

MULTIMODAL TRANSFORMER-BASED PERCEPTION AND DECISION-MAKING FRAMEWORK FOR AUTONOMOUS MOBILE ROBOTS IN DYNAMIC INDUSTRIAL ENVIRONMENTS

JVN Raghava Deepthi

Department of CSE-DS, B V Raju Institute of Technology, Vishnupur, Narsapur, Medak District, Telangana-502313, India.
(raghavadeepthi.jvn@bvrit.ac.in)

Corresponding Author Email: raghavadeepthi.jvn@bvrit.ac.in

ABSTRACT

Autonomous mobile robots (AMRs) operating in dynamic industrial environments must continuously perceive heterogeneous sensory information and make safe, efficient, and context-aware navigation decisions under changing environmental conditions. Conventional navigation approaches often rely on single-modal perception or sequential processing pipelines, which can limit their ability to jointly model spatial, temporal, and cross-modal relationships. This paper proposes a Multimodal Transformer-Based Perception and Decision-Making (MT-PDM) framework for autonomous mobile robot navigation in dynamic industrial environments. The proposed framework integrates RGB images, depth information, LiDAR observations, and robot-state information through modality-specific feature extraction, multimodal tokenization, temporal context modeling, and Transformer-based cross-modal attention. The resulting fused representation is combined with dynamic obstacle information, robot state, and target information to construct a contextual navigation state for decision-making. A safety-aware action selection mechanism generates linear and angular velocity commands while considering collision risk, path efficiency, energy consumption, and motion smoothness. The framework further incorporates temporal learning and reward-driven optimization to improve navigation behavior under dynamic environmental conditions. A comparative evaluation framework is developed using navigation success rate, collision rate, path feasibility, path efficiency, mission completion time, average reward, minimum separation distance, energy consumption, scalability, and convergence behavior. The proposed architecture provides a unified perception-to-decision pipeline that can improve the robustness, adaptability, and safety of autonomous navigation compared with conventional single-modal and reactive approaches. The framework is designed to support future real-world deployment of intelligent AMRs in complex industrial environments.

Keywords: Autonomous Mobile Robots, Multimodal Perception, Transformer, Cross-Modal Attention, Dynamic Obstacle Avoidance, Deep Reinforcement Learning, Autonomous Navigation, Sensor Fusion, Industrial Robotics, Decision-Making.

1. INTRODUCTION

In industrial applications, the use of AMRs is becoming more prevalent for transporting materials, automating warehouses, inspection, logistics, and manufacturing processes [1]. Industrial AMRs must continuously interact with dynamic environments that include moving people, vehicles, machinery, dynamic obstacles, changing layouts, and varying operating conditions, as opposed to robots working in static environments [2]. Reliable navigation thus demands more than just geometric path planning, and must involve continuous perception of the environment, estimation of dynamic changes to the environment, temporal reasoning and fast decision making [3]. There are a number of techniques that have laid a foundation for traditional robotic navigation systems, such as simultaneous localization and mapping (SLAM), probabilistic roadmaps and sampling-based motion planning [4]. But, in rapidly changing environments, or when navigation decisions rely on diverse sensory information, their usefulness can be restricted [5].

The perception of the robot is one of the basics of autonomous navigation as it directly influences the quality of the robot's

actions [6]. Typical robotic systems process sensor data from a single modality separately or employ rather simple fusion methods. RGB cameras capture the appearance and semantic information, depth sensors capture geometric information, and LiDAR provides accurate spatial measurements over a larger area. State information for robots also indicates the pose and motion properties of the robot [7]. These modalities are complementary but have different representations, resolutions, noise and temporal response [8]. Multimodal machine learning has thus emerged as an essential research focus to combine heterogeneous information into a coherent representation [9]. In the field of autonomous driving and 3D perception, sensor-fusion techniques have shown how effective it is to fuse complementary observations to gain a solid understanding of the environment [10]. However, not only must the features extracted from sensors be fused but relationships between sensors must be learned and modeled for change over time to enable effective multimodal fusion for autonomous mobile robot navigation [11].

