Appendix

A. BlackBird Dataset

The Blackbird unmanned aerial vehicle (UAV) dataset is an indoor dataset [28] designed to capture aggressive flight maneuvers for fully autonomous drone racing. It collects data using a Blackbird quadrotor platform with an Xsens MTi-3 IMU. The blackbird takes place in a motion capture room and follows a predefined periodic trajectory, each lasting approximately 3-4 minutes at high speed.

1) Seen and Unseen sequences separation: In our experiments, we define two distinct groups of trajectories: **SEEN** and **UNSEEN** sequences. SEEN sequences appear in both the training and testing phases, while UNSEEN sequences do not appear in any phase of the training process. Specifically, for SEEN sequences, we use the same five sequences that were used in IMO: `clover`, `egg`, `halfMoon`, `star`, and `winter`, with peak velocities of 5, 8, 4, 5, 4$\mathrm{m\,s^{-1}}$, respectively. Each trajectory is split into training, validation, and testing sets. Since these trajectories appear in both the training and testing sets, we term them as **SEEN sequences**. To further evaluate the model's ability to adapt to unseen trajectories, we also select five additional trajectories from the Blackbird dataset: `ampersand`, `sid`, `oval`, `sphinx`, and `bentDice`, with peak velocities of 2, 5, 4, 4, 3$\mathrm{m\,s^{-1}}$, respectively. Compared to the SEEN sequences, these new trajectories never appear in training or validation; therefore, we refer to them as **UNSEEN sequences**. By comparing results on both SEEN and UNSEEN sequences, we could gain a comprehensive understanding of the model's robustness and generalization capabilities.

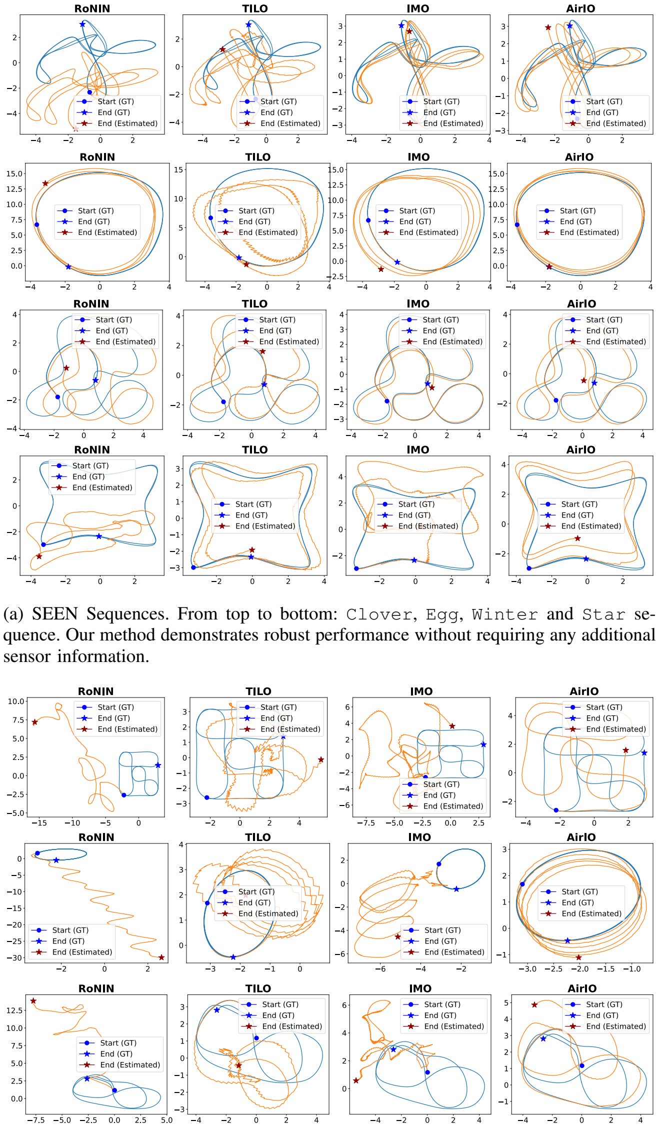
2) Training and Testing Sequences Separation: In our experiments, we follow the same dataset-splitting strategy presented in IMO: for each trajectory, the data is allocated as 70% for training, 15% for validation, and the remaining 15% for testing. We use the SEEN sequences' training and validation set to train our model, making our training setup identical to IMO's. For testing, we employ both the SEEN sequences' testing set and the UNSEEN sequences' testing set. The comprehensive testing setup allows us to evaluate our method's generalization to new trajectories.

