# EES: A Data-Driven End-to-End Escorting System via Spatiotemporal Feature Fusion

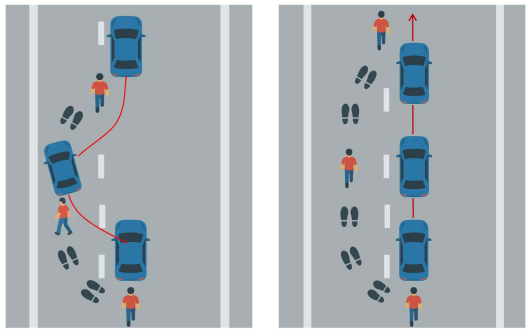
Youjin Yu, Junxiang Li, Bowen Li, Tao Wu, and Huijing Zhao, *Senior Member, IEEE*

***Abstract*—This letter presents a technique that allows unmanned vehicles to escort a human to their destinations. Current human-centered following methods depend solely on human movement, which presents significant limitations. The complexity of human movement during tactical maneuvers can lead to erratic vehicle motion. Additionally, the static relative positioning between the human and vehicle creates a rigid following pattern, thereby constraining the vehicle's ability to dynamically adjust its position for optimal coverage. To address these limitations, we propose a data-driven end-to-end escorting system (EES) that takes into account both environmental information and human movement to achieve adaptive escorting. We propose a soft-coding paradigm to replace the traditional hard-coding intent modeling to address the inconsistency of human intention and vehicle motion, and establish human-scene following through a cross-modal attention gating network. We conducted experiments in the CARLA simulation and the real world. The results demonstrate that the proposed EES reduces prediction errors by 41.2% during overall processes and by 54.5% during cornering. Additionally, EES can adapt to various positions and dynamically adjust the relative positions between humans and unmanned systems to adapt to complex scenarios.**

***Index Terms*—Human-robot collaboration, autonomous vehicle navigation, motion and path planning.**

## I. Introduction

With the widespread application of the unmanned vehicles, the human-following system is designed to escort and assist personnel [1], providing support in various tasks and scenarios. It can enhance the safety and efficiency of personnel in performing tasks, especially in high-risk missions such as urban combat and border patrol. Existing research has predominantly employed human-centered approaches, which strictly track human movement to guide robotic actions, positioning humans as the central element in the following strategy. While these approaches are effective in many situations, such as medical robots, service robots, and luggage-carrying robots [2], [3], they are not suited well for military applications. In military environments, soldiers often exhibit erratic motion patterns, such as zigzag movements and sharp stops, in response to potential hazards. These complex and unpredictable human trajectories result in unstable vehicle motion, posing significant challenges for traditional human-centered following approaches. Moreover, the reliance on a static relative position between the human and the vehicle results in an inflexible following pattern. This rigidity prevents the vehicle from dynamically adjusting its position to provide optimal cover or support, thereby diminishing the efficacy of the following system. Inspired by [1], [4], [5], we propose an innovative end-to-end escorting system that integrates environmental information with human movement patterns, creating a flexible and adaptive following strategy for unmanned vehicles to offer escort and support in complex environments.



(a) Human-centered Following (b) Escorting System

Fig. 1. **Overview of the task.** (a) In human-centered following, the vehicle tracks the movement of the human by considering the future trajectyories and intentions; (b) In escorting system, the movement of vehicle not only relies on human movement, but also environmental information.

We develop a data-driven escorting system that takes full account of environmental information to enable more flexible and adaptive following. Our contributions have three aspects:

1) We propose an end-to-end escorting architecture fusing vehicle perceptual information with human motion patterns. The architecture can realize robust escorting regardless of different positions and orientations of the human.
2) Our soft-coding paradigm encodes human trajectory features as high-level navigation commands, reducing the inconsistency between long-term human intention and short-term vehicle motion. Additionally, a cross-modal attention mechanism is introduced to balance environmental perception with human movement.

Received 21 May 2025; accepted 11 September 2025. Date of publication 3 October 2025; date of current version 21 October 2025. This article was recommended for publication by Associate Editor Andrea Pupa and Editor Angelika Peer upon evaluation of the reviewers' comments. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62103431, Grant U21A20518, and Grant U22A2061. *(Corresponding authors: Junxiang Li; Tao Wu.)*

Youjin Yu, Junxiang Li, Bowen Li, and Tao Wu are with the College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail: yuyoujin23@nudt.edu.cn; lijunxiang@nudt.edu.cn; wutao@nudt.edu.cn).

Huijing Zhao is with the Institute for Artificial Intelligence, Peking University, Beijing 100871, China (e-mail: zhaohj@cis.pku.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/LRA.2025.3617728, provided by the authors.