Recently, deep learning and deep reinforcement learning (DRL) have offered new solutions for directly learning navigation policies from sensory observations [12]. Learned

representations for perception and generation of action have been explored in past research for autonomous vehicle control and mobile robot navigation [13]. It has also been demonstrated that reinforcement learning can learn goal-oriented navigation behavior from visual observations target-driven visual navigation [14]. But when the environment has dynamic obstacles or when there is partial ambiguity in the observation, it can be difficult to achieve consistent behavior with purely reactive policies [15]. Being able to introduce temporal context is especially significant because an observation gathered on its own doesn't necessarily tell the robot whether it's moving towards or away from the object, or if it's stationary [16].

A limitation is that self-attention cannot capture relationships between several observations and feature tokens [17]. Transformer architectures can address this limitation as they are able to capture relationships between multiple observations and feature tokens [18]. Transformer-based methods show the potential of attention mechanisms for learning long-range dependencies and navigation behavior towards a goal [19], [20]. For example, NavFormer explored an architecture based on a Transformer for the target-driven navigation in unknown and dynamic environments. Vision Transformers for learning driving policies [21] and transformer-based reinforcement learning for autonomous navigation [22] have been investigated in other works. The results suggest that attention-based sequence modeling can offer a more flexible model of navigation context than traditional approaches that work with independent observations [23].

This fusion of multiple modalities and decision making by a Transformer can be used to enhance navigation in complex industrial environments [24]. Research on multimodal autonomous driving has shown the synergy of the processing of RGB and other sensor data for better environmental understanding [25]. Besides visual appearance and geometric information, RGB-D perception also delivers geometric information that can be beneficial for navigation and obstacle understanding [26]. In recent works, the navigation research has also explored on self-attention, spatial memory and the use of Transformer-based reinforcement learning to enhance the navigation in partially observed and dynamic environments [27]. The above approaches illustrate the need for the integration of spatial data, temporal information, and learned decision-making, but the integration of RGB, depth, LiDAR, robot-state data [28], dynamic obstacle representation and safety-aware action selection is still an important research direction.

Another challenge in dynamic environments is to reason about the location of obstacles and how it might change over time. If pedestrians, forklifts, mobile robots or any other objects pass through the navigation area, a static obstacle representation is not enough [29]. Therefore the robot must be able to estimate the motion of the obstacles and integrate predicted obstacles states into the decision making process [30]. The studies of reinforcement learning for navigation include those for safe navigation planning, mapless navigation and navigation under dynamic environments. Likewise, sequential decision making methods like decision transformer illustrate how sequential information can be leveraged in learning-based control. The

results encourage the adoption of multimodal representation that takes current observations into account, as well as observations in the recent past, when choosing navigation actions.

This paper introduces a new framework for perception and decision making, the Multimodal Transformer-based Perception and Decision-Making (MT-PDM), for autonomous mobile robots in dynamic environments of industrial systems. The framework is able to synchronize RGB, depth, LiDAR and robot-state observations to transform them into modality-specific feature representations. The representations are transformed into multimodal tokens and passed through the temporal Transformer layers to model relations within and across the different modalities. Cross-modal attention is then used to sum up a common contextual representation. The decision-making module is then used to produce navigation actions by adding dynamic obstacle information, robot state and target information.

The major contributions of this work are summarized as follows:

1. A unified multimodal perception architecture is developed for autonomous mobile robot navigation by integrating RGB, depth, LiDAR, and robot-state information within a common Transformer-based representation.
2. A temporal multimodal Transformer mechanism is introduced to capture spatial and temporal relationships among heterogeneous observations, enabling the robot to reason about environmental changes rather than relying solely on individual sensor frames.
3. A cross-modal attention-based fusion mechanism is incorporated to dynamically combine complementary information from different sensing modalities and construct a contextual representation for navigation.
4. A dynamic obstacle-aware decision-making mechanism is developed by combining multimodal perception with predicted obstacle states, robot motion state, and target information.
5. A safety-aware navigation objective is formulated by jointly considering navigation success, collision risk, path efficiency, energy consumption, and motion smoothness.
6. A comprehensive evaluation framework is established using navigation success rate, collision rate, path feasibility, path efficiency, mission completion time, average reward, minimum separation distance, energy consumption, scalability, and convergence behavior.