TABLE IV: Separation of trajectory sequences into SEEN and UNSEEN categories, and their respective allocations to training, validation, and testing sets.

SEEN	clover	Egg	halfMoon	Star	Winter
training (70%)	✔	✔	✔	✔	✔
validation (15%)	✔	✔	✔	✔	✔
testing (15%)	✔	✔	✔	✔	✔
UNSEEN	Ampersand	Sid	Oval	Sphinx	BentDice
training (70%)	✘	✘	✘	✘	✘
validation (15%)	✘	✘	✘	✘	✘
testing (15%)	✔	✔	✔	✔	✔

B. Qualitative Evaluation for Blackbird dataset

We present more details on the evaluation of the Blackbird dataset. As shown in 11, we showcase seven additional trajectories that further highlight our method's performance. Our method achieves superior performance in the SEEN sequences. For more than half of these sequences, it outperforms existing methods that rely on additional information in addition to IMU measurements, while our method uses only IMU data. When evaluating UNSEEN sequences, our model outperforms all existing methods on all sequences, demonstrating its remarkable adaptability.



(a) SEEN Sequences. From top to bottom: `Clover`, `Egg`, `Winter` and `Star` sequence. Our method demonstrates robust performance without requiring any additional sensor information.

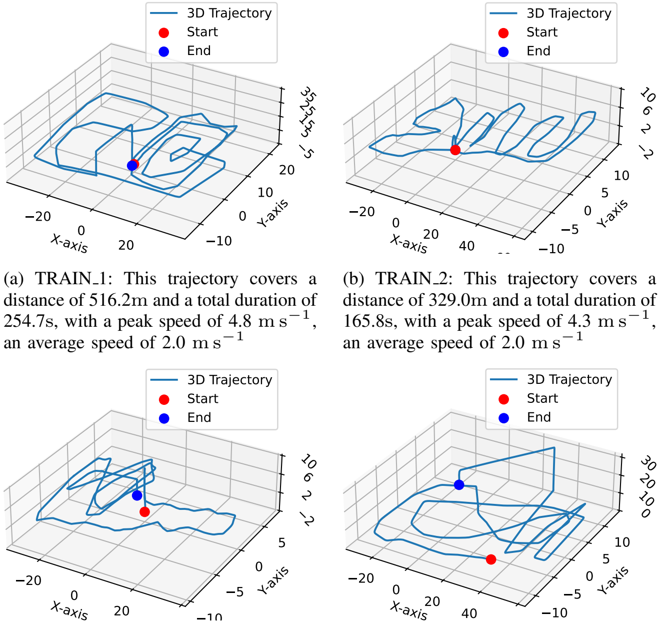
(b) UNSEEN Sequences. From top to bottom: `bentDice`, `Oval`, and `Sphinx` sequence. Our method demonstrates remarkable adaptability to trajectories it has never seen before.

Fig. 11: Estimated trajectories of Blackbird dataset by RoNIN, TLIO, IMO and AirIO (Ours).

C. Pegasus Dataset

We collected a simulation dataset in the open-source Pegasus Simulator [29] to evaluate our proposed method under controlled conditions. Pegasus is a framework built on top of NVIDIA Omniverse and Isaac Sim. It is designed for multirotor simulation and supports integration with PX4 firmware, as well as Python control interfaces. In our setup, we used QGroundControl to control the multirotor and also employed the quadratic thrust curve and linear drag model, ensuring the generated flight dynamics closely resemble real-world conditions.