Digital Object Identifier 10.1109/LRA.2025.3617728

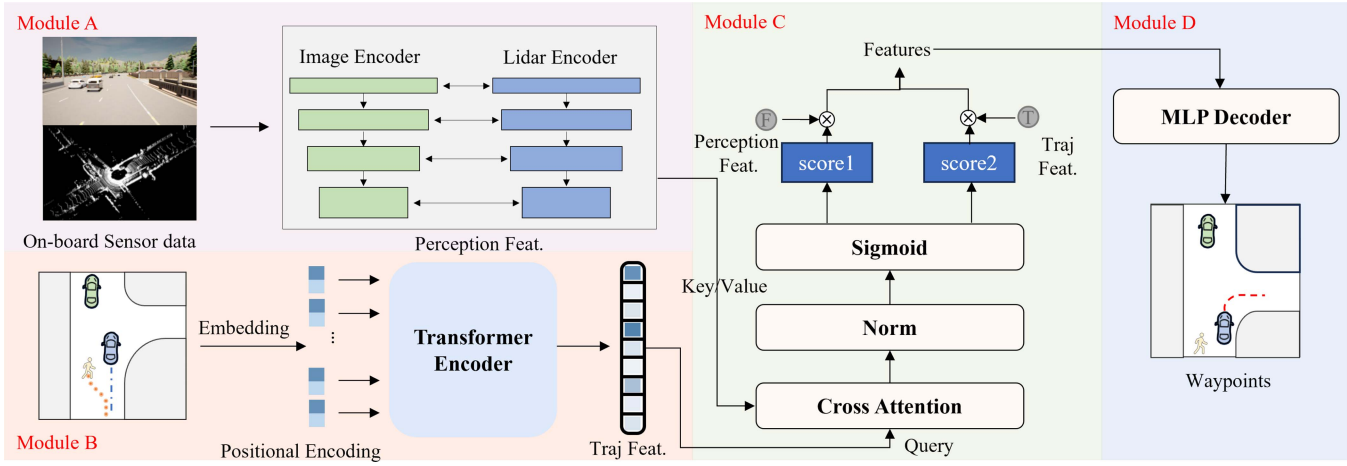



Fig. 2. **Overview architecture of EES.** EES maps the inputs of vehicle perceptual information and human historical trajectories to vehicle waypoints. The model includes modules for perceptual feature extraction, human trajectory feature extraction, and cross-modal attention gating structure.

TABLE I
RESULTS ON CARLA DATASETS

| Exp | ADE ↓ | FDE ↓ | ADE_turn ↓ | FDE_turn ↓ |
|---|---|---|---|---|
| EKF | 1.6765 | 2.9237 | 2.4362 | 4.6929 |
| Trajectory-based | 1.0684 | 1.6499 | 1.3884 | 2.4130 |
| Intention-based | 1.1757 | 1.9298 | 1.7070 | 3.1942 |
| EES(ours)† | **0.9863** | **1.4026** | **1.1092** | **1.7295** |

† To maintain comparability, this is the result of template matching the output of EES.

3) We conducted experiments in simulated and real vehicle environments to validate the method's effectiveness. The experimental results show that the proposed EES outperforms existing methods and can effectively improve the performance of human-machine collaboration.

## II. RELATED WORK

### A. Human-Centered Following Methods

The core of human-centered following lies in real-time sensing of human position, speed, direction, and other dynamic characteristics. Ho et al. [4] developed a nonholonomic motion model using a Kinect camera, which estimates human motion direction, speed, and turning rate through Kalman filtering. To improve turning accuracy, Conte and Furukawa [1] proposed using head orientation as an implicit indicator of human intent, leveraging the rotational trend of the head before a change in direction. Nikdel et al. [6] integrated RGB-D and LiDAR measurements, dynamically accounted for the robot's localization uncertainty, and introduced a reliability verification mechanism to enhance accuracy.

The learning-based methods have shown good performance in human state estimation. Chongyu et al. [7] developed a CNN-LSTM hybrid network that captures gait features from thermal imaging data to predict human intentions. To enhance the accuracy of prediction, Wang et al. [8] described human movement by predicting human pose. Building on this, Mahdavian et al. [9] simultaneously predicted human trajectories and poses, with the two complementing each other to further improve prediction accuracy. However, the aforementioned methods focus solely on human movement. Jiang et al. [10] represented the environment as an occupancy map and incorporated environmental information to constrain human movement, thereby further enhancing prediction accuracy.

In summary, previous work primarily emphasizes predicting human movement, but our approach integrates environmental information to improve escorting performance.

### B. End-to-End Following Methods

The end-to-end approach uses a deep neural network to directly map environmental information to navigation, avoiding cumulative errors from optimizing multiple modules separately [11]. Pierre et al. [12] pioneered end-to-end mapping from images to control using a convolutional neural network, but their model lacked depth information, limiting its performance in obstacle avoidance and handling occlusions. Nikdel et al. [13] used deep reinforcement learning to output waypoints based on relative distance and angle, yet they underutilized environmental information. To address the critical challenge of obstacle and occlusion avoidance, Leisiazar et al. [14] represented the environment as an occupancy map and combined Monte Carlo Tree Search with deep reinforcement learning to generate more reliable navigation results. To further address the variability of human movement, they adopted a multimodal representation that provides probabilistic predictions, thereby enhancing decision-making [15].

In summary, previous research has mainly concentrated on fixed following patterns, which show limited adaptability in complex scenarios. The proposed escorting system can follow from various positions and can dynamically adjust as needed.

## III. METHOD

This section introduces the proposed end-to-end escorting system, as illustrated in Fig. 2. The model utilizes onboard

