# PG-MoE: Physically Grouped Mixture-of-Experts with Soft Track Generation for Multi-Horizon Typhoon Track Prediction

**Anonymous authors**
Paper under double-blind review

## Abstract

Accurate multi-horizon typhoon track forecasting is crucial for disaster prevention and mitigation, yet existing deep learning methods for ensemble forecast post-correction suffer from two key limitations: they discard probabilistic uncertainty from ensemble members, leading to error accumulation, and single networks lack capacity for diverse typhoon dynamics, while conventional MoE lacks physical priors. We propose PG-MoE, integrating soft-track probabilistic encoding and physically grouped mixture-of-experts. The soft-track module converts multi-source ensemble forecasts into standardized probabilistic representations via a downsampling-free fully convolutional network and confidence ellipse fitting, preserving uncertainty evolution. The MoE partitions experts into 23 homogeneous groups based on six physical attributes (e.g., lead time, intensity, moving distance) with a hierarchical sparse router and two-stage decoupled training. On a Northwest Pacific typhoon test dataset (2024–2025), PG-MoE significantly outperforms mainstream deep learning baselines and ECMWF deterministic forecasts within the critical 12–72H window. Ablations validate the core components. Our method offers a general probabilistic input pipeline and a physics-informed MoE paradigm, balancing accuracy and interpretability. Current framework supports MLP experts and track data only, with future extensions to multimodal observations and adaptive partitioning.

## 1 Introduction

Accurate multi-horizon typhoon track forecasting is fundamental to disaster mitigation and early warning in the Northwest Pacific. Operational ensembles, such as those from the ECMWF and the NCEP Global Ensemble Forecast System (GEFS), generate dozens of candidate tracks to quantify forecast uncertainty Vannitsem et al. (2020). Nevertheless, deep learning-based post-processing methods suffer from noticeable performance degradation within the critical 12–72H warning window. This degradation arises from two interconnected limitations. Fig. 1 illustrates this divergence using Typhoon 2405, along with the predictors employed for physical expert grouping, offering a concrete view of real-case forecast uncertainty.

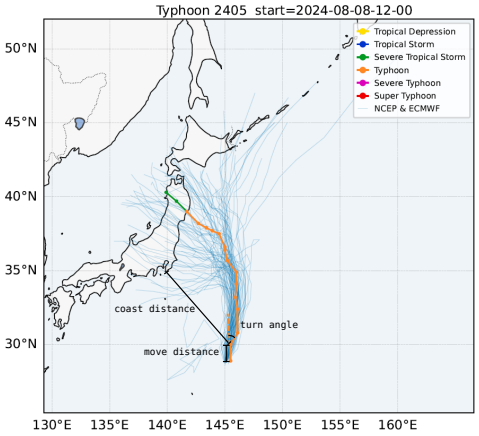



Figure 1: Illustration of ensemble divergences and some physical attributes for Typhoon 2405.

Traditional post-processing pipelines typically average or filter discrete ensemble members, thereby discarding anisotropic spatial uncertainty in forecast errors Rasp & Lerch (2018). In the absence of standardized probabilistic track representations, position errors accumulate rapidly with increasing lead time. Existing statistical uncertainty frameworks rely on



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.

rigid circular confidence bounds, which fail to capture directional track spread or asymmetric error distributions Luo et al. (2025); Yoo & Ho (2024). Meanwhile, typhoon evolution is governed by diverse atmospheric and topographic factors, resulting in highly heterogeneous dynamical regimes that a monolithic network cannot adequately model Liu et al. (2022). Critically, these two limitations are mutually reinforcing: inadequate spatial uncertainty modeling compounds the difficulty of fitting heterogeneous typhoon dynamics, ultimately restricting both predictive accuracy and interpretability.

Mixture-of-Experts (MoE) architectures in meteorology partially alleviate the shortcomings of single-network models, yet they remain hampered by flawed expert partitioning. Most existing designs derive expert groups from purely unsupervised clustering without incorporating cyclone physical priors. This practice yields redundant subnetworks and imbalanced allocation, where some experts are overactivated while others are underutilized. Consequently, expert specialization is not guaranteed, and the learned partitions may conflict with physical regimes, undermining interpretability and generalization Gan et al. (2025). Although ensemble averaging and unsupervised MoE serve as efficient operational baselines, their lack of physical constraints and fine-grained uncertainty representation limits further performance gains.

To address these bottlenecks, we propose a Physically Grouped Mixture-of-Experts (PG-MoE) framework, coupled with a Soft-Track probabilistic encoder for refined ensemble-based typhoon track correction. Soft-Track employs a fully convolutional network (FCN) without downsampling to generate 0.25° high-resolution spatial probability heatmaps from raw ensemble tracks. By fitting 75% bivariate normal confidence ellipses, we extract seven standardized anchor points, which serve as compact yet expressive uncertainty descriptors for downstream routing and correction. This design preserves full ensemble probabilistic information while mitigating numerical inconsistencies across forecast batches. Unlike fully unsupervised partitioning, our framework constructs 23 homogeneous expert groups using six physically observable cyclone attributes with overlapping buffer intervals, ensuring physically consistent task division. We further devise a two-layer sparse hierarchical router and a decoupled two-stage training strategy: experts are pre-trained and frozen, with only the routing parameters optimized during training. This explicit physically constrained training paradigm is absent in conventional end-to-end MoE pipelines Fedus et al. (2022).

Our contributions are twofold: (1) Soft-Track, a probabilistic post-processing pipeline for typhoon ensemble forecasts that provides fine-grained anisotropic uncertainty quantification; and (2) a novel MoE partitioning scheme based on physically observable attributes, which simultaneously balances prediction accuracy, numerical stability, and model interpretability.

## 2 Related Work

This work is closely related to three research directions: deep learning-based TC track forecasting, mixture-of-experts (MoE) models, and uncertainty modeling of ensemble forecasts.

**Deep Learning-Based TC Track Forecasting.** Deep learning methods for tropical cyclone track forecasting rely on historical tracks, satellite imagery, or reanalysis fields, generating forecasts via RNNs or spatiotemporal graph networks (Wang et al., 2023). Weather foundation models such as Pangu (Bi et al., 2022) and TC-specific foundation models (Liu et al., 2022) have advanced forecasting, yet remain centered on deterministic single-point trajectories, leaving ensemble post-processing and probabilistic information underexploited.

**Mixture-of-Experts for Meteorological Forecasting.** MoE enhances model capacity through dynamic routing and has been validated in large-scale models (Shazeer et al., 2017; Fedus et al., 2022). In meteorology, MoE has been applied to multi-scale forecasting and nowcasting (Gan et al., 2025). However, expert partitioning often relies on unsupervised clustering without physical priors, leading to weak interpretability and inadequate decoupling of heterogeneous typhoon dynamics.

**Uncertainty Modeling of Ensemble Forecasts.** Ensemble forecasting is standard in operational uncertainty quantification, supported by statistical post-processing methods (Gneiting & Raftery, 2007; Vannitsem et al., 2020). Deep learning frameworks commonly use ensemble means as deterministic inputs. Although some studies investigate multi-member temporal fusion (Luo et al., 2025), they use discrete member trajectories without explicit spatial probability encoding and face incon-

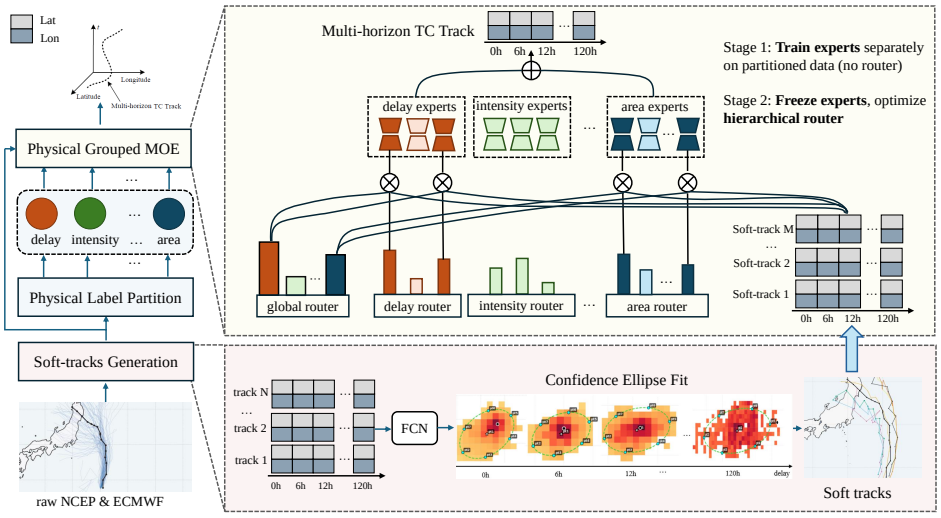



Figure 2: Overview of PG-MoE. It generates soft-tracks via confidence ellipse fitting, partitions tasks by physical attributes, and adopts a two-stage training strategy.

sistent member counts and alignment difficulties (Rasp & Lerch, 2018). A standardized probabilistic track representation for ensemble post-processing is still lacking.

In summary, existing research lacks an organic integration of probabilistic ensemble encoding, physics-guided MoE, and end-to-end track correction. This paper proposes structured uncertainty encoding via soft-track generation and decoupled dynamic modeling via physically grouped MoE to address these issues.

## 3 METHOD

We formulate the multi-lead-time typhoon track forecasting task as follows. Given input ensemble forecasts over lead times from 0h to 120h at 6h intervals, the model predicts the future track sequence $\hat{\mathcal{Y}}$ to minimize the discrepancy against the ground-truth $\mathcal{Y}$:

$$\min_{\theta} \mathcal{L}(\hat{\mathcal{Y}}, \mathcal{Y}),$$

where $\theta$ denotes all learnable parameters and $\mathcal{L}(\cdot)$ is the trajectory regression loss.