In our experiment, we collected a total of seven trajectories datasets, named **Pegasus Dataset**. We divided the Pegasus dataset into training and testing sets. Specifically, four trajectories are selected for training and the remaining three trajectories are testing. They are illustrated in 12 and 13. We provide a detailed overview of each trajectory as follows.



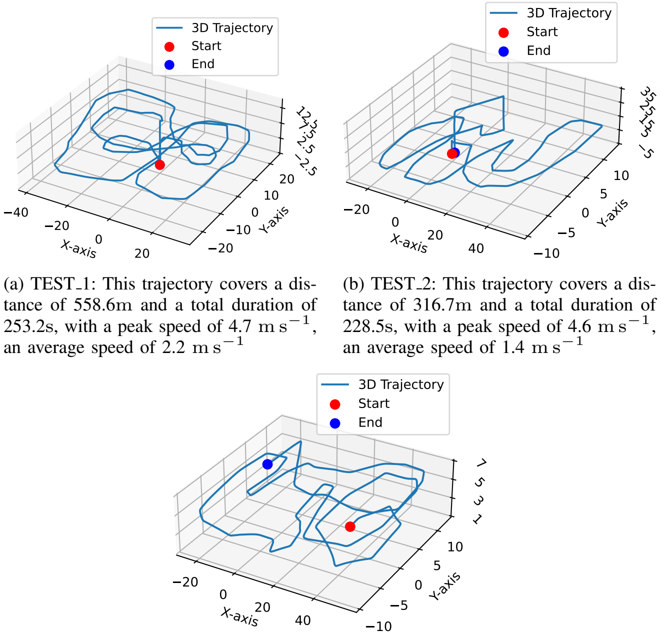
(a) TRAIN_1: This trajectory covers a distance of 516.2m and a total duration of 254.7s, with a peak speed of 4.8 m s^{-1}, an average speed of 2.0 m s^{-1}

(b) TRAIN_2: This trajectory covers a distance of 329.0m and a total duration of 165.8s, with a peak speed of 4.3 m s^{-1}, an average speed of 2.0 m s^{-1}

(c) TRAIN_3: This trajectory covers a distance of 263.5m and a total duration of 355.5s, with a peak speed of 4.6 m s^{-1}, an average speed of 0.7 m s^{-1}

(d) TRAIN_4: This trajectory covers a distance of 452.1m and a total duration of 160.4s, with a peak speed of 4.9 m s^{-1}, an average speed of 2.8 m s^{-1}

Fig. 12: Training set



(a) TEST_1: This trajectory covers a distance of 558.6m and a total duration of 253.2s, with a peak speed of 4.7 m s^{-1}, an average speed of 2.2 m s^{-1}

(b) TEST_2: This trajectory covers a distance of 316.7m and a total duration of 228.5s, with a peak speed of 4.6 m s^{-1}, an average speed of 1.4 m s^{-1}

(c) TEST_3: This trajectory covers a distance of 402.1m and a total duration of 148.8s, with a peak speed of 4.9 m s^{-1}, an average speed of 2.7 m s^{-1}

Fig. 13: Testing set

D. EuRoC Dataset

The EuRoC datasets are the well-known benchmarks for odometry and SLAM algorithms. They are collected by a micro aerial vehicle: an AscTec Firefly hex-rotor helicopter. There are 11 trajectories collected in two scenarios: an industrial environment and a motion capture room. We selected `MH_01_easy`, `MH_03_medium`, `MH_05_difficult`, `V1_02_medium`, `V2_01_easy`, `V2_03_difficult` for training, and the rest for testing.

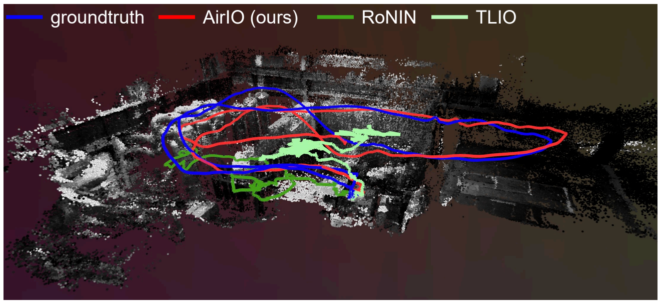


Fig. 14: The MH_04_difficult trajectories from the EuRoC dataset visualized within its 3D reconstruction map. While RoNIN (dark green) and TLIO (light green) fail, AirIO (red) retains a coherent trajectory shape.