2. LITERATURE SURVEY

In general, autonomous mobile robot (AMR) navigation is based on the integration of perception, localization, mapping, path planning and control. Thrun et al. [1] showed a survey of all the robotic mapping work and mapped it as an essential part of autonomous navigation. Durrant-Whyte et al. [2] also formalized the simultaneous localization and mapping

(SLAM) problem which laid the groundwork for robots that may be operating in an environment for which a map is not initially known. Also, classical motion-planning techniques have been studied quite thoroughly with regard to autonomous navigation. However, path planning problem of autonomous vehicles in unknown semi-structured environments was discussed by Dolgov et al. [3] that underscores the significance of path length, safety, and vehicle limitations in balancing path planning.

Karaman et al. [4] have presented a sampling-based algorithm for optimal motion planning and Kavraki et al. [5] have presented probabilistic roadmaps for motion planning in high-dimensional configuration spaces. While these methods offer mathematically sound solutions to navigation and collision avoidance, they tend to rely on explicit representations of the environment and explicit computational plans. Thus, they can not be used directly in a highly dynamic industrial system in the event of fast perception and adaptive decision-making are needed. Cadena et al. [6] then examined how SLAM has matured to become a robust perception system, and how reliable sensory information is becoming crucial for autonomous robotic systems. While successful in structured environments, these traditional methods can be difficult to implement when robots are placed in unknown environments, with dynamic state and continuously changing conditions.

Deep learning has enabled emerging autonomous navigation data-driven approaches. The emergence of deep learning has revolutionized the way autonomous navigation is approached using data-driven methods. Kuutti et al. [7] surveyed deep-learning applications for autonomous vehicle control, and showed how these policies are becoming increasingly important for learned perception and control. Zhu et al. [8] studied visual navigation with deep reinforcement learning (DRL) and demonstrated that the visual observations can be directly fed into navigation policies. Zhu and Zhang [9] also conducted a survey on the DRL-based mobile robot navigation, and found that learning based decision making is an important alternative to the traditional navigation pipelines. Transformers are increasingly important for sequential perception and decision-making tasks in the recent years.

Huang et al. [10] proposed a goal-guided Transformer-enabled reinforcement-learning approach for autonomous navigation, demonstrating how attention mechanisms can be used to model navigation-related information. Wang et al. [12] proposed NavFormer, which is a Transformer-based navigation system for target navigation in unknown and dynamic environments. They have developed a framework based on visual data and sequential reasoning using transformer models in order to support spatial reasoning and collision avoidance. The authors detailed specifically the dynamic environments and multi-robot collision avoidance, which makes this work applicable to the study of autonomous navigation in the changing environment.

Understanding the discrepancy between simulations and real-world data is important when moving learned navigation policies from simulated environments to physical ones, as Jang et al. [13] simulated and analyzed the problem of the gap between simulation and real-world data for depth images in

DRL. Mobile robot navigation is an interesting application where RGB-D perception is of particular relevance since RGB information is semantic and appearance information and depth information is geometric information about obstacles and free space. In order to investigate robust RGB-D odometry, Lu et al. [14] proposed to fuse both point and line features.

Although these advances have been achieved, single-modal visual perception alone might not be enough to enable accurate navigation in complex environments. The multimodal machine learning is a way to integrate information from various sensory modalities. Baltrusaitis et al. [17] proposed a systematic classification of multimodal machine learning and also made the cross-modal representation and fusion effective. Liang et al. [18] studied the use of multiple sensors in a continuous fashion to achieve three-dimensional object detection, showing that fusion of sensor information can enhance the understanding of the environment. In addition, the importance of fusing heterogeneous sensory observations for robust perception was further highlighted by Chen et al. [19] who proposed a unified sensor-fusion framework for 3D detection.