Fig. 2 illustrates our three-stage pipeline, which transforms raw ensemble forecasts into corrected trajectories. Soft-Track Generation first encodes ensemble uncertainty into standardized features. Physical Label Partition extracts physical priors to inform the router. Physical Grouped MoE then applies its experts to the Soft-Track features for trajectory correction, while its router selects and aggregates them using the extracted features. Detailed descriptions are provided below.

### 3.1 SOFT-TRACK GENERATION

To fully exploit the spatial uncertainty of ensemble forecasts and resolve member inconsistency, we design the Soft-Track Generation module (Fig. 2, lower right). It comprises two sequential components: heatmap generation via FCN and confidence ellipse fitting.

**Heatmap Generation via FCN.** We discretize the Northwest Pacific into uniform grids with a spatial resolution of $\delta$. A downsampling-free fully convolutional network (FCN) extracts multi-scale representations. Output scores are converted into a probability heatmap via global spatial Softmax. The network optimizes the lead-time-wise Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{B} \sum_{b=1}^{B} \sum_{u} Y_{b,u} \left( \log Y_{b,u} - \log(P_{b,u} + \epsilon) \right), \tag{1}$$

Table 1: Details of physical label partition

| Label | expert 1 | expert 2 | expert 3 | buffer | unit |
|---|---|---|---|---|---|
| delay | $[0, 24)$ | $[24, 72)$ | $[72, 120]$ | $\pm 6$ | $h$ |
| intensity | $(0, 24.5)$ | $[24.5, 41.4]$ | $(41.4, +\infty)$ | $\pm 3.5$ | $m/s$ |
| move_distance | $(0, 60)$ | $[60, 150]$ | $(150, +\infty)$ | $\pm 15$ | $km$ |
| turn_angle | $(0, 20)$ | $[20, 45]$ | $(45, 90)$ | $\pm 5$ | $^\circ$ |
| coast_distance | $(-\infty, 0)$ | $[0, 100]$ | $(100, +\infty)$ | $\pm 20$ | $km$ |

| area | expert 1 | expert 2 | expert 3 | expert 4 | expert 5 | expert 6 | expert 7 | expert 8 | lat: °N lon: °E |
|---|---|---|---|---|---|---|---|---|---|
| | $[105, 120]$ & $[0, 23]$ | $[120, 135]$ & $[5, 25]$ | $[135, 150]$ & $[5, 20]$ | $[150, 180]$ & $[0, 20]$ | $[120, 135]$ & $[25, 35]$ | $[135, 180]$ & $[20, 40]$ | $[100, 180]$ & $[0, 5]$ | else | |

**Confidence Ellipse Fitting.** High-dimensional heatmaps are unsuitable for direct input into downstream expert networks. We fit a confidence ellipse to extract $N_p + 1$ standardized key points: a weighted center and $N_p$ boundary points. The center $\boldsymbol{\mu}$ and covariance matrix $\Sigma$ are computed as:

$$\boldsymbol{\mu} = \mathbb{E}_P \left[ \begin{bmatrix} j \\ i \end{bmatrix} \right], \quad \Sigma = \mathbb{E}_P \left[ \left( \begin{bmatrix} j \\ i \end{bmatrix} - \boldsymbol{\mu} \right) \left( \begin{bmatrix} j \\ i \end{bmatrix} - \boldsymbol{\mu} \right)^T \right], \tag{2}$$

where $j$ and $i$ denote longitude and latitude grid indices. Eigen-decomposition of $\Sigma$ yields the axis lengths. We scale the axes by a $q$-th chi-squared quantile. To resolve orientation ambiguity, we align boundary points according to the azimuth between consecutive centers.

### 3.2 Physical Label Partition

To address heterogeneous typhoon dynamics and enforce expert specialization based on physical observable attributes, we design the Physical Label Partition module. It computes six physical attributes derived from Soft-Track features and the initial state: delay, intensity, move_distance, turn_angle, coast_distance, and area. Tab. 1 details the corresponding interval boundaries.

Delay reflects the shift from local to synoptic-scale controls (Liu & Tan, 2026; Zhang et al., 2025; Yuan et al., 2025). Intensity governs vortex–steering flow interactions (Wu et al., 2005). Move_distance captures steering strength (Chan & Gray, 1982). Turn_angle corresponds to circulation adjustments (Holland, 1984). Coast_distance isolates land friction and topographic blocking (Lin et al., 2002; Bender et al., 1987). Area accounts for regional oceanographic variability (Gray, 1977; Tong et al., 2024).

To prevent sharp discontinuities, we adopt overlapping buffers specified in Tab. 1. The five continuous attributes are each partitioned into three groups. Specifically, delay follows operational short-, medium-, and extended-range norms (Zhang et al., 2025; Yuan et al., 2025). Intensity adopts thresholds from the China national grade scale. Move_distance is defined based on climatological translation speeds (Kossin, 2018). Turn_angle is based on statistical deflection analyses (Hsu et al., 2018). Coast_distance is based on the onset scale of orographic effects (Lin et al., 2002; Bender et al., 1987) and area schemes (Gray, 1977). These definitions balance data distribution while preserving physical interpretability.

Within each physical expert group, samples are mutually exclusive except for boundary overlaps, whereas samples across different groups may overlap due to the multi-attribute nature of typhoons. These grouped experts are pre-trained independently in the first training stage (Sec. 3.3).

### 3.3 Physical Grouped Mixture-of-Experts

To enable expert specialization and dynamic routing, we design the PG-MoE module (Fig. 2, top right). The architecture contains an expert bank and a hierarchical routing network. We detail the core components below.

**Physically Partitioned Expert Bank.** We instantiate 23 experts based on the strategy in Sec. 3.2. Each expert is a 2-layer MLP. It maps the $2(N_p + 1)$-dimensional Soft-Track features to a 2D coordinate correction $\boldsymbol{O}_{e,g}$, where $e$ indexes the expert and $g$ indexes its physical group.

**Hierarchical Sparse Routing Network.** The routing network has two levels: a Global Router and six group-specific sub-routers. The Global Router processes the 7-dimensional physical vector. It evaluates the overall influence of all physical attributes to determine which physical regimes best match the current typhoon state. It selects the top-$K_g$ groups and normalizes the retained scores to obtain the sparse group weights $\tilde{w}_g$. Each sub-router exclusively takes its corresponding single physical attribute to preserve independent decision-making. The sub-router computes the routing scores for all experts within its group. It then selects the top-$K_e$ experts and normalizes the retained scores to obtain the intra-group weights $\tilde{w}_{e,g}$. The two-level design decouples global regime selection from local expert refinement.

The sparse selection mechanism follows a standard Top-K operation. We compute routing scores for all candidates. We retain only the top-K scores and set the rest to zero. The retained scores are then re-normalized via Softmax to produce the final weights $\tilde{w}_g$ and $\tilde{w}_{e,g}$.

The final prediction aggregates the selected experts:

$$\hat{\boldsymbol{y}} = \sum_{g=1}^{6} \tilde{w}_g \cdot \underbrace{\sum_{e \in \mathcal{E}_g} \tilde{w}_{e,g} \cdot \boldsymbol{O}_{e,g}}_{\boldsymbol{O}_g}, \tag{3}$$

where $\mathcal{E}_g$ is the set of activated experts in group $g$, and $\boldsymbol{O}_g$ is the aggregated output of that group.

**Training Objective.** We adopt a decoupled two-stage training strategy. In Stage 1, each expert is pre-trained on its designated subset with the Haversine loss. They are subsequently frozen to preserve physical specialization. In Stage 2, only the routing network is optimized. The total objective is:

$$\mathcal{L} = \mathcal{L}_{\text{haversine}} + \alpha_1 \mathcal{L}_{\text{global}} + \alpha_2 \mathcal{L}_{\text{group}}, \tag{4}$$

where $\mathcal{L}_{\text{haversine}}$ measures the great-circle distance:

$$\mathcal{L}_{\text{haversine}} = \frac{1}{B} \sum_{i=1}^{B} 2R \cdot \arcsin\left( \sqrt{\sin^2 \frac{\Delta\phi_i}{2} + \cos\hat{\phi}_i \cos\phi_i \sin^2 \frac{\Delta\lambda_i}{2}} \right), \tag{5}$$

with $\Delta\lambda_i$ and $\Delta\phi_i$ being the longitude and latitude differences, and $R = 6371\,\text{km}$. Auxiliary losses enforce load balancing. $\mathcal{L}_{\text{global}} = G \sum_{g=1}^{G} p_g^2$ encourages balanced group-level activation. $\mathcal{L}_{\text{group}} = \frac{1}{G} \sum_{g=1}^{G} E_g \sum_{e=1}^{E_g} p_{g,e}^2$ prevents over-activation of individual experts within a group. Here, $p_g$ and $p_{g,e}$ denote the normalized routing probabilities.

# 4 Experiments

## 4.1 Experimental Setup

We use typhoon data from the Northwest Pacific basin ($85.0°\text{E} - 200.0°\text{E}, 0.0°\text{N} - 65.0°\text{N}$), with ground-truth tracks from BABJ and ensemble inputs from ECMWF and NCEP GEFS. The domain is discretized into uniform grids with a spatial resolution of $\delta = 0.25°$. Training/validation sets cover 2015–2023 (7:3 split), and the test set covers 2024–2025. Raw ensembles are processed via the Soft-Track pipeline (Sec. 3.1) to fit a $q = 75\%$ confidence ellipse, from which we extract $N_p = 6$ boundary points and a weighted center, yielding $2(N_p + 1) = 14$-dimensional inputs. Six physical partition attributes are computed as in Sec. 3.2. For the PG-MoE module (Sec. 3.3) , we set the global router to select top-$K_g = 3$ groups and each sub-router to select top-$K_e = 2$ experts within its group, with auxiliary loss coefficients $\alpha_1 = 0.01$ and $\alpha_2 = 0.001$. All features are Z-score normalized using training set statistics. Model architecture, training hyperparameters, and baseline implementation details are provided in Appendix A. Appendix B provides additional experiments, including routing and training ablations, boundary expansion, intra-cluster expert count, and analyses of the grid resolution ($\delta$), chi-squared quantile ($q$), number of boundary sampling points ($N_p$), global routing Top-$K_g$, and intra-cluster routing Top-$K_e$.