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

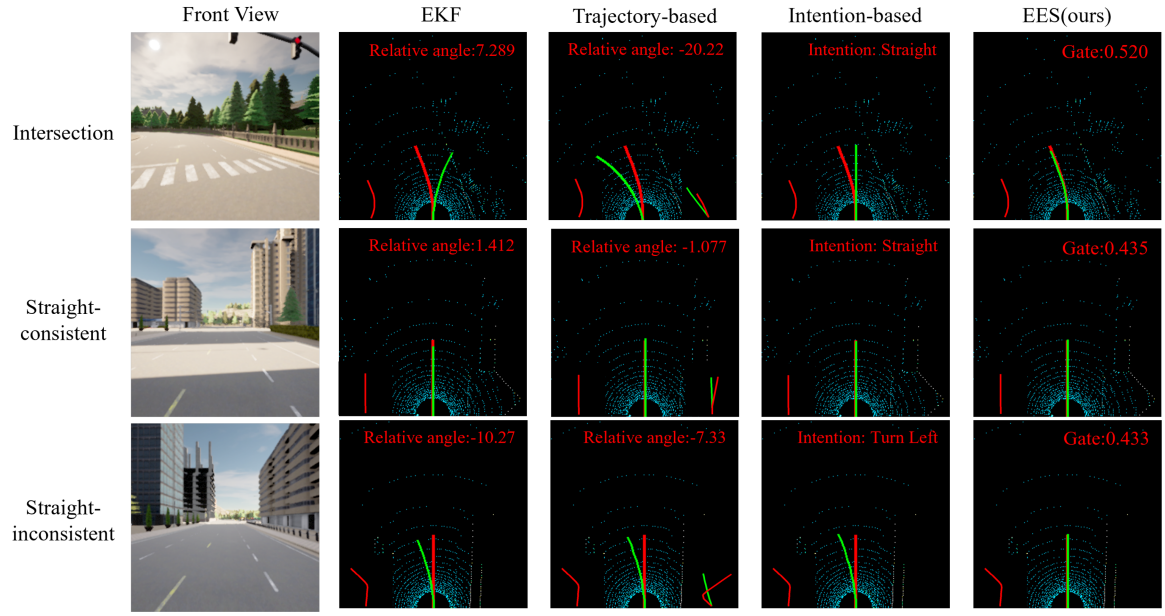



Fig. 3. **Results visualization.** We chose the typical scenes. Row 1: turn at intersection; Row 2: go straight (human and vehicle movement is consistent); and Row 3: human interference (human and vehicle movement is inconsistent). In the LiDAR image, the red line in the lower left represents the human historical trajectories, the red line in the center represents the ground truth of a real vehicle, and the green line represents the waypoints. In the trajectory-based method, the green and red colors in the lower right represent predicted human trajectories and ground truth, respectively.

TABLE II
ABLATION EXPERIMENTS RESULTS

| Exp | ADE ↓ | FDE ↓ | ADE_turn ↓ | FDE_turn ↓ |
|---|---|---|---|---|
| Intention-1 | 0.7865 | 1.2915 | 0.7725 | 1.2546 |
| | 1.9500 | 3.7675 | 1.9031 | 3.6833 |
| Intention-2 | 1.0874 | 1.8762 | 0.9926 | 1.7361 |
| | 1.1888 | 2.0580 | 1.0252 | 1.7973 |
| Gate-1 | 0.7463 | 1.2883 | 0.8786 | 1.5604 |
| Gate-2 | 0.8092 | 1.3365 | 1.0814 | 1.8990 |
| Gate-3 | 0.9306 | 1.6342 | 0.9732 | 1.7370 |
| Concat | 0.9763 | 1.1603 | 0.8869 | 1.5118 |
| EES(ours) | 0.7185 | 1.2193 | 0.8194 | 1.4062 |

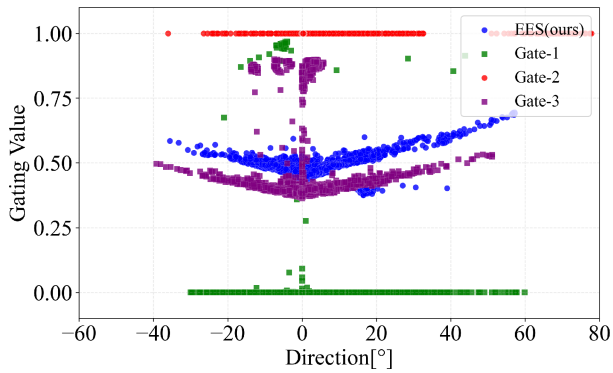



Fig. 4. **Visualization of gating value and vehicle motion direction.** As the Angle increases, the gating value of EES also increases accordingly. Gate-3 also shows a similar pattern, while Gate-1 and Gate-2 do not correlate significantly.

camera images, LiDAR point clouds, and human historical trajectories as inputs to generate waypoints.

### A. Problem Formulation

This work considers an autonomous vehicle tasked with escorting a human adaptively. The escort policy is defined by:

$$\begin{cases} x_r &= x_p + D(t)\cos\theta_r(t), \\ y_r &= y_p + D(t)\sin\theta_r(t), \end{cases} \tag{1}$$

where, the subscript $()_p$ and $()_r$ denote the human and vehicle. Here, $D(t)$ represents a time-varying distance that adjusts within a predefined range to enable adaptive escort. The angle $\theta_r(t)$ is determined by the human direction $\theta_p$, their relative positions $\theta_m$ (indicating whether the human is in front of, behind, to the side of the vehicle), and road constraints $\gamma(t)$. The road constraints is implicit expressed by the environmental information and the weighted parameter $\alpha(t)$ is learned from the training data.

$$\theta_r(t) = (1-\alpha(t))\gamma(t) + \alpha(t)(\theta_p) + \theta_m, \tag{2}$$

Unlike human-centered following models, where the vehicle is strictly determined by the human (formulated as $\theta_r(t) = \theta_p + \theta_m$). The goal of this letter is to develop a learning-based motion planning framework that effectively realizes this adaptive escort policy in dynamic environments.