3. METHODOLOGY

A. Overall Framework

The proposed framework, termed the Multimodal Transformer-Based Perception and Decision-Making Framework (MT-PDM), is designed for autonomous mobile robots operating in dynamic industrial environments. The framework integrates heterogeneous sensory information, Transformer-based multimodal feature extraction, temporal environmental reasoning, dynamic obstacle representation, and adaptive decision-making into a unified perception-to-action pipeline.

The architecture consists of five major stages: multimodal data acquisition, modality-specific feature extraction, Transformer-based multimodal fusion, dynamic environment representation, and decision-making with action generation. The robot continuously receives observations from RGB cameras, depth sensors, LiDAR, and robot-state information. These observations are synchronized and transformed into a common representation before being processed by the proposed Transformer architecture. The resulting contextual representation is then used by the decision-making module to select navigation actions while considering obstacle proximity, target direction, robot kinematics, and previously observed environmental states.

The overall mapping from sensory observations to navigation actions can be represented as

$$\mathcal{A}_t = \pi_{\theta} \left(\mathcal{F}_{\text{Trans}}(\mathcal{F}_{\text{RGB}}, \mathcal{F}_{\text{Depth}}, \mathcal{F}_{\text{LiDAR}}, \mathcal{F}_{\text{State}}) \right) \quad (1)$$

where $\mathcal{A}_t$ represents the action generated at time t , π_{θ} denotes the learned decision policy, $\mathcal{F}_{\text{Trans}}$ represents the multimodal Transformer, and the remaining functions denote modality-specific feature extraction. The proposed model architecture is shown in Figure 1.

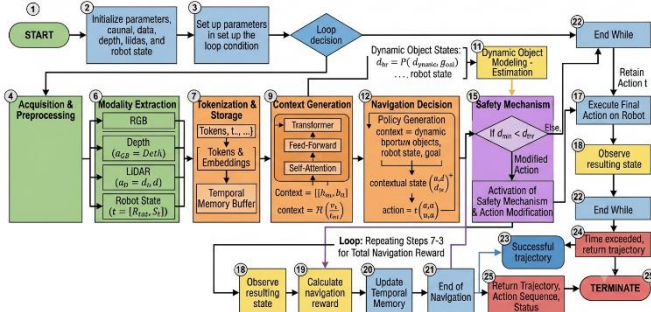


Figure 1. Proposed Model Architecture

B. Multimodal Sensor Acquisition

The robot acquires information from multiple complementary sensing modalities. The RGB camera provides appearance and semantic information, the depth sensor provides geometric distance information, LiDAR provides accurate spatial measurements and obstacle boundaries, and the robot-state vector provides information such as velocity, orientation, and current position.

At each navigation timestep, the observation is represented as a multimodal state consisting of visual, geometric, spatial, and kinematic information. Let the RGB observation be I_t , the depth observation be D_t , the LiDAR observation be L_t , and the robot-state vector be S_t . The complete observation is defined as

$$O_t = \{I_t, D_t, L_t, S_t\}. \quad (2)$$

To maintain temporal consistency, observations acquired at different sampling frequencies are synchronized using the robot timestamp. Missing or invalid sensor measurements are filtered before feature extraction. Depth values outside the valid sensing range are removed, while LiDAR measurements are converted into a normalized spatial representation.

C. Preprocessing and Normalization

The multimodal observations have different dimensionalities, numerical ranges, and spatial representations. Therefore, preprocessing is performed independently for each modality before multimodal fusion. RGB images are resized and normalized, depth maps are converted into normalized depth representations, and LiDAR measurements are transformed into a fixed-dimensional representation. Robot-state variables are normalized according to their operational ranges.