We compare against operational forecasts (ECMWF, NCEP) and sequential baselines (BiLSTM, ConvLSTM, TMPI Luo et al. (2025)). Specifically, ECMWF_A and NCEP_A denote the ensemble

Table 2: $\Delta R$ (km) on 2024 and 2025 test sets.

| Horizon(2024) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
|---|---|---|---|---|---|---|---|---|---|---|
| BiLSTM | 53.62 | 53.08 | 66.81 | 70.6 | 85.98 | 111.79 | 133.55 | 176.0 | 243.55 | 452.81 |
| ConvLSTM | 50.12 | 47.86 | 58.79 | 65.67 | 83.45 | 107.78 | 131.64 | 168.8 | 253.38 | 465.34 |
| TMPI | 47.39 | 45.84 | 58.67 | 65.24 | 79.36 | 105.18 | 129.41 | 165.77 | 242.21 | 443.57 |
| ECMWF_A | 75.65 | 100.22 | 106.04 | 115.06 | 137.38 | 153.02 | 172.78 | 201.36 | 232.0 | 349.0 |
| ECMWF_D | 45.51 | 46.18 | 51.76 | 57.76 | 74.34 | 91.59 | 114.73 | 140.07 | 222.59 | 355.97 |
| NCEP_A | 49.69 | 55.21 | 60.15 | 70.88 | 90.61 | 113.65 | 138.29 | 170.96 | 243.95 | 360.1 |
| NCEP_D | 52.18 | 54.12 | 60.22 | 71.69 | 93.53 | 111.53 | 135.59 | 164.93 | 227.72 | **314.52** |
| **PG-MOE (Ours)** | **42.96** | **43.28** | **49.59** | **56.3** | **70.58** | **86.07** | **108.63** | **138.23** | **218.41** | 348.11 |
| Horizon(2025) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
| BiLSTM | 55.65 | 58.33 | 72.27 | 70.82 | 82.46 | 104.29 | 137.48 | 176.52 | 284.82 | 413.76 |
| ConvLSTM | 51.93 | 51.66 | 61.86 | 68.36 | 78.28 | 100.7 | 131.78 | 170.85 | 295.14 | 411.39 |
| TMPI | 50.86 | 49.9 | 61.19 | 66.41 | 79.79 | 97.8 | 127.07 | 166.11 | 289.58 | 398.15 |
| ECMWF_A | 98.28 | 103.21 | 112.41 | 118.73 | 138.27 | 166.78 | 210.0 | 232.05 | 314.87 | 406.56 |
| ECMWF_D | 42.22 | 46.9 | 51.01 | 58.36 | 73.34 | 90.83 | 113.4 | **140.4** | 258.25 | 462.64 |
| NCEP_A | 42.44 | 47.0 | 56.79 | 64.54 | 87.01 | 116.71 | 155.25 | 187.85 | 289.73 | 545.44 |
| NCEP_D | 46.25 | 52.06 | 66.82 | 69.89 | 88.94 | 118.55 | 159.89 | 190.55 | 341.58 | 489.04 |
| **PG-MOE (Ours)** | **40.24** | **44.28** | **48.45** | **55.44** | **69.57** | **86.79** | **111.41** | 141.13 | **230.34** | **388.65** |

Best in bold, second-best underlined. This notation applies to all tables below.

mean forecasts from ECMWF and NCEP, respectively, while ECMWF_D and NCEP_D denote the deterministic forecasts from ECMWF and NCEP, respectively. For ablations, we include a single-expert MLP (Single-E) and a dense MoE (Dense-MoE), and further conduct a leave-one-out analysis of the six physical expert clusters, assessing both their contribution and activation frequency. The primary metric is the Average Track Error (ATE), denoted by $\Delta R$ in tables, computed via the Haversine distance and reported in km as the mean over three independent runs.

### 4.2 Main Results

We compare PG-MoE against sequential deep learning baselines and operational forecasts (ECMWF, NCEP). ATE are reported in Tab. 2. Error increases with lead time across all models.

PG-MoE consistently achieves the lowest track errors across the vast majority of forecast lead times from 6H to 120H on both test sets, with only two instances of second-best results (2024-120H and 2025-72H). Overall, PG-MoE exhibits stable and competitive accuracy over the operationally relevant 6H–120H window, and its improvements are particularly pronounced in the short-to-medium range. These results validate the effectiveness of soft-track probabilistic encoding and physically grouped MoE in enhancing forecast accuracy and robustness.

### 4.3 Ablation Study on the SoftTrack Probabilistic Prior Module

To verify the SoftTrack component, we compare three variants: NoSoftTrack (raw discrete trajectories), SoftTrack-probability (top-7 probability points directly from the heatmap without ellipse fitting), and PG-MoE (full SoftTrack). As shown in Tab. 3, NoSoftTrack consistently underperforms both probabilistic variants at all lead times from 12H to 120H, with the gap widening notably as the forecast horizon increases. This confirms that global probability distributions effectively constrain trajectory drift over extended ranges.

Table 3: $\Delta R$ (km) for SoftTrack ablation on 2024 and 2025 test sets. Results for additional lead times are provided in Tab. 13.

| | 2024 | | | | | 2025 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Horizon | 12H | 24H | 48H | 72H | 120H | 12H | 24H | 48H | 72H | 120H |
| NoSoftTrack | 47.02 | 60.37 | 105.23 | 184.75 | 508.03 | 48.01 | 64.14 | 114.21 | 200.03 | 444.78 |
| SoftTrack-probability | 45.12 | 56.69 | 88.50 | 139.12 | **345.87** | 45.11 | 57.13 | 91.84 | 148.44 | 410.70 |
| PG-MoE | **43.28** | **56.30** | **86.07** | **138.23** | 348.11 | **44.28** | **55.44** | **86.79** | **141.13** | **388.65** |

Table 4: $\Delta R$ (km) for sparse MoE architecture ablation on the 2024–2025 test sets. Results for additional lead times are provided in Tab. 14.

| | 2024 | | | | | 2025 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Horizon | 12H | 24H | 48H | 72H | 120H | 12H | 24H | 48H | 72H | 120H |
| Single-E | 64.38 | 73.55 | 96.59 | 141.14 | **332.98** | 70.95 | 78.12 | 99.38 | 140.72 | **382.48** |
| Dense-MoE | 43.56 | 56.39 | 86.59 | **135.60** | 339.27 | **43.77** | 55.54 | 88.23 | **140.62** | 391.84 |
| PG-MoE | **43.28** | **56.30** | **86.07** | 138.23 | 348.11 | 44.28 | **55.44** | **86.79** | 141.13 | 388.65 |

The key advantage lies in data retention: NoSoftTrack discards members with mismatched timestamps, causing substantial sample loss, whereas probabilistic variants preserve full ensemble uncertainty. SoftTrack-probability and PG-MoE achieve comparable errors, indicating that the heatmap itself captures essential information; the additional ellipse fitting further extracts stable, low-dimensional, and spatially aligned features, enabling PG-MoE to achieve a slightly better overall balance. Overall, SoftTrack robustly reduces track errors in the critical 24H–120H window, making it an essential module for medium-to-long-range typhoon forecasting.

## 4.4 PG-MOE Ablation

### 4.4.1 MoE Architecture Ablation

To validate the effectiveness of the sparse MoE architecture, we compare PG-MoE against Single-E (a single-expert MLP with the same activation compute as PG-MoE) and Dense-MoE (Tab. 4). Multi-expert models significantly outperform Single-E within the core 12H–72H window, demonstrating that a single network structure lacks the specialized modeling capacity for heterogeneous typhoon dynamics. At the 120H ultra-long lead time, the gaps among models narrow considerably, with Single-E unexpectedly achieving the lowest errors in both years, likely due to reduced overfitting under high uncertainty. Dense-MoE achieves competitive performance at some short-to-medium lead times, occasionally matching or slightly surpassing PG-MoE (e.g., 72H in 2024), but exhibits larger fluctuations across horizons. In contrast, PG-MoE's physics-aware hierarchical sparse routing activates only relevant experts, yielding the most consistent performance across the operationally critical 12H–72H window.

### 4.4.2 Single Physical Expert Cluster Removal

To quantify the contribution of each physical grouping, we conduct a leave-one-out ablation study. PG-MoE constructs six expert clusters based on lead time, translation distance, turning amplitude, intensity, coastal distance, and activity basin. In each variant, one cluster is disabled during both training and inference, with the full PG-MoE as the baseline (Tab. 5). Removing a cluster increases track error in many evaluated settings, although its effect depends on the year and forecast horizon. At 12H–24H, removal degrades performance in most cases, whereas at 72H–120H, some variants achieve errors comparable to or slightly below the full model (e.g., w/o-area-group at 120H in 2024). This suggests that the contributions of some clusters overlap or become less beneficial at longer lead times. Among individual clusters, the time-group is particularly important for short-term forecasting: its removal yields the largest error increase at 12H in both years. The other clusters have less consistent individual effects and are more substitutable, especially at longer lead times. Overall, combining all six physical priors provides a balance of short-term accuracy and long-term robustness.