E. Ablation Study in Pegasus and EuRoC dataset

TABLE V: Ablation study on the EuRoC and Pegasus datasets comparing different feature representations. Evaluation metric: RTE (Unit: m).

Seq.	Global − gravity	Global	Body − gravity	Global +Attitude	Body	Body +Attitude
EuRoC						
MH02	1.684	1.575	2.346	1.542	1.314	**0.972**
MH04	2.618	1.961	2.525	1.707	1.329	**1.009**
V103	1.407	**1.352**	1.613	1.485	1.623	1.512
V202	1.721	1.789	2.176	1.723	1.373	**1.263**
V101	1.801	1.991	1.463	1.498	1.122	**1.104**
Avg.	1.846	1.734	2.025	1.591	1.352	**1.172**
Pegasus						
TEST_1	2.783	2.971	2.134	1.694	1.137	**1.017**
TEST_2	2.704	3.007	1.961	2.339	2.162	**1.905**
TEST_3	3.274	3.350	1.298	1.969	0.584	**0.396**
Avg.	2.920	3.109	1.798	2.001	1.294	**1.106**
Blackbird						
Ampersand	8.933	6.845	11.379	8.374	1.254	**1.185**
Sid	7.768	6.151	3.575	2.72	0.737	**0.504**
Oval	4.154	3.936	1.094	4.555	0.690	**0.558**
Sphinx	4.637	3.647	1.695	2.875	**0.888**	1.021
BentDice	6.998	5.587	5.427	5.578	1.389	**0.871**
Avg.	6.498	5.233	4.634	4.82	0.992	**0.828**

F. Ablation Study on Model Compression

To evaluate the compressibility of different representations, we introduced additional two lightweight models **Light A** and **Light B**. The light models keep the same layer structure but shrink the dimensions of each layer's hidden units. Finally, the encoder's latent feature dimension is reduced from 256 to 128, and then further to 64, yielding progressively smaller models.

To quantify the performance degradation as the model is compressed, we define the degradation ratio for ATE and RTE. A higher degradation ratio indicates a larger drop in performance. As shown in VI, the model under body frame shows smoother degradation in both ATE and RTE.

TABLE VI: Ablation study on the Blackbird, Pegasus, and EuROC datasets comparing compressibility of models under body frame and global frame. Evaluation metric: ATE (Unit: m), RTE(Unit: m), and Degradation.

Model		**Regular**		**Light A**		**Light B**	
Feature Size		256×1		128×1		64×1	
Model Size		2.524 MB		0.641 MB		0.175 MB	
Metrics		ATE	Degradation	ATE	Degradation	ATE	Degradation
Blackbird	Body	0.647	-	0.755	16.8%	0.931	44.0%
	Global	0.837	-	1.238	47.9%	1.522	81.8%
Pegasus	Body	4.670	-	10.118	116.6%	15.192	225.3%
	Global	17.278	-	30.950	79.1%	69.236	300.7%
EuRoC	Body	4.730	-	5.447	15.2%	6.875	45.4%
	Global	10.096	-	14.033	39.0%	38.236	278.7%
Metrics		RTE	Degradation	RTE	Degradation	RTE	Degradation
Blackbird	Body	0.345	-	0.454	31.3%	0.510	47.7%
	Global	0.525	-	0.583	11.0%	0.983	87.1%
Pegasus	Body	1.516	-	2.226	46.9%	2.203	45.3%
	Global	3.109	-	3.422	10.0%	4.858	56.2%
EuRoC	Body	1.352	-	1.359	0.5%	1.297	-4.1%
	Global	1.734	-	2.176	25.5%	4.468	157.8%