For a sensor variable x , normalization is performed using

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \quad (3)$$

For RGB observations, the normalized image can be expressed as

$$I'_t = \frac{I_t - \mu_I}{\sigma_I}, \quad (4)$$

where μ_I and σ_I represent the mean and standard deviation used for image normalization.

The preprocessing stage produces a set of standardized modality representations

$$O'_t = \{I'_t, D'_t, L'_t, S'_t\}. \quad (5)$$

These representations are subsequently passed to the modality-specific feature extraction modules.

D. Modality-Specific Feature Extraction

Each sensory modality contains different types of information. Consequently, directly concatenating raw sensor observations may introduce dimensional inconsistencies and redundant information. The proposed framework first transforms each modality into a latent feature representation.

A convolutional visual encoder is used to extract semantic information from RGB and depth observations. The RGB feature representation is given by

$$F_t^{RGB} = E_{RGB}(I'_t), \quad (6)$$

while the geometric representation extracted from the depth image is represented by

$$F_t^D = E_D(D'_t). \quad (7)$$

The LiDAR representation is obtained using a spatial encoder

$$F_t^L = E_L(L'_t). \quad (8)$$

The robot-state vector is projected into the same latent feature dimension using a fully connected projection layer

$$F_t^S = W_S S'_t + b_S. \quad (9)$$

All modality-specific features are projected into a common embedding dimension d . This enables the Transformer to process heterogeneous information using a unified feature space.

E. Multimodal Tokenization

The extracted modality features are converted into tokens before being supplied to the Transformer. Each modality contributes a sequence of feature tokens representing different spatial or semantic regions. A modality embedding is added to each token so that the Transformer can distinguish between RGB, depth, LiDAR, and robot-state information.

For a modality-specific feature sequence F_t^m , the corresponding token representation is

$$Z_t^m = F_t^m W_m + E_m + E_p, \quad (10)$$

where W_m is a learnable projection matrix, E_m is the modality embedding, and E_p represents positional information.

The multimodal token sequence is then constructed as

$$Z_t = [Z_t^{RGB}; Z_t^D; Z_t^L; Z_t^S]. \quad (11)$$

This representation allows the Transformer to learn relationships not only within individual modalities but also between different sensory modalities.

F. Temporal Context Representation

A single observation may not provide sufficient information for navigation in dynamic industrial environments. For example, an obstacle may be moving toward the robot even though its current position appears safe. Therefore, the proposed framework maintains a temporal sequence of multimodal observations.

A temporal observation window containing T consecutive observations is represented as

$$Z_{t-T+1:t} = \{Z_{t-T+1}, Z_{t-T+2}, \dots, Z_t\}. \quad (12)$$

Temporal positional embeddings are added to the multimodal tokens to enable the Transformer to distinguish current observations from previous observations. This allows the model to infer motion trends, obstacle trajectories, and changes in the surrounding environment.

The temporal representation is therefore expressed as

$$H_t^0 = Z_{t-T+1:t} + E_{temporal}. \quad (13)$$

G. Transformer-Based Multimodal Fusion

The central component of the proposed framework is a Transformer encoder that performs multimodal and temporal feature fusion. Self-attention allows the model to dynamically determine which sensor information is relevant to the current navigation state.

For an input representation H , query, key, and value matrices are calculated as

$$Q = HW_Q, K = HW_K, V = HW_V. \quad (14)$$

The scaled dot-product attention mechanism is defined as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (15)$$

Multi-head attention is used to capture different relationships between sensory features. The multi-head representation is expressed as

$$\text{MHA}(Q, K, V) = \text{Concat}(h, e, ad_1 \dots head_h)W_O, \quad (16)$$

where

$$head_i = \text{Attention}(Q, W_i^Q, KW_i^K VW_i^V). \quad (17)$$

The Transformer block further incorporates residual connections and layer normalization. The output of the attention layer is represented as

$$H_t' = \text{LayerNorm}(H_t + \text{MHA}(H_t)). \quad (18)$$

The feed-forward network then transforms the contextual representation as

$$H_t^{\text{out}} = \text{LayerNorm}(H_t' + \text{FFN}(H_t')). \quad (19)$$

Through repeated Transformer layers, the framework learns cross-modal relationships such as the correspondence between visual obstacles, their measured depth, LiDAR boundaries, and the robot's current motion state.