Table 5: $\Delta R$ (km) for single-cluster removal ablations on 2024 and 2025 test sets. Results for additional lead times are provided in Tab. 15.

| | 2024 | | | | | 2025 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Horizon | 12H | 24H | 48H | 72H | 120H | 12H | 24H | 48H | 72H | 120H |
| w/o-time-group | 44.45 | 56.82 | 86.63 | **137.07** | 355.70 | 45.03 | 56.29 | 88.26 | **137.63** | **376.78** |
| w/o-move-group | 43.64 | 56.48 | 86.92 | 137.83 | 347.40 | 44.02 | 55.82 | 89.46 | 140.63 | 389.40 |
| w/o-turn-group | 43.56 | 56.80 | 87.01 | 138.69 | 355.56 | **43.87** | 55.93 | 87.45 | 140.49 | 387.19 |
| w/o-intensity-group | 43.85 | 56.63 | 87.63 | 137.40 | 348.21 | 44.05 | **55.39** | 89.19 | 142.54 | 386.42 |
| w/o-coast-group | 43.38 | **56.20** | 87.18 | 137.51 | 348.56 | 44.11 | 55.73 | 89.55 | 140.73 | 383.68 |
| w/o-area-group | 43.70 | 56.21 | 87.29 | 137.88 | **341.33** | 43.97 | 56.02 | 90.30 | 142.42 | 389.11 |
| PG-MoE | **43.28** | 56.30 | **86.07** | 138.23 | 348.11 | 44.28 | 55.44 | **86.79** | 141.13 | 388.65 |

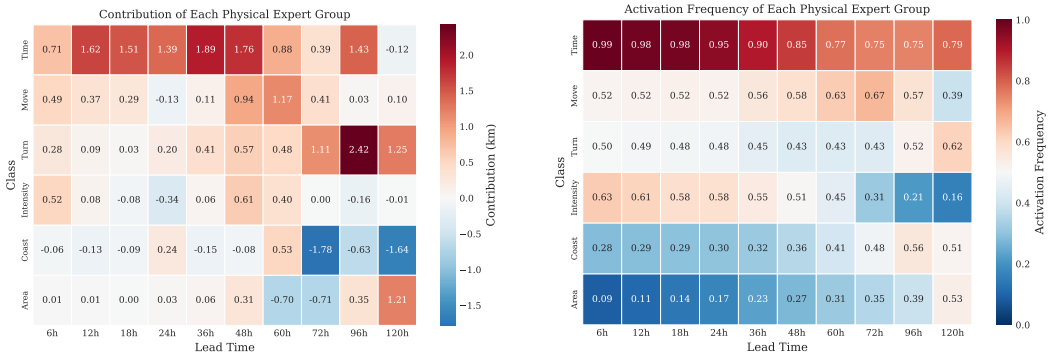


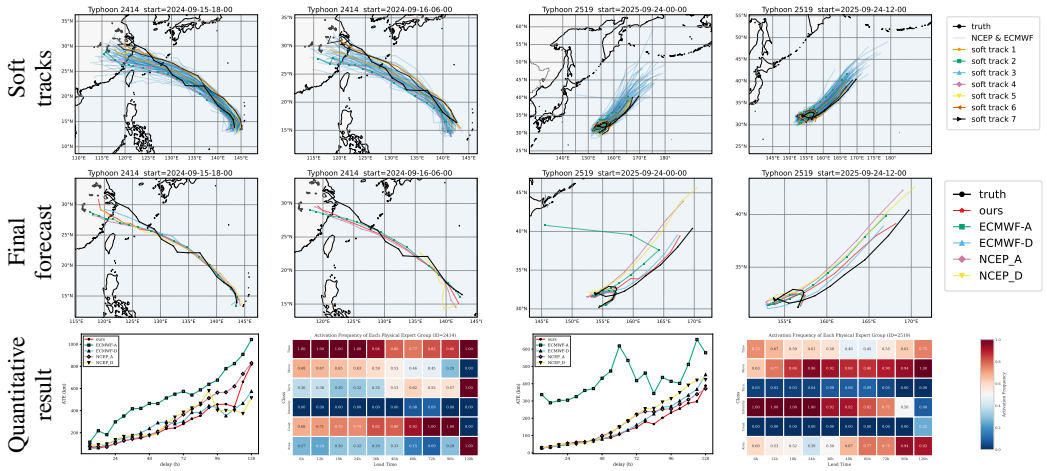

Figure 3: Contribution (left) and activation frequency (right) of each physical expert group across lead times, averaged over all 2024–2025 test typhoons. Contribution is defined as $C_{g,h} = \Delta R_{\text{w/o } g,h} - \Delta R_{\text{full},h}$. In the left heatmap, red/blue denote positive/negative contribution; in the right heatmap, darker red indicates higher activation frequency.

Fig. 3 shows the contribution and activation frequency of each physical expert group across lead times, averaged over all typhoon samples in the 2024 and 2025 test sets. The two panels exhibit a clear positive correspondence: groups with larger contributions tend to be activated more frequently, and vice versa. This consistency indicates that the hierarchical sparse router allocates more routing weight to physical groups that are more informative for the current lead time. It also explains the leave-one-out results: removing a highly contributed and frequently activated group (e.g., the time group at short lead times) causes a clear error increase, whereas removing a low-contribution, low-activation group has a weaker effect and may even be compensated by other groups at long lead times.

### 4.5 Case Study on Representative Typhoons

To verify the generalization and physical interpretability of PG-MoE, we select two representative typhoons with divergent evolutionary characteristics: landfalling and recurving Typhoon 2414 (PULASAN) and long-lived open-ocean Typhoon 2519 (NEOGURI). Fig. 4 illustrates the ensemble forecast tracks and Soft-Track extracted trajectories, track comparisons, ATE curves, and expert activation frequencies.

For Typhoon 2414, the cyclone moves northwestward, makes landfall near Zhejiang–Shanghai, and then recurves northeastward. Soft-Track effectively preserves the ensemble uncertainty near the coast. During the landfall–recurvature stage, PG-MoE closely follows the ground truth within the 12–72H window, outperforming operational baselines. Activation statistics show that the lead-time and coast-distance experts remain highly activated for most lead times, whereas the move-distance expert decays rapidly after the early period. The intensity expert is almost inactive, consistent with the cyclone remaining a tropical storm with little intensity variation. In contrast, the turn-angle and area experts become more active later, in line with its recurvature.




(a) 2414 PULASAN (b) 2519 NEOGURI

Figure 4: Case visualization for Typhoon 2414 (a) and Typhoon 2519 (b). Rows correspond to Soft tracks, Final forecast, Quantitative result respectively.

For Typhoon 2519, the cyclone moves westward, executes a circulation gyration, and then sharply turns northeast. Soft-Track effectively mitigates trajectory drift at medium-to-long lead times. During the gyration, PG-MoE exhibits slower error growth than operational forecasts across the 72–108H window, with ATE remaining below 300 km. Activation statistics show that the move-distance and area clusters remain highly activated, while the intensity cluster dominates short lead times but decays at longer horizons. In contrast, the coast-distance cluster is rarely activated, consistent with its long open-ocean lifetime. The turn-angle cluster also remains largely inactive, leading to insufficient feature capture for large-scale looping scenarios, which likely contributes to its error growth.

In summary, these cases validate Soft-Track's capability in encoding ensemble uncertainty and the physical grouping MoE with adaptive routing. Within the 12–72H window, PG-MoE consistently outperforms ECMWF and NCEP operational forecasts, demonstrating robust generalization for multi-horizon typhoon track prediction in the Northwest Pacific.

## 5 Limitations and Future Work

Despite its competitive performance, our framework has several limitations. First, the Soft-Track representation is tailored to MLP experts, which limits compatibility with sequential models. Second, fixed grouping thresholds derived from meteorological priors may not adapt optimally to extreme events, such as sharp turns or landfalls. Third, evaluations are restricted to the Northwest Pacific basin. Finally, the model relies solely on ensemble track data, excluding multi-modal inputs like satellite imagery or reanalysis fields. Future work will extend the framework to multiple basins and support joint prediction of intensity and wind-rain hazards.

## 6 Conclusion

We propose PG-MoE, which integrates Soft-Track probabilistic encoding with physically grouped mixture-of-experts. It addresses insufficient ensemble uncertainty utilization and the limited capacity of single networks for heterogeneous typhoon dynamics. On the Northwest Pacific test set, PG-MoE outperforms deep learning baselines and ECMWF deterministic forecasts within the critical 12–72H window. Extensive ablation studies verify the necessity of each core component. Our work provides a generic probabilistic pipeline for ensemble post-processing. The physics-guided MoE paradigm also offers a reference for fine-grained forecasting of other meteorological variables.

### AI USE STATEMENT

We used generative AI tools to assist with drafting and polishing the manuscript, retrieving and identifying relevant literature, and discussing research ideas, methodological choices, and experimental design. The authors reviewed AI-assisted suggestions and take responsibility for the citations, methods, experimental results, and final content of this paper.

### ETHICS STATEMENT

This study uses a dataset assembled by the authors from meteorological forecasts and tropical cyclone tracks. It does not involve human subjects or personal information. Because track prediction can affect real-world decisions, our results are intended for research evaluation and should not be used as the sole basis for operational warnings.

## REFERENCES

Morris A Bender, Robert E Tuleya, and Yoshio Kurihara. A numerical study of the effect of island terrain on tropical cyclones. *Monthly Weather Review*, 115(1):130–155, 1987.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*, 2022.

Johnny CL Chan and William M Gray. Tropical cyclone movement and surrounding flow relationships. *Monthly Weather Review*, 110(10):1354–1374, 1982.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Lihao Gan, Xin Man, Chenghong Zhang, and Jie Shao. Ewmoe: An effective model for global weather forecasting with mixture-of-experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 210–218, 2025.

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

William M Gray. Tropical cyclone genesis in the western north pacific. *Journal of the Meteorological Society of Japan. Ser. II*, 55(5):465–482, 1977.

Greg J Holland. Tropical cyclone motion. a comparison of theory and observation. *Journal of Atmospheric Sciences*, 41(1):68–75, 1984.

Li-Huan Hsu, Shih-Hao Su, Robert G Fovell, and Hung-Chi Kuo. On typhoon track deflections near the east coast of taiwan. *Monthly Weather Review*, 146(5):1495–1510, 2018.

James P Kossin. A global slowdown of tropical-cyclone translation speed. *Nature*, 558(7708): 104–107, 2018.