### B. Perception Feature Extraction

To make the escorting system flexible, understanding the surrounding environment is crucial. We utilize camera images and LiDAR point clouds as inputs. Considering the complexity of the model and the limitations of the onboard computing platform, inspired by the Transfuser [16], we propose a cross-modal fusion framework based on the Transformer architecture. Our approach aims to capture comprehensive and multi-scale

TABLE III
COMPARISON OF CLOSED-LOOP RESULTS

| Scenario | Category | Model | Smoothness ↓ | | | Interventions ↓ | Distance | Angle |
|---|---|---|---|---|---|---|---|---|
| | | | Cur Std | Omega Mean | Omega Std | | | |
| CARLA | Front | EKF | 3.547 | 0.807 | 0.371 | 2 | $11.8 \pm 3.7$ | $5.5 \pm 6.3$ |
| | | Trajectory-based | 3.594 | 1.413 | 0.460 | 4 | $12.3 \pm 2.4$ | $4.1 \pm 5.6$ |
| | | Intention-based | 3.676 | 0.397 | 0.201 | 1 | $11.3 \pm 3.1$ | $8.3 \pm 9.0$ |
| | | EES(ours) | **1.847** | **0.287** | **0.069** | **0** | $11.0 \pm 3.1$ | $8.8 \pm 13.4$ |
| | Side | EKF | 3.205 | 0.566 | 0.338 | 6 | $4.1 \pm 2.1$ | $4.9 \pm 7.0$ |
| | | Trajectory-based | 5.698 | 1.366 | 0.349 | 2 | $8.0 \pm 2.0$ | $6.3 \pm 5.1$ |
| | | Intention-based | 4.973 | 0.446 | 0.264 | 5 | $6.7 \pm 3.8$ | $7.0 \pm 8.5$ |
| | | EES(ours) | **3.086** | **0.196** | **0.070** | **0** | $5.8 \pm 1.8$ | $7.4 \pm 7.6$ |
| | Back | EKF | 2.297 | 0.603 | 0.341 | 3 | $9.5 \pm 4.1$ | $8.2 \pm 11.7$ |
| | | Trajectory-based | 5.735 | 0.739 | 0.293 | 3 | $10.3 \pm 4.8$ | $9.9 \pm 16.1$ |
| | | Intention-based | 7.064 | 0.517 | 0.276 | 4 | $12.4 \pm 4.2$ | $13.7 \pm 13.7$ |
| | | EES(ours) | **1.564** | **0.084** | **0.051** | **0** | $6.8 \pm 2.6$ | $12.6 \pm 16.0$ |
| | S-movement | EKF | 3.656 | 0.609 | 0.297 | - | $6.8 \pm 1.8$ | $28.4 \pm 17.5$ |
| | | Trajectory-based | 4.816 | 0.444 | 0.154 | - | $11.8 \pm 3.6$ | $21.4 \pm 12.9$ |
| | | Intention-based | 3.797 | 0.289 | 0.176 | - | $9.6 \pm 2.5$ | $18.7 \pm 11.1$ |
| | | EES(ours) | **3.106** | **0.012** | **0.016** | **-** | $7.5 \pm 2.8$ | $28.4 \pm 17.5$ |
| Real-World | Scene-1 | EKF | 0.781 | 0.964 | 1.029 | - | $8.7 \pm 1.3$ | $24.0 \pm 35.4$ |
| | | Trajectory-based | 1.579 | 1.292 | 1.543 | 14 | $8.8 \pm 2.1$ | $26.1 \pm 37.8$ |
| | | Intention-based | 1.100 | 0.911 | 0.997 | 5 | $8.4 \pm 2.9$ | $15.9 \pm 31.3$ |
| | | EES (ours) | **0.516** | **0.691** | **0.997** | **0** | $8.4 \pm 2.9$ | $15.9 \pm 31.3$ |
| | Scene-2 | EKF | 0.700 | 0.832 | **0.716** | 1 | $9.4 \pm 2.1$ | $48.2 \pm 46.5$ |
| | | Trajectory-based | 0.395 | 0.863 | 1.072 | 1 | $10.4 \pm 1.2$ | $53.2 \pm 52.7$ |
| | | Intention-based | 0.333 | 1.378 | 1.430 | 1 | $8.2 \pm 1.7$ | $59.8 \pm 54.7$ |
| | | EES (ours) | **0.224** | **0.786** | 0.858 | **0** | $12.5 \pm 0.6$ | $39.6 \pm 41.4$ |

*Note:* EKF is the core method in [4] and [6]. Trajectory-based is inspired by [8], [9], and [10]. Intention-based is motivated by [1], [7] and [21].


(a) Scene-1 (b) Scene-2

Fig. 5. **The routes for the real-vehicle experiments.** To verify the generalization ability of the model, two scenarios were selected: the data collected from the blue areas was used for training, and the model was evaluated in the red.

features by integrating information from perspective views and bird's-eye view (BEV) across different resolutions.

