAM with Python 3.10.12 and PyTorch 2.6.0+cu124. Training data are shuffled, whereas validation and test origins are evaluated in fixed order.