Yuh-Lang Lin, Darrell B Ensley, Sen Chiao, and Ching-Yuang Huang. Orographic influences on rainfall and track deflection associated with the passage of a tropical cyclone. *Monthly weather review*, 130(12):2929–2950, 2002.

Hao-Yan Liu and Zhe-Min Tan. On the forecast errors of tropical cyclone intensity in hwrf over the western north pacific. *Geophysical Research Letters*, 53(1):e2025GL117542, 2026.

Zili Liu, Kun Hao, Xiaoyi Geng, Zhengxia Zou, and Zhenwei Shi. Dual-branched spatio-temporal fusion network for multihorizon tropical cyclone track forecast. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:3842–3852, 2022.

Lei Luo, Qifeng Qian, Jiahao Luan, Liming Mao, Xing Liao, Xuanzhi Chen, Yajing Xu, and Jun Guo. Time-member progressive inference for long-term tropical cyclone track forecasting. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

Stephan Rasp and Sebastian Lerch. Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11):3885–3900, 2018.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Haoyu Tong, Tim Li, and Xiao Pan. Meridional shift of climatological tropical cyclone genesis location in the western north pacific. *Atmospheric Science Letters*, 25(11):e1263, 2024.

Stéphane Vannitsem, John Bjørnar Bremnes, Jonathan Demaeyer, Gavin R Evans, Jonathan Flowerdew, Stephan Hemri, Sebastian Lerch, Nigel Roberts, Susanne Theis, Aitor Atencia, et al. Statistical postprocessing for weather forecasts–review, challenges and avenues in a big data world. *arXiv preprint arXiv:2004.06582*, 2020.

Liang Wang, Bingcheng Wan, Shaohui Zhou, Haofei Sun, and Zhiqiu Gao. Forecasting tropical cyclone tracks in the northwestern pacific based on a deep-learning model. *Geoscientific Model Development*, 16(8):2167–2179, 2023.

Liguang Wu, Bin Wang, and Scott A Braun. Impacts of air–sea interaction on tropical cyclone track and intensity. *Monthly weather review*, 133(11):3299–3314, 2005.

Seungwoo Yoo and Chang-Hoi Ho. Development of interpretable probability ellipse in tropical cyclone track forecasts using multiple operational ensemble prediction systems. *Journal of Geophysical Research: Atmospheres*, 129(4):e2023JD039295, 2024.

Huanmujin Yuan, Hong Wang, Yubin Li, Kevin KW Cheung, and Zhiqiu Gao. Forecast errors of tropical cyclone track and intensity by the china meteorological administration from 2013 to 2022. *Atmospheric and Oceanic Science Letters*, pp. 100675, 2025.

Xiping Zhang, Guomin Chen, Zifeng Yu, and Xiangxiang Zhang. A study of tropical cyclone strike probability forecasts in three ensemble prediction systems. *Weather and Forecasting*, 40(12): 2621–2635, 2025.

# A Detailed Experimental Setup

**Dataset and Preprocessing (Full Details).** We conduct experiments using typhoon data from the Northwest Pacific basin ($85.0°\text{E} \sim 200.0°\text{E}$, $0.0°\text{N} \sim 65.0°\text{N}$). The ground-truth tracks are obtained from the BABJ best-track dataset, and ensemble forecast inputs are merged from ECMWF and NCEP GEFS. Due to data limitations, NCEP ensemble records are missing for a portion of samples.

The study domain is discretized onto a regular $0.25°$ latitude–longitude grid with spatial dimensions $H = 260, W = 460$. We adopt a strict time-extrapolation protocol for dataset splitting. The training and validation sets include all typhoon samples from 2015 to 2023, split at a ratio of $7 : 3$. The test set comprises typhoon samples occurring in 2024 and 2025. Samples sharing the same typhoon ID and initial forecast time are grouped together. A group is discarded if ground-truth positions are missing at required lead times; if ground truth is unavailable for later lead times within a group, the trajectory sequence is truncated accordingly.

The model takes ensemble forecasts at the initial time as input and outputs optimized typhoon center tracks for lead times ranging from 0 h to 120 h with a 6 h interval. Raw ensemble forecasts are preprocessed following the Soft-Track pipeline described in Sec. 3.1, with the following parameters: the probability threshold is 0.001, the chi-squared quantile for 75% confidence is 2.773, and six boundary points are sampled at angles $-60°, -90°, -120°, 120°, 90°, 60°$, which are combined with the probability-weighted mean center to form a set of seven coordinate features. Subsequently, the six physical partitioning features are computed following the definitions in Sec. 3.2, with overlapping buffers applied to continuous features, while the discrete area zones have no buffers.

The seven coordinate features fed into the MoE module are standardized via Z-score normalization. Mean and variance statistics are calculated solely on the 2015–2023 training set to avoid temporal

leakage. Let $i$ and $j$ denote the grid indices for latitude and longitude, respectively. The predicted geographic coordinates $(\hat{\phi}, \hat{\lambda})$ are computed as:

$$\hat{\phi} = \phi_{\min} + i \cdot \Delta\phi + 0.125,$$
$$\hat{\lambda} = \lambda_{\min} + j \cdot \Delta\lambda + 0.125,$$

where $\phi_{\min}$ and $\lambda_{\min}$ are the lower bounds of the grid, and $\Delta\phi = \Delta\lambda = 0.25°$. Coastline distance computation uses the high-resolution GSHHS coastline dataset `GSHHS_shp/h/GSHHS_h_L1.shp`.

**Model Architecture Hyperparameters (Full Details).** The downsampling-free FCN consists of 4 feature extraction convolutional blocks followed by 1 output convolution layer, comprising a total of 5 `Conv2d` layers with $3 \times 3$ kernels, stride 1, and padding 1, without any pooling layers, and a fixed spatial resolution of $260 \times 460$. The channel sequence is configured as $1 \to 32 \to 64 \to 128 \to 64 \to 1$. `Dropout2d(0.1)` is applied only after the first three convolutional blocks, and convolution weights adopt PyTorch's default Kaiming initialization. For the Physical Grouped Mixture-of-Experts (PG-MoE), each expert is implemented as a 2-layer MLP with an input dimension of 14, a hidden dimension of 64, outputting 2D longitude–latitude corrections, with ReLU activation.

The hierarchical gating network is configured as follows: the global group gate follows a $7 \to 32 \to 32 \to 6$ structure, selecting the top-3 groups with renormalized weights. For the group-specific sub-gates, continuous features use a $1 \to 16 \to 3$ structure, while for the discrete area ID group, the maximum area ID is max_area_id $= 7$, with an embedding dimension emb_dim $= 8$, followed by an MLP $8 \to 16 \to 8$. Within each group, the top-2 experts are selected and renormalized.

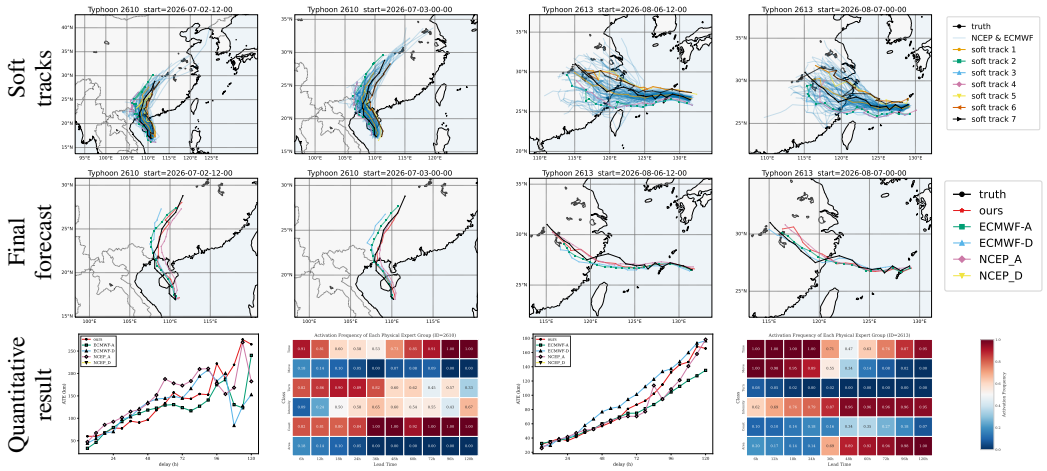
The overall loss function for gate training is formulated as $\mathcal{L} = \mathcal{L}_{\text{haversine}} + \alpha_1 \mathcal{L}_{\text{global}} + \alpha_2 \mathcal{L}_{\text{group}}$, where $\alpha_1 = 0.01$ and $\alpha_2 = 0.001$. Stage 1 expert pre-training solely uses $\mathcal{L}_{\text{haversine}}$, whereas in Stage 2, the expert weights are fully frozen and all three loss terms are jointly optimized. The Haversine spherical distance computation adopts an Earth radius of 6371 km.

**Training Optimization Details.** For the Soft-Track FCN, we set the learning rate to $10^{-4}$, batch size to 8, and maximum epoch to 16, with an early stopping patience of 10 monitored on the validation KL loss. For MoE expert pre-training, the learning rate is $10^{-3}$, batch size is 32, and maximum epoch is 32, without early stopping. For MoE gate training, the learning rate is $10^{-3}$, batch size is 64, and maximum epoch is 32, also without early stopping. No learning rate scheduler or gradient clipping is used. We employ the Adam optimizer with PyTorch default hyperparameters and zero weight decay.

All quantitative experiments are repeated three times with fixed random seeds $\{42, 1234, 911\}$, and results are reported as mean. Randomness of PyTorch, NumPy, and CUDA is fully locked, and `cudnn.benchmark` is disabled.

Experiments are conducted on a single RTX 4090 (24 GB) GPU under Linux, CUDA 12.4, and driver version 550.78. The implementation is based on PyTorch with auxiliary libraries including NumPy, Pandas, and scikit-learn. Training uses full FP32 precision without automatic mixed precision. During inference, experts are set to `eval()` and `requires_grad=False`. A small constant of $10^{-8}$ is added to avoid numerical instability in logarithms and division operations.