Specifically, LiDAR point clouds within a 48m× 48 m frontal area are projected into a 256× 256 BEV pseudo-image. Hierarchical feature tensors $F_L^d$ are extracted through the ResNet backbone [17], where superscript $d$ denotes the stage-level depth index. Similarly, front-view camera images are processed by another ResNet [17], denoted as $F_I^d$. A transformer architecture fuses multi-scale global features from LiDAR and image modalities. At each hierarchical level, the BEV LiDAR features $F_L^d$ and front-view image features $F_I^d$ are fed into a self-attention module for interaction. The multi-scale feature fusion can be expressed as:

$$F_P^d = \text{Attention}(F_L^d, F_I^d), \tag{3}$$

where $F_P^d$ represents the perception features at hierarchical level $d$. After each self-attention module, the features are reshaped to match the input dimensions and added to the original inputs via residual connections. Finally, the perception features are passed through convolutional layers to produce feature vector $F_P \in \mathbb{R}^{1\times1\times256}$:

$$F_P = \text{ConvLayer}(F_P^d) \tag{4}$$

### C. Human Trajectory Feature Extraction

Human movement is too complex to be directly translated into vehicle navigation commands due to the lack of a one-to-one mapping between human intention and vehicle motion. Inspired by previous works [18], [19], we adopt a soft fusion strategy: the human trajectory features act as high-level navigation commands. In order to better derive the human intention from historical trajectories, we employ the Transformer module to extract rich features from the context and encode them into feature vectors.

Regarding the historical trajectories of humans, we incorporate twenty historical points into our analysis at time $t$, $J = \{(x_i, y_i) | i \in \mathbb{Z}, t-19 \leq i \leq t\}$. Initially, we use a multilayer perceptron (MLP) to embed these points into 256-dimensional vectors $E(J)$. Positional encoding $PE$ is added to $E(J)$ to incorporate temporal order information. Subsequently, we employ a self-attention mechanism to capture the temporal dependencies among the embedded features, resulting in the trajectory features $F_T \in \mathbb{R}^{1\times1\times256}$:

$$F_T = \text{Attention}(E(J) + P) \tag{5}$$

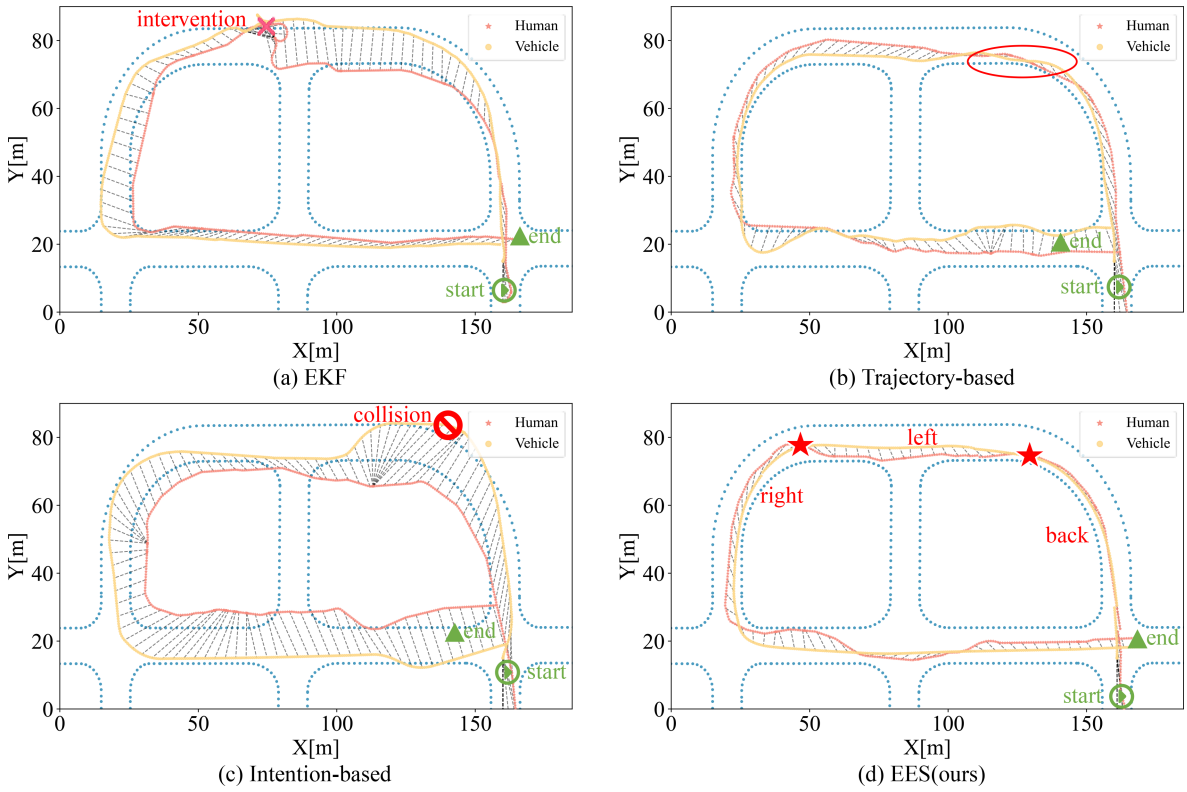



Fig. 6. **Results of the EES model in CARLA.** We dynamically adjusted the relative positions (front, back, left, right) between humans and vehicles, and the red stars in the figure indicate successful adjustment. Blue dots indicate road edges, prohibition symbols indicate vehicle-road collisions, and crosses indicate intervention. The straight line between a human and a vehicle indicates the relative positions of both at the same moment in time.

### D. Cross-Modal Attention Gating

The cross-modal attention gating module is designed to dynamically balance human historical trajectories and environment-aware features, making the escorting more adaptive.