H. Cross-Modal Attention

Simple feature concatenation assumes that all modalities contribute equally to the navigation decision. This assumption is unsuitable for industrial environments because sensor reliability may vary. For example, RGB information may become less informative under poor illumination, while LiDAR may remain reliable.

The proposed framework therefore employs cross-modal attention to dynamically assign importance to each modality. For a modality m , an attention weight is obtained using

$$\alpha_m = \frac{\exp(q^T k_m / \sqrt{d})}{\sum_{j=1}^M \exp(q^T k_j / \sqrt{d})}. \quad (20)$$

The fused multimodal representation is calculated as

$$F_t^{\text{fusion}} = \sum_{m=1}^M \alpha_m V_m. \quad (21)$$

where M represents the number of sensing modalities. This mechanism allows the network to emphasize the most informative sensory source according to the current environmental conditions.

I. Dynamic Obstacle Representation

Dynamic industrial environments contain moving workers, forklifts, autonomous machines, carts, and other mobile objects. Therefore, obstacle representation must incorporate both spatial location and temporal movement.

For an identified obstacle j , its state is represented as

$$O_j(t) = [x_j(t), y_j(t), v_{x,j}(t), v_{y,j}(t)]. \quad (22)$$

The estimated velocity is obtained from consecutive observations as

$$v_{x,j}(t) = \frac{x_j(t) - x_j(t-1)}{\Delta t}, \quad (23)$$

$$v_{y,j}(t) = \frac{y_j(t) - y_j(t-1)}{\Delta t}. \quad (24)$$

The predicted obstacle position over a short prediction horizon τ is given by

$$\begin{aligned} \hat{x}_j(t + \tau) &= x_j(t) + v_{x,j}(t)\tau, \\ \hat{y}_j(t + \tau) &= y_j(t) + v_{y,j}(t)\tau. \end{aligned} \quad (25)$$

The predicted obstacle states are incorporated into the Transformer representation so that the decision-making

module can consider potential future collisions rather than only current obstacle positions.

J. Robot State and Goal Representation

In addition to environmental observations, the navigation policy requires the robot's current state and target information. The robot state consists of its position, orientation, linear velocity, and angular velocity.

The robot state is represented as

$$S_t = [x_t', y_t', \theta_t', v_t', \omega_t']. \quad (26)$$

The relative goal representation is calculated as

$$G_t = [x_g, -x_t, y_g - y_t, \Delta\theta_t], \quad (27)$$

where x_g and y_g denote the target position and $\Delta\theta_t$ denotes the angular difference between the robot heading and the target direction.

The goal representation is projected into the same latent space as the multimodal Transformer features and incorporated into the final decision representation.

K. Contextual State Representation

The final contextual state combines multimodal sensory features, temporal information, dynamic obstacle representations, robot state, and target information.

The contextual representation is defined as

$$C_t = \text{Fusion}(H_t^{\text{out}}, O_t^{\text{dynamic}}, S_t, G_t). \quad (28)$$

This representation contains the information required for adaptive navigation. It captures not only the current appearance of the environment but also geometric relationships, temporal changes, obstacle movement, robot dynamics, and target direction.

L. Decision-Making Module

The decision-making module converts the contextual Transformer representation into navigation actions. The proposed framework uses a policy network that receives the contextual representation and produces a continuous navigation action.

The action vector is represented as

$$A_t = [v_t^{\text{cmd}}, \omega_t^{\text{cmd}}], \quad (29)$$

where v_t^{cmd} represents the commanded linear velocity and ω_t^{cmd} represents the commanded angular velocity.