**Baseline Models (Full Details).** We compare against four categories of baselines. The first category comprises operational forecasting baselines, including ensemble mean and deterministic forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) and the U.S. National Centers for Environmental Prediction (NCEP), denoted as ECMWF-EnsMean, ECMWF-Determ, NCEP-EnsMean, and NCEP-Determ, respectively. Due to missing NCEP data for a portion of samples, NCEP-related baselines can only be evaluated on a reduced subset with complete NCEP records and serve solely as reference. The second category consists of sequential deep learning baselines, including BiLSTM, ConvLSTM, and TMPI. Their input preprocessing follows the TMPI scheme Luo et al. (2025), but omits the Soft-Track probabilistic encoding proposed in this work. The forecast lead times are aligned with those of our method. The third category comprises MoE variants for ablation, specifically Single-E (single MLP expert), Dense-MoE (dense PG-MoE),




(a) 2610 MAYSAK (b) 2613 DOLPHIN

Figure 5: Case visualization for Typhoon 2610 (a) and Typhoon 2613 (b). Rows correspond to Soft tracks, Final forecast, Quantitative result respectively.

NoPrior-R (unprior grouped MoE), E2E-Train (end-to-end training), and Flat-R (flat single-layer router). To ensure a fair comparison, all MoE variants take the identical 14-dimensional ellipse features extracted by Soft-Track as input.

**Evaluation Metrics.** The primary metric is Average Track Error (ATE) computed via spherical Haversine distance in kilometers. Results of learned models are reported as mean over three independent runs; operational forecasts are deterministic outputs without stochastic variation. In the ablation and interpretability analysis, we quantify expert load balancing and routing decision quality by examining the expert cluster activation frequency and Top-1 routing selection accuracy.

# B Additional Experiments

## B.1 Illustrative Case Studies of Two 2026 Typhoons

To further illustrate the generalization ability and physical interpretability of PG-MoE, we select two influential typhoons in 2026, Typhoon 2610 (MAYSAK) and Typhoon 2613 (DOLPHIN), for qualitative case studies. It should be noted that the test set used in the main experiments consists of typhoons from 2024 and 2025. Therefore, these two 2026 cases are not included in the test set and are not used in any quantitative metric computation or comparison; they are presented only for visual inspection and physical interpretation. Fig. 5 illustrates the ensemble tracks, Soft-Track trajectories, ATE curves, and expert activation frequencies.

For Typhoon 2610, the cyclone first recurves over the ocean and then moves northward before making landfall. The ensemble tracks become strongly dispersed near the recurvature point and the coast, while Soft-Track largely preserves the uncertainty. Within the 12–72H window, PG-MoE shows relatively low errors that are qualitatively comparable to mainstream operational forecasts. In terms of activation, the lead-time and coast-distance experts remain highly activated throughout; the turn-angle expert is active early but declines later; the intensity expert strengthens from the medium range; and the move-distance and area experts are almost inactive. This pattern echoes Maysak's oceanic recurvature, coastal landfall, and mid-term intensity enhancement.

For Typhoon 2613, a long-lived open-ocean typhoon, the cyclone moves over a long distance from the open ocean to landfall in China with no obvious recurvature, crossing multiple regions. Soft-Track mitigates trajectory drift at medium-to-long lead times. Within the 12–72H window, PG-MoE exhibits smooth error growth qualitatively comparable to mainstream operational forecasts. In terms of activation, the lead-time and intensity experts remain highly activated throughout; the move-distance expert is active early and decays rapidly later; the area expert gradually increases and

finally dominates at medium-to-long lead times; while the turn-angle and coast-distance experts are almost inactive. This reflects its long open-ocean movement, limited recurvature, and cross-region approach to landfall.

In summary, these two 2026 cases illustrate Soft-Track's capability in encoding ensemble uncertainty and the adaptive routing of MoE within the 12–72H window.

### B.2 Routing and Training Ablations

To validate the two-stage training pipeline, dedicated physical prior inputs, and the hierarchical two-level router, we compare PG-MoE with three variants: E2E-Train (end-to-end training), NoPrior-R (router without physical prior inputs), and Flat-R (single-layer router). Quantitative errors are reported in Tab. 6, while Fig. 6 illustrates expert activation frequencies and Top-1 routing accuracy.

E2E-Train eliminates expert pre-training and jointly optimizes all weights, which breaks the predefined physical grouping structure and causes ambiguous expert responsibilities. This is reflected in Fig. 6 as severely imbalanced activation and low routing accuracy. NoPrior-R discards physical prior features, leaving the gating network without sufficient guidance to discriminate across cyclone scenarios. Flat-R achieves the lowest or second-lowest errors across most horizons, yet its rigid hard-selection rules lead to extreme load imbalance and volatile routing accuracy (Fig. 6). In contrast, PG-MoE's hierarchical router imposes two-level constraints—coarse-grained group filtering and fine-grained expert selection—which naturally balances activation loads and stabilizes routing decisions. Collectively, these ablations verify that the synergy of decoupled training, physical prior inputs, and hierarchical routing is essential for interpretable expert grouping and robust long-term generalization.

### B.3 Boundary Expansion and Intra-Cluster Expert Count Ablations

To mitigate interval-boundary discontinuities, we compare four variants: single-expert without/with expansion (Top1-No, Top1-W) and dual-expert without/with expansion (Top2-No, PG-MoE). Tab. 7 and Fig. 7 report quantitative errors, activation frequencies, and routing accuracy. Single-expert variants force one network to fit the entire interval, yielding poor boundary modeling and imbalanced activation; dual-expert variants instead allow two subnetworks to separately specialize on central and boundary samples, smoothing discontinuities.

Boundary expansion significantly reduces errors under single-expert architectures by supplying missing transitional samples, and retains its benefits under dual-expert designs, though with a less pronounced effect. While PG-MoE demonstrates stronger short-term performance on the 2025 test set, the core advantage lies in the synergy of dual-experts and boundary expansion: together, they eliminate feature discontinuities from physical partitioning and achieve a balanced trade-off between short-term accuracy and long-term robustness.

### B.4 Grid Resolution Ablation

We compare three spatial resolutions for the region-partitioning scheme, corresponding to the hyperparameter $\delta$ defined in Sec. 3.1: $\delta = 0.2^\circ$, $\delta = 0.25^\circ$, and $\delta = 0.3^\circ$. Tab. 8 reports the mean $\Delta R$ (km) on the 2024 and 2025 test sets.

Based on these results, we adopt $0.25^\circ$ as the final resolution. It demonstrates the best or second-best accuracy for most lead times in both years, effectively balancing short- to medium-range precision with long-lead reliability. Specifically, the $0.2^\circ$ resolution introduces a lower initial spherical distance error due to finer grid partitioning, leading to better performance in the early horizons. However, as the forecast lead time extends, the accumulated long-term error itself far exceeds the initial advantage provided by the grid resolution. Furthermore, the higher resolution divides the probability distribution into smaller spatial units, which reduces the probability value in each individual region and consequently weakens the model's capability to capture long-term uncertainty, resulting in substantial errors at 120H. In contrast, $0.3^\circ$ only excels at the longest range at the cost of sacrificing accuracy in the earlier steps. Therefore, $0.25^\circ$ offers the most favourable trade-off between overall accuracy, stability, and computational efficiency.

Table 6: $\Delta R$ (km) for routing and training ablations on the 2024–2025 test sets.

| Horizon(2024) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
|---|---|---|---|---|---|---|---|---|---|---|
| E2E-Train | 45.07 | 44.63 | 50.02 | 57.24 | 71.38 | 86.89 | 111.38 | 145.59 | 245.43 | 375.72 |
| NoPrior-R | 45.52 | 44.79 | 50.84 | 57.41 | 72.46 | 88.6 | 110.64 | 138.37 | 216.89 | 339.23 |
| Flat-R | **42.94** | 43.34 | **48.97** | **55.57** | **70.23** | **85.35** | **106.32** | **133.8** | **212.14** | **335.87** |
| PG-MoE | 42.96 | **43.28** | 49.59 | 56.3 | 70.58 | 86.07 | 108.63 | 138.23 | 218.41 | 348.11 |
| Horizon(2025) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
| E2E-Train | 42.8 | 46.7 | 51.13 | 59.46 | 75.76 | 94.43 | 117.52 | 148.53 | 272.61 | 459.68 |
| NoPrior-R | 40.64 | 44.89 | 49.64 | 56.85 | 72.2 | 89.4 | **109.24** | **137.34** | **229.2** | **382.3** |
| Flat-R | **38.27** | **43.38** | **47.99** | **55.24** | 70.06 | 87.73 | 110.88 | 140.03 | 230.6 | 389.17 |
| PG-MoE | 40.24 | 44.28 | 48.45 | 55.44 | **69.57** | **86.79** | 111.41 | 141.13 | 230.34 | 388.65 |

Table 7: $\Delta R$ (km) for boundary and intra-cluster expert ablations on 2024 and 2025 test sets.

| Horizon(2024) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
|---|---|---|---|---|---|---|---|---|---|---|
| Top1-NoExpand | 55.84 | 46.6 | 52.25 | 58.01 | 72.25 | 90.18 | 111.79 | 139.37 | 219.56 | 343.87 |
| Top1-WExpand | 45.01 | 45.65 | 51.33 | 57.43 | 72.11 | 87.73 | **107.97** | **134.67** | **210.14** | **333.47** |
| Top2-NoExpand | 43.68 | 43.94 | 49.75 | **55.92** | **70.28** | 86.75 | 109.32 | 139.35 | 219.71 | 348.84 |
| PG-MoE | **42.96** | **43.28** | **49.59** | 56.3 | 70.58 | **86.07** | 108.63 | 138.23 | 218.41 | 348.11 |
| Horizon(2025) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
| Top1-NoExpand | 52.95 | 46.31 | 51.03 | 58.02 | 72.95 | 89.81 | 111.94 | 141.44 | 242.24 | 401.92 |
| Top1-WExpand | 42.59 | 46.95 | 50.95 | 57.57 | 72.74 | 89.25 | **109.65** | 139.38 | 237.08 | 399.02 |
| Top2-NoExpand | **39.51** | **44.25** | 48.78 | 55.57 | 70.73 | 87.29 | 110.03 | **137.95** | 231.8 | 390.83 |
| PG-MoE | 40.24 | 44.28 | **48.45** | **55.44** | **69.57** | **86.79** | 111.41 | 141.13 | **230.34** | **388.65** |

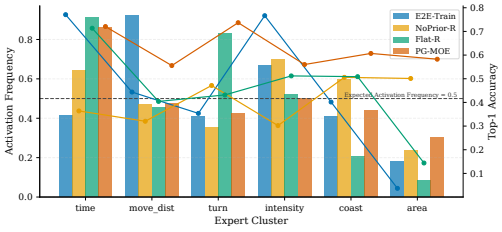



Figure 6: Activation frequency (bars, left axis) and top-1 routing accuracy (solid lines, right axis) across six physical expert clusters; the black dashed line marks the balanced expected activation frequency of 0.5.