The module takes perception features $F_P$ and trajectory features $F_T$ as inputs. First, multi-head attention computes queries from $F_T$ and keys/values from $F_P$, allowing $F_T$ to query trajectory-relevant regions in $F_P$ actively. The attention output is processed by a sigmoid function to generate channel-wise gating values $G \in \mathbb{R}^{1\times1\times256}$ :

$$G = \sigma\left(\text{MHSA}(Q = F_T, K = F_P, V = F_P)\right), \tag{6}$$

where $\sigma$ denotes the sigmoid function. Each element of the gating values $G$ corresponds to the weight of each channel. This gating mechanism amplifies perception regions critical to waypoints while suppressing conflicting or redundant features. Specifically:

$$F_{\text{fused}} = G \odot F_P + (1 - G) \odot F_T, \tag{7}$$

where $\odot$ denotes element-wise multiplication. The fused feature $F_{\text{fused}} \in \mathbb{R}^{1\times1\times256}$ explicitly models the interaction between perception and trajectory features.

Unlike the concatenate operation, where the perception features and human trajectory features are fixed after training and remain constant across all scenes, our gating structure dynamically adjusts the fusion weights based on the specific characteristics of each scene. This approach allows for a more adaptive and context-aware integration of the features.

### E. Waypoint Regression

Finally, a MLP decoder is used to predict waypoints in BEV space, represented as 2D coordinates. We use the L1 loss function to quantify the discrepancy between the ground truth and the predictions of the model.

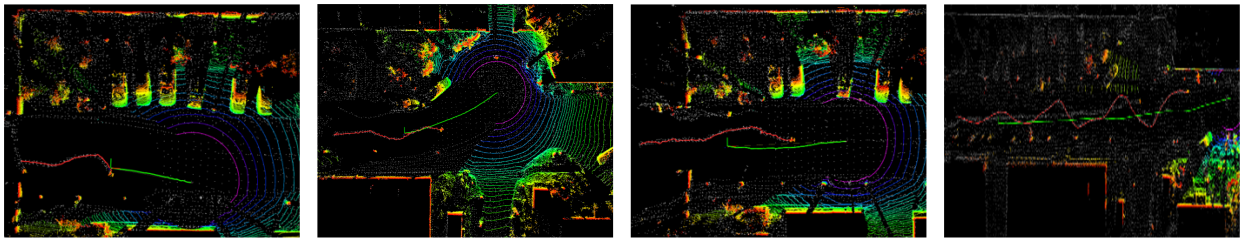
$$L = \frac{1}{N}\sum_{t=1}^{N} \|w_t - w_{gt,t}\|_1, \tag{8}$$

where $N$ denotes the predicted steps, $w_t \in W$ denotes the predictions of the model, and $w_{gt,t} \in W_{gt}$ denotes the ground truth.

## IV. Experiments

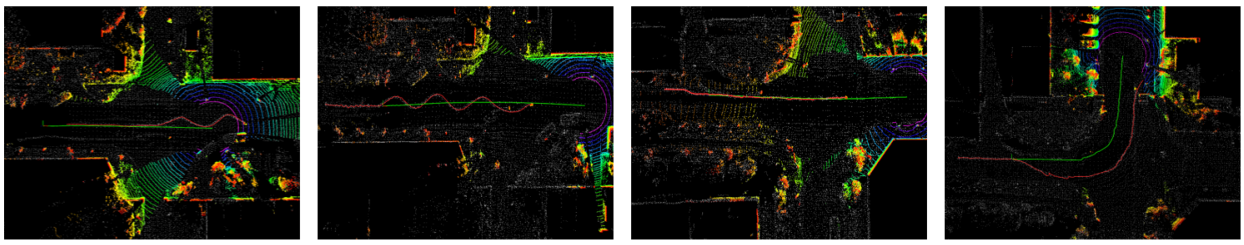
### A. Experimental Setup

*1) Datasets:* We employed CARLA [20] to manually drive a vehicle, collecting approximately 80,000 synchronized frames. Each frame includes LiDAR point clouds, a forward-facing RGB image, and the recent positions of humans as they moved in front of, behind, or alongside the vehicle. We designated towns 01, 02, 03, 04, 07, and 10 as our training set, with town 06 serving as



(a) EKF (b) Trajectory-based (c) Intention-based (d) EES(ours)

Fig. 7. **Comparison of experimental results in Scene-2.** We used Simultaneous Localization and Mapping (SLAM) to draw the environment as well as the trajectories. The red lines indicate the trajectories of humans, while the green lines indicate the trajectories of vehicles.



(a) Forwad (b) S-move/Backword (c) Go Straight at Intersection (d) Turn Left at the same Intersection

Fig. 8. **Results of the EES model in scene-2.** (a) and (b) show the following effects for humans at different locations. (c) and (d) show the results at the same intersection under different human guidance behaviors.

the validation set and town 05 being used for testing. For ground truth, we recorded the trajectories traversed by the vehicle.

*2) Closed-Loop Setup:* For closed-loop experiments, we utilized a single-object tracking algorithm to obtain the human's relative position with respect to the vehicle. Subsequently, we converted the historical position to the current coordinate via the vehicle wheel speed odometer to obtain trajectories. In the experiments, we manually controlled human motion and generated vehicle waypoints through the proposed EES. These waypoints were then executed using a pure-pursuit controller to facilitate low-level commands.

*3) Evaluation Metrics:* We evaluate the prediction errors on the dataset using two metrics: average displacement error (ADE) and final displacement error (FDE). ADE measures the mean displacement error across the entire predicted path, while FDE focuses on the displacement error at the final prediction points, calculated based on the last three guidance points of the path.

$$\begin{cases} \text{ADE} = \frac{1}{N}\sum_{t=1}^{N} \|w_t - w_{gt,t}\|_2 \\ \text{FDE} = \frac{1}{M}\sum_{t=N-M+1}^{N} \|w_t - w_{gt,t}\|_2, \end{cases} \tag{9}$$

where M represents the number of final points used to calculate the directional error.