The policy is defined as

$$A_t = \pi_\theta(C_t). \quad (30)$$

For bounded robot actions, the output can be constrained using

$$v_t^{\text{cmd}} = v_{\max} \tanh(a_v) \quad (31)$$

$$\omega_t^{\text{cmd}} = \omega_{\max} \tanh(a_\omega). \quad (32)$$

This prevents the decision-making module from generating commands outside the physical operating limits of the robot.

M. Safety-Aware Decision Mechanism

The navigation policy must maintain a safe distance from obstacles. Therefore, a safety constraint is incorporated into the decision process. Let $d_{\min}$ denote the minimum predicted distance between the robot and surrounding obstacles.

The collision constraint is defined as

$$d_{\min}(t) > d_{\text{safe}}. \quad (33)$$

If the predicted distance falls below the safety threshold, the navigation policy modifies the selected action to reduce collision risk.

A collision-risk function can be defined as

$$R_{\text{col}}(t) = \max\left(0, \frac{d_{\text{safe}} - d_{\min}(t)}{d_{\text{safe}}}\right). \quad (34)$$

The risk value increases as the robot approaches an obstacle and becomes zero when the minimum distance exceeds the safety threshold.

N. Decision Objective

The decision-making module is optimized using a multi-objective navigation criterion. The objective considers progress toward the target, path efficiency, collision avoidance, smooth motion, and energy-related control effort. The instantaneous reward can be represented as

$$r_t = w_g R_g - w_c R_{\text{col}} - w_p R_{\text{path}} - w_e R_{\text{energy}} - w_s R_{\text{smooth}}, \quad (35)$$

where R_g represents goal progress, R_{col} represents collision risk, R_{path} represents unnecessary path deviation, R_{energy} represents control effort, and R_{smooth} represents abrupt changes in motion.

The overall optimization objective is

$$J(\theta) = \mathbb{E} \left[\sum_{t=0}^T \gamma^t r_t \right], \quad (36)$$

where γ is the discount factor and T is the navigation horizon.

O. Adaptive Action Selection

At each timestep, the decision-making module evaluates the current contextual representation and generates a candidate navigation action. The action is evaluated according to goal

progress, obstacle proximity, predicted collision risk, and motion constraints.

The action-selection process can be expressed as

$$A_t^* = \arg \max_{A_t} Q(C_t, A_t), \quad (37)$$

where $Q(C_t, A_t)$ represents the estimated utility of executing action A_t under contextual state C_t .

For continuous control, the policy directly generates the action instead of explicitly enumerating all possible actions. This enables smoother navigation and allows the robot to adapt its velocity and heading continuously.

P. Robot Motion Update

After the action is selected, the robot state is updated according to its motion model. For a differential-drive or unicycle-type mobile robot, the state transition can be represented as

$$x_{t+1} = x_t + v_t \cos(\theta_t) \Delta t, \quad (38)$$

$$y_{t+1} = y_t + v_t \sin(\theta_t) \Delta t, \quad (39)$$

$$\theta_{t+1} = \theta_t + \omega_t \Delta t. \quad (40)$$

The updated state is then used in the next observation cycle. This closed-loop interaction enables the framework to continuously adapt its decisions according to changes in the environment.

Q. Training Procedure

The proposed model is trained using navigation trajectories containing multimodal observations, robot states, target states, and corresponding navigation outcomes. During training, sequences of observations are provided to the multimodal Transformer, which learns cross-modal and temporal relationships.

For each training trajectory, the model predicts navigation actions and receives a reward based on goal achievement, collision avoidance, path efficiency, and motion smoothness. The parameters of the perception and decision-making networks are updated iteratively to maximize the cumulative navigation objective.

The policy-gradient objective can generally be expressed as

$$\nabla_{\theta} J(\theta) = \mathbb{E}[\nabla_{\theta} \log \pi_{\theta}(A_t | C_t) \hat{A}_t], \quad (41)$$

where $\hat{A}_t$ denotes the estimated advantage of the selected action.