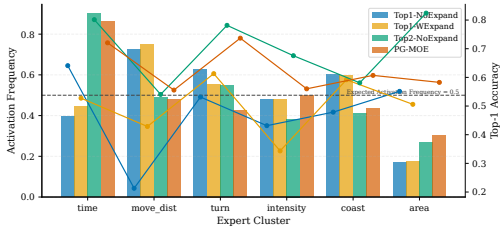



Figure 7: Expert cluster activation frequency (left y-axis) and top-1 routing selection accuracy (right y-axis) for boundary ablation variants. The dashed line marks the balanced expected activation frequency of 0.5.

Table 8: $\Delta R$ (km) for different grid resolutions at selected horizons on the 2024 and 2025 test sets.

| Horizon(2024) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
|---|---|---|---|---|---|---|---|---|---|---|
| $\delta = 0.2°$ | **42.78** | 43.47 | **48.06** | **55.53** | **70.37** | 85.36 | **106.25** | **134.05** | 218.16 | 372.09 |
| $\delta = 0.25°$ | 42.96 | **43.28** | 49.59 | 56.3 | 70.58 | 86.07 | 108.63 | 138.23 | 218.41 | 348.11 |
| $\delta = 0.3°$ | 42.89 | 43.69 | 49.03 | 56.55 | 71.11 | **84.29** | 107.39 | 137.43 | **212.86** | **345.15** |
| Horizon(2025) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
| $\delta = 0.2°$ | **38.58** | **43.67** | **47.8** | 55.64 | 70.87 | 89.77 | 115.38 | 142.15 | 228.45 | 418.06 |
| $\delta = 0.25°$ | 40.24 | 44.28 | 48.45 | **55.44** | **69.57** | 86.79 | **111.41** | **141.13** | 230.34 | 388.65 |
| $\delta = 0.3°$ | 40.46 | 45.02 | 49.17 | 56.19 | 70.3 | **86.51** | 112.2 | 142.02 | **219.45** | **360.76** |

Table 9: $\Delta R$ (km) for different chi-squared quantiles at selected horizons on the 2024 and 2025 test sets.

| Horizon(2024) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
|---|---|---|---|---|---|---|---|---|---|---|
| $q = 50\%$ | **42.72** | **43.27** | **49.1** | **55.78** | **70.53** | 86.62 | 109.49 | 138.9 | 219.95 | **347.55** |
| $q = 75\%$ | 42.96 | 43.28 | 49.59 | 56.3 | 70.58 | **86.07** | **108.63** | **138.23** | **218.41** | 348.11 |
| $q = 90\%$ | 43.23 | 43.39 | 49.25 | 55.96 | 70.76 | 87.12 | 110.45 | 139.81 | 221.81 | 352.46 |
| Horizon(2025) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
| $q = 50\%$ | **38.21** | **43.45** | **47.76** | **55.04** | 71.09 | 89.4 | 113.06 | 141.92 | 232.93 | **379.74** |
| $q = 75\%$ | 40.24 | 44.28 | 48.45 | 55.44 | **69.57** | **86.79** | **111.41** | **141.13** | **230.34** | 388.65 |
| $q = 90\%$ | 39.1 | 43.74 | 48.06 | 55.15 | 70.82 | 90.18 | 113.6 | 141.18 | 235.49 | 392.16 |

### B.5 Chi-Squared Quantile Ablation

We investigate the effect of the chi-squared quantile used for ellipse fitting, which determines the coverage range and the number of points included in each regional expert. Three quantiles corresponding to the hyperparameter $q$ defined in Sec. 3.1 are evaluated: $q = 50\%$, $q = 75\%$, and $q = 90\%$. Tab. 9 reports the mean $\Delta R$ (km) at selected horizons on the 2024 and 2025 test sets.

The experimental results indicate that the performance differences across different confidence levels are relatively small, with all variations falling within a reasonable range. Specifically, while the $q = 90\%$ quantile encompasses a broader set of effective probability regions, it simultaneously incorporates a substantial number of low-probability or noise regions, which introduces redundant information and slightly hampers generalization. Conversely, the $q = 50\%$ quantile produces a more compact ellipse that focuses on the core probability area, but may over-constrain the representation of spatial uncertainty, thereby limiting its generalization capability. Overall, the $q = 75\%$ quantile achieves the optimal balance between probability coverage and feature representation stability, avoiding noise from over-expansion while preserving sufficient spatial uncertainty information. Based on this trade-off, we adopt $q = 75\%$ as the default configuration.

### B.6 Number of Boundary Sampling Points Ablation

We compare three settings for the number of boundary points sampled from the ellipse boundary, corresponding to the hyperparameter $N_p$ defined in Sec. 3.1: $N_p = 4$ (cardinal directions: up, down, left, right), $N_p = 6$ (left, right, and four diagonal directions), and $N_p = 8$ (all cardinal and diagonal directions). Note that the total number of extracted points is $N_p + 1$ (including the weighted center). Tab. 10 reports the mean $\Delta R$ (km) at selected horizons on the 2024 and 2025 test sets.

Table 10: $\Delta R$ (km) for different numbers of boundary sampling points at selected horizons on the 2024 and 2025 test sets.

| Horizon(2024) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
|---|---|---|---|---|---|---|---|---|---|---|
| $N_p = 4$ | 43.75 | 44.2 | 49.95 | 56.57 | 71.16 | 87.58 | 109.05 | 140.11 | **218.33** | **337.39** |
| $N_p = 6$ | **42.96** | **43.28** | 49.59 | 56.3 | **70.58** | **86.07** | **108.63** | **138.23** | 218.41 | 348.11 |
| $N_p = 8$ | 43.28 | 43.9 | **49.55** | **56.26** | 70.77 | 87.34 | 110.47 | 139.63 | 221.41 | 350.39 |
| Horizon(2025) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
| $N_p = 4$ | 39.5 | **43.63** | **48.0** | 55.13 | 70.31 | 90.24 | 113.77 | 142.64 | 233.58 | **378.71** |
| $N_p = 6$ | 40.24 | 44.28 | 48.45 | 55.44 | **69.57** | **86.79** | **111.41** | **141.13** | **230.34** | 388.65 |
| $N_p = 8$ | **38.73** | 43.66 | 48.23 | **55.08** | 70.89 | 90.04 | 114.42 | 145.48 | 234.26 | 386.2 |

The experimental results indicate that the number of sampling points has a relatively minor impact on overall performance. Specifically, the $N_p = 4$ setting retains only core directional information, which may lead to insufficient characterization of spatial uncertainty and limited generalization capability. Conversely, the $N_p = 8$ setting provides a more comprehensive coverage of directional ranges and more clearly delineates the boundary, but the excessive peripheral points may cause the model to over-focus on the boundary range rather than the central trajectory during subsequent prediction, thereby introducing noise and degrading predictive accuracy. In contrast, the $N_p = 6$ setting achieves a better balance between sufficient shape representation and focused attention on the central trajectory, avoiding both the information loss from under-sampling and the attention shift caused by over-sampling. Therefore, we adopt $N_p = 6$ as the final choice, as it provides adequate elliptical shape representation while maintaining robust generalization performance.

### B.7 Global Routing Top-K Ablation

We investigate the effect of the number of expert clusters activated by the global router, corresponding to the hyperparameter $K_g$ defined in Sec. 3.3. With a total of $G = 6$ clusters, we evaluate $K_g = 1$ to 6. Tab. 11 reports the mean $\Delta R$ (km) at selected horizons on the 2024 and 2025 test sets.

The experimental results indicate that $K_g = 1$ represents the primary performance bottleneck, as activating only one expert cluster prevents the model from effectively capturing diverse typhoon dynamics, resulting in extremely large errors. When $K_g$ increases to 2, performance improves significantly. Notably, the performance differences between $K_g = 2$ and $K_g = 6$ are relatively small overall, all within a reasonable range. Although larger $K_g$ values occasionally outperform $K_g = 3$ at individual points, $K_g = 3$ consistently delivers stable and superior performance across the vast majority of horizons and both test years, particularly at medium and long lead times. Therefore, we select $K_g = 3$ as the optimal configuration, achieving a favourable balance between accuracy, generalisation, and computational efficiency.

### B.8 Intra-Cluster Routing Top-K Ablation

We evaluate the effect of the number of experts activated within each cluster, corresponding to the hyperparameter $K_e$ defined in Sec. 3.3. We test $K_e = 1$, 2, and 3. Tab. 12 reports the mean $\Delta R$ (km) at selected horizons on the 2024 and 2025 test sets.