In the closed-loop experiments, we use smoothness and intervention to quantify performance. In our experiments, vehicle motion is guided by human movement. As a result, the vehicle's trajectory fluctuates in response to changes in human movement. Given the difficulty of ensuring consistent human movement across experiments, we use a ratio to quantify the effect of human movement on the vehicle's trajectory fluctuations. Smoothness refers to the general trend of change, rather than emphasizing the characteristics of momentum.

$$S = \frac{\overline{\theta}_r}{\overline{\theta}_p}, \tag{10}$$

where $\overline{\theta}_r$ is characterized by metrics such as curvature standard deviation, average angular velocity, or standard deviation of angular velocity, and $\overline{\theta}_p$ is the corresponding indicator for human. Intervention refers to the occurrence of collisions. It is required when a vehicle collides with other traffic participants and stops with high collison risks.

*4) Baseline and Comparison Methods:* Existing research has primarily focused on predicting human states to provide guidance for vehicles. Due to differences in sensor setups and applications, we re-implemented several methods with necessary adaptations. (1) EKF-based method: Following [4], [6], we defined a four-dimensional state space. The current position information was used as input to predict the next position and velocity. Based on the predicted velocity and position, we generated waypoints for the upcoming period. (2) Trajectory-based method: Inspired by [8], [9], [10], we used historical trajectories as input, encoded them into feature vectors via a self-attention mechanism, and decoded these vectors to directly regress the future ten-step coordinates in an end-to-end manner. (3) Intention-based method: Following [1], [7], [21], we first clustered future trajectories into eight representative patterns and assigned an intention label to each pattern, which served as the ground truth. We then employed a neural network for intention prediction. The input data were encoded using a self-attention encoder, and the intention category was predicted.

For all methods, a unified template matching approach was used to convert the predicted human movement into vehicle motion commands. This ensured that the comparison focused

on the quality of the human state prediction and interpretation, rather than downstream planning differences.

### B. Comparative Experiments

We first tested on the dataset, with visualization results shown in Fig. 3. The three rows of results correspond respectively to the intersection scene, the straight road scene, and the straight road driving when the human movement is inconsistent with the vehicle. The EKF method struggles with predicting correct directions at intersections, resulting in collisions with the roadside. On straight roads, the system tends to follow human movement, even when they do not align with the vehicle's intended path, resulting in unintended turns. Similarly, both trajectory-based and intention-based methods are vulnerable to human movement, particularly when the human movement does not coincide with the vehicle motion. In contrast, our method demonstrates improved performance by utilizing perceptual information to make accurate determinations, even when there is inconsistency between human and vehicle movement. Moreover, it effectively utilizes human guidance to navigate turns at intersections. Table I shows the average ADE and FDE metrics for the entire and turning process. Our method achieves improvements of 41.2% in ADE and 52.0% in FDE for the entire process, and 54.5% in ADE and 63.2% in FDE for the turning process.

### C. Ablation Experiments

To further validate the effectiveness of our designed algorithm, we conducted a series of ablation experiments.

*1) Intention and Trajectory Feature Ablation:* We validated the effectiveness of using trajectory features in place of intention as advanced navigation targets. Our implementation included two scenarios: In Intention-1, the vehicle motion intention served as input to predict the waypoint, while in Intention-2, the human motion intention was used for the same purpose. Simultaneously, we estimated the scrambled intent inputs for testing. The results, presented in Table II , show that the ADE and FDE are higher for human intent inputs compared to vehicle intent inputs. Additionally, the increase in these metrics is less pronounced when incorrect intent is utilized. These findings suggest that human intentions do not perfectly align with vehicle motion, supporting our earlier analysis.

*2) Gating Structure Ablation:* This section is dedicated to evaluating the effectiveness of the gating network module. We conducted a series of experiments: The Gate-1 model incorporates a bidirectional attention mechanism, utilizing perception features and trajectory features as distinct queries. In contrast, the Gate-2 model implements hard attention through Gumbel-Softmax, resulting in binary gating values. Meanwhile, Gate-3 employs a straightforward linear layer to establish the gating structure, which is also compared against simple concatenation. The experimental results indicate that our model outperforms the others.

Furthermore, we analyze the relationship between the gating values and vehicle motion direction, as illustrated in Fig. 4. The blue dots clearly demonstrate a correlation between the gating values and the vehicle motion direction in our model. This suggests that the gating network structure enhances the model's ability to learn specific features that influence waypoint regression. Notably, despite Gate-3 showing weaker performance on certain metrics, it still displayed a similar pattern in this regard.

### D. Experiments in CARLA Simulation

We further conducted closed-loop experiments in CARLA simulation to validate the effectiveness of the algorithm. In town 05, we considered the following experimental scenarios, including humans moved in front of, behind, and alongside with the vehicle. We also tested the movement of vehicle when humans performed S-curve movements. The results of these experiments are summarized in Table III. Our proposed EES enables omnidirectional escorting, adapting to human movement in various directions while maintaining exceptional smoothness without any intervention. In our S-movement experiments, our method demonstrated the best performance in smoothness, highlighting the model's impressive ability to balance human movement with environmental information to achieve effective adaptive escorting.