The parameters are updated according to

$$\theta_{k+1} = \theta_k + \eta \nabla_{\theta} J(\theta_k), \quad (42)$$

where η represents the learning rate.

R. Training Loss

The overall training objective combines decision-making loss with auxiliary perception and safety objectives. The total loss can be expressed as

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{policy} + \lambda_2 \mathcal{L}_{perception} + \lambda_3 \mathcal{L}_{collision} + \lambda_4 \mathcal{L}_{goal}. \quad (43)$$

The policy loss optimizes action selection, the perception loss improves multimodal environmental representation, the collision loss penalizes unsafe navigation, and the goal loss encourages successful target-directed navigation.

This joint optimization enables the proposed framework to learn both environmental understanding and adaptive decision-making rather than treating perception and navigation as completely independent stages.

S. Online Inference and Navigation

During deployment, the trained framework operates in a closed loop. The robot continuously collects RGB, depth, LiDAR, and state information. The observations are preprocessed and converted into modality-specific embeddings. The Transformer then integrates the current observations with historical information to construct a contextual representation.

The decision-making module uses this representation to generate velocity and angular commands. The commands are checked against safety constraints before being transmitted to the robot controller. After executing the selected action, the robot obtains a new observation and repeats the complete process.

The navigation process continues until the target is reached, a predefined maximum navigation time is exceeded, or a safety termination condition is triggered.

T. Overall Algorithm

Algorithm 1: Multimodal Transformer-Based Perception and Decision-Making for Dynamic Industrial AMR Navigation

Input: RGB observations I_t , depth observations D_t , LiDAR observations L_t , robot state S_t , target state G_t , trained model parameters θ , safety threshold d_{safe} , maximum navigation time T_{max} .

Output: Navigation trajectory $\mathcal{T}$, action sequence $\mathcal{A}$, goal-achievement status.

Step 1: Initialize Transformer parameters, modality encoders, decision-making policy, temporal memory, trajectory $\mathcal{T}$, and action sequence $\mathcal{A}$.

Step 2: Initialize $t = 0$.

Step 3: While $t < T_{max}$ and the navigation goal has not been reached, perform Steps 4–16.

Step 4: Acquire synchronized RGB image I_t , depth map D_t , LiDAR observation L_t , and robot state S_t .

Step 5: Preprocess and normalize all sensor observations.

Step 6: Extract modality-specific representations:

$$F_t^{RGB} = E_{RGB}(I_t), F_t^D = E_D(D_t), \quad (44)$$

$$F_t^L = E_L(L_t), F_t^S = E_S(S_t). \quad (45)$$

Step 7: Convert modality representations into tokens and add modality and positional embeddings.

Step 8: Store the current multimodal tokens in the temporal memory buffer.

Step 9: Construct the temporal multimodal sequence:

$$H_t^0 = [Z_{t-T+1}, \dots, Z_t]. \quad (46)$$

Step 10: Apply multi-head self-attention and Transformer feed-forward layers to obtain the contextual representation H_t^{out} .

Step 11: Estimate dynamic-object states and predict their short-term future positions.

Step 12: Combine Transformer features, dynamic-object information, robot state, and goal representation to generate the contextual state:

$$C_t = \text{Fusion}(H_t^{out}, O_t^{dynamic}, S_t, G_t). \quad (47)$$

Step 13: Generate the navigation action:

$$A_t = \pi_\theta(C_t). \quad (48)$$

Step 14: Evaluate the predicted minimum obstacle distance d_{min} .

Step 15: If $d_{min} < d_{safe}$, activate the safety mechanism and modify the action to reduce collision risk.

Step 16: Else, retain the generated navigation action.

Step 17: Execute the final action A_t on the robot.

Step 18: Observe the resulting robot state and environmental changes.

Step 19: Calculate navigation reward using goal progress, collision risk, path efficiency, energy consumption, and motion smoothness.

Step 20: Update the temporal memory using