The experimental results indicate that the overall performance differences between $K_e = 1$, 2, and 3 are relatively small. Specifically, $K_e = 1$ occasionally achieves the best performance at very long lead times (e.g., 2024-120H), but exhibits noticeably larger errors in short- and medium-range forecasts, suggesting that a single expert lacks sufficient capacity to capture early-stage dynamics. $K_e = 3$ provides only marginal improvements over $K_e = 2$ at a few horizons, and the third expert introduced often lacks sufficient training on the current samples, potentially injecting unnecessary noise into the prediction. In contrast, $K_e = 2$ consistently delivers stable and competitive performance across the majority of lead times in both years, particularly excelling at short-to-medium

Table 11: $\Delta R$ (km) for different global routing top-$K_g$ values at selected horizons on the 2024 and 2025 test sets.

| Horizon(2024) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
|---|---|---|---|---|---|---|---|---|---|---|
| $K_g = 1$ | 112.56 | 112.88 | 116.75 | 120.69 | 133.92 | 144.71 | 164.25 | 193.91 | 261.71 | 392.36 |
| $K_g = 2$ | 42.99 | 43.35 | 49.33 | **55.86** | **70.11** | 86.89 | 110.76 | 141.32 | 225.32 | 351.59 |
| $K_g = 3$ | **42.96** | **43.28** | 49.59 | 56.3 | 70.58 | **86.07** | **108.63** | 138.23 | 218.41 | 348.11 |
| $K_g = 4$ | 43.24 | 43.52 | 49.71 | 56.48 | 71.19 | 87.01 | 109.35 | 138.94 | 223.97 | 360.63 |
| $K_g = 5$ | 43.1 | 43.29 | **49.2** | 56.07 | 70.78 | 86.57 | 109.45 | 138.59 | 224.47 | 355.05 |
| $K_g = 6$ | 43.1 | 43.5 | 49.36 | 56.33 | 70.45 | 86.32 | 108.68 | **137.07** | **217.26** | **342.68** |
| Horizon(2025) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
| $K_g = 1$ | 111.09 | 110.11 | 108.74 | 113.61 | 127.36 | 144.58 | 165.05 | 196.4 | 282.25 | 420.35 |
| $K_g = 2$ | **38.76** | **43.4** | **47.83** | **54.96** | 70.22 | 87.42 | **110.98** | 140.47 | 240.62 | 404.81 |
| $K_g = 3$ | 40.24 | 44.28 | 48.45 | 55.44 | **69.57** | **86.79** | 111.41 | 141.13 | **230.34** | **388.65** |
| $K_g = 4$ | 39.22 | 43.93 | 48.88 | 56.26 | 71.31 | 89.76 | 111.99 | **139.5** | 237.37 | 395.69 |
| $K_g = 5$ | 39.22 | 43.93 | 48.51 | 56.17 | 71.46 | 90.49 | 112.82 | 140.16 | 234.0 | 392.76 |
| $K_g = 6$ | 38.96 | 43.84 | 48.45 | 56.2 | 71.38 | 89.15 | 111.18 | 140.38 | 235.84 | 399.57 |

Table 12: $\Delta R$ (km) for different intra-cluster routing top-$K_e$ values at selected horizons on the 2024 and 2025 test sets.

| Horizon(2024) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
|---|---|---|---|---|---|---|---|---|---|---|
| $K_e = 1$ | 45.01 | 45.65 | 51.33 | 57.43 | 72.11 | 87.73 | **107.97** | **134.67** | **210.14** | **333.47** |
| $K_e = 2$ | **42.96** | **43.28** | 49.59 | **56.3** | **70.58** | **86.07** | 108.63 | 138.23 | 218.41 | 348.11 |
| $K_e = 3$ | 43.14 | 43.89 | **49.59** | 56.6 | 71.22 | 87.95 | 111.53 | 139.73 | 217.25 | 347.52 |
| Horizon(2025) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
| $K_e = 1$ | 42.59 | 46.95 | 50.95 | 57.57 | 72.74 | 89.25 | **109.65** | **139.38** | 237.08 | 399.02 |
| $K_e = 2$ | 40.24 | 44.28 | 48.45 | **55.44** | **69.57** | **86.79** | 111.41 | 141.13 | **230.34** | **388.65** |
| $K_e = 3$ | **39.29** | **44.26** | **48.35** | 55.55 | 70.15 | 87.61 | 111.66 | 139.85 | 235.05 | 397.64 |

ranges (12H–48H) while maintaining good robustness at long lead times. This balance between accuracy and generalization makes $K_e = 2$ the preferred choice.

## C Complete Experimental Results

This section reports quantitative track error $\Delta R$ (km) results for all baseline models and ablation variants, covering a more complete set of forecast lead times across the two test years (2024, 2025). In contrast to the main-text table, which only reports mean errors for a limited selection of horizons, the tables below provide extended temporal coverage.

Table 13: Full version of Main Text Tab. 3 — SoftTrack ablation

| Horizon(2024) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
|---|---|---|---|---|---|---|---|---|---|---|
| NoSoftTrack | **36.0** | 47.02 | 52.74 | 60.37 | 83.68 | 105.23 | 140.87 | 184.75 | 290.37 | 508.03 |
| SoftTrack-probability | 44.52 | 45.12 | 50.25 | 56.69 | 72.48 | 88.5 | **108.5** | 139.12 | 227.48 | **345.87** |
| PG-MoE | 42.96 | **43.28** | **49.59** | **56.3** | **70.58** | **86.07** | 108.63 | **138.23** | **218.41** | 348.11 |
| Horizon(2025) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
| NoSoftTrack | **39.04** | 48.01 | 54.28 | 64.14 | 82.11 | 114.21 | 153.17 | 200.03 | 351.93 | 444.78 |
| SoftTrack-probability | 40.7 | 45.11 | 49.92 | 57.13 | 73.7 | 91.84 | 119.71 | 148.44 | 259.38 | 410.7 |
| PG-MoE | 40.24 | **44.28** | **48.45** | **55.44** | **69.57** | **86.79** | **111.41** | **141.13** | **230.34** | **388.65** |

Table 14: Full version of Main Text Tab. 4 — MoE architecture ablation

| Horizon(2024) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
|---|---|---|---|---|---|---|---|---|---|---|
| Single-E | 66.17 | 64.38 | 69.08 | 73.55 | 84.82 | 96.59 | 117.22 | 141.14 | **214.41** | **332.98** |
| Dense-MoE | 43.24 | 43.56 | **49.4** | 56.39 | 70.86 | 86.59 | **107.62** | **135.6** | 216.3 | 339.27 |
| PG-MoE | **42.96** | **43.28** | 49.59 | **56.3** | **70.58** | **86.07** | 108.63 | 138.23 | 218.41 | 348.11 |
| Horizon(2025) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
| Single-E | 69.99 | 70.95 | 74.23 | 78.12 | 86.99 | 99.38 | 115.51 | 140.72 | **227.1** | **382.48** |
| Dense-MoE | **39.07** | **43.77** | **48.16** | 55.54 | 70.57 | 88.23 | **111.23** | **140.62** | 236.05 | 391.84 |
| PG-MoE | 40.24 | 44.28 | 48.45 | **55.44** | **69.57** | **86.79** | 111.41 | 141.13 | 230.34 | 388.65 |

Table 15: Full version of Main Text Tab. 5 — Single cluster ablation

| Horizon(2024) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
|---|---|---|---|---|---|---|---|---|---|---|
| w/o-time-group | 43.82 | 44.45 | 50.22 | 56.82 | 70.68 | 86.63 | **108.16** | **137.07** | 217.26 | 355.7 |
| w/o-move-dist-group | 43.3 | 43.64 | 49.47 | 56.48 | 70.97 | 86.92 | 108.23 | 137.83 | 219.75 | 347.4 |
| w/o-turn-group | 43.24 | 43.56 | 49.57 | 56.8 | 71.61 | 87.01 | 109.62 | 138.69 | 222.16 | 355.56 |
| w/o-intensity-group | 43.1 | 43.85 | 49.82 | 56.63 | 71.5 | 87.63 | 109.22 | 137.4 | **216.27** | 348.21 |
| w/o-coast-group | **42.96** | 43.38 | **49.23** | **56.2** | 71.01 | 87.18 | 108.58 | 137.51 | 219.87 | 348.56 |
| w/o-area-group | 43.33 | 43.7 | 49.51 | 56.21 | 70.91 | 87.29 | 109.41 | 137.88 | 216.69 | **341.33** |
| PG-MoE | 42.96 | **43.28** | 49.59 | 56.3 | **70.58** | **86.07** | 108.63 | 138.23 | 218.41 | 348.11 |
| Horizon(2025) | 6H | 12H | 18H | 24H | 36H | 48H | 60H | 72H | 96H | 120H |
| w/o-time-group | 40.01 | 45.03 | 49.17 | 56.29 | 71.28 | 88.26 | **110.0** | **137.63** | **226.0** | **376.78** |
| w/o-move-dist-group | 39.33 | 44.02 | **48.32** | 55.82 | 70.52 | 89.46 | 113.66 | 140.63 | 231.1 | 389.4 |
| w/o-turn-group | 39.21 | **43.87** | 48.5 | 55.93 | 70.44 | 87.45 | 112.38 | 140.49 | 231.81 | 387.19 |
| w/o-intensity-group | 39.31 | 44.05 | 48.48 | **55.39** | 70.42 | 89.19 | 112.59 | 142.54 | 234.08 | 386.42 |
| w/o-coast-group | 39.21 | 44.11 | 48.63 | 55.73 | 70.07 | 89.55 | 113.05 | 140.73 | 231.65 | 383.68 |
| w/o-area-group | **39.15** | 43.97 | 48.61 | 56.02 | 71.47 | 90.3 | 112.58 | 142.42 | 234.55 | 389.11 |
| PG-MoE | 40.24 | 44.28 | 48.45 | 55.44 | **69.57** | **86.79** | 111.41 | 141.13 | 230.34 | 388.65 |