We conducted a dynamic adjustment experiment to modify the relative positions of the human and the vehicle during the trial, with the results illustrated in Fig. 6. As seen in Fig. 6(d), our approach successfully adjusts the human position relative to the vehicle (transitioning from back to left and right) without significantly impacting the vehicle's motion. In contrast, methods like EKF face challenges in adjusting the lateral distance between a human and a vehicle. When a person shifts laterally away from or towards the vehicle, the vehicle tends to follow suit, risking a drift off the road. Even when adjustments are made successfully, the vehicle's motion remains affected under the trajectory-based approach, as evidenced in the red-circled area of Fig. 6(b).

### E. Experiments in Real Vehicle

We conducted real-world experiments on an unmanned vehicle, a Lynx hybrid 6x6 all-terrain assault vehicle equipped with a Hesai 80-beam LiDAR, a forward-looking camera, and an integrated inertial navigation system. Our model was optimized with NVIDIA TensorRT for accelerated inference and deployed on an onboard computer to process sensor data and output waypoints in real time.

To assess the generalization ability of the model, we trained it on the blue areas (as shown in Fig. 5) and tested it in the red areas without fine-tuning. As shown in Table III, our model outperforms algorithms such as EKF in real-world scenarios. Fig. 7 visualizes the actual testing process. Our model continues to function normally during S-movement, while algorithms like EKF exhibit instability and are prone to collisions. We further visualize the vehicle motion under different human guiding behaviors in Fig. 8. The vehicle can follow the human whether they are positioned in front of or behind it. In panel (b), we demonstrate that when the human performs an S-movement on a straight road, the vehicle maintains a straight trajectory. In panels (c) and (d), at the intersection, we observe that varying human guiding behaviors can lead to different vehicle motion, which aligns with our previous conclusions.

## V. Conclusion

In this letter, we present a data-driven end-to-end learning framework for escorting, which integrates environmental information and human historical trajectories. Our model enables human-following in front of, behind, and to the side of the vehicle, with dynamic switching between these positions, achieving adaptive escorting. We conduct experiments in simulation and real-world scenarios to demonstrate the effectiveness of our approach.

Although our model already fuses human trajectory features with environmental information, it still underutilizes richer human behavioral signals. Future work will integrate modalities such as hand gestures and full-body actions to deepen human–machine collaboration.

## References

[1] D. Conte and T. Furukawa, "Autonomous robotic escort incorporating motion prediction and human intention," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 3480–3486.

[2] R. Gockley, J. Forlizzi, and R. Simmons, "Natural person-following behavior for social robots," in *Proc. 2nd ACM/IEEE Int. Conf. Hum.-Robot Interaction*, 2007, pp. 17–24.

[3] E.-J. Jung, B.-J. Yi, and S. Yuta, "Control algorithms for a mobile robot tracking a human in front," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 2411–2416.

[4] D. M. Ho, J.-S. Hu, and J.-J. Wang, "Behavior control of the mobile robot for accompanying in front of a human," in *Proc. IEEE/ASME Int. Conf. Adv. Intell. Mechatron.*, 2012, pp. 377–382.

[5] L. G. Fletcher, P. Perali, A. Beathard, and J. M. O'Kane, "A visibility-based escort problem," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 4804–4811.

[6] P. Nikdel, R. Shrestha, and R. Vaughan, "The hands-free push-CART: Autonomous following in front by predicting user trajectory around obstacles," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 4548–4554.

[7] Z. Chongyu, G. Wenzhi, W. Rongwei, Z. Wang, and C. Wu, "Deep learning-driven front-following within close proximity: A hands-free control model on a smart walker," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 812–818.

[8] A. Wang, Y. Makino, and H. Shinoda, "Machine learning-based human-following system: Following the predicted position of a walking human," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 4502–4508.

[9] M. Mahdavian, P. Nikdel, M. TaherAhmadi, and M. Chen, "STPOTR: Simultaneous human trajectory and pose prediction using a non-autoregressive transformer for robot follow-ahead," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 9959–9965.

[10] Q. Jiang, B. Susam, J.-J. Chao, and V. Isler, "Map-aware human pose prediction for robot follow-ahead," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2024, pp. 13031–13038.

[11] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10164–10183, 2024.

[12] J. M. Pierre, "End-to-end deep learning for robotic following," in *Proc. 2nd Int. Conf. Mechatron. Syst. Control Eng.*, 2018, pp. 77–85.

[13] P. Nikdel, R. Vaughan, and M. Chen, "BGP:Learning based goal planning for autonomous following in front," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 3140–3146.

[14] S. Leisiazar, E. J. Park, A. Lim, and M. Chen, "An MCTS-DRL based obstacle and occlusion avoidance methodology in robotic follow-ahead applications," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 221–228.

[15] S. Leisiazar, S. R. R. Rohani, E. J. Park, A. Lim, and M. Chen, "Adapting to frequent human direction changes in autonomous frontal following robots," *IEEE Robot. Automat. Lett.*, vol. 10, no. 3, pp. 2934–2941, Mar. 2025.

[16] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "TransFuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12878–12895, Nov. 2023.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[18] W. Ou, T. Wu, J. Li, J. Xu, and B. Li, "RREV: A robust and reliable end-to-end visual navigation," *Drones*, vol. 6, no. 11, 2022, Art. no. 344.

[19] Y. Wang, Y. Sun, J. Li, and M. Shi, "Cross-modal fusion-based prior correction for road detection in off-road environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 12239–12246.

[20] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Conf. Robot Learn.*, 2017, pp. 1–16.

[21] G. P. Moustris and C. S. Tzafestas, "Intention-based front-following control for an intelligent robotic rollator in indoor environments," in *Proc. IEEE Symp. Ser. Comput. Intell.*, 2016, pp. 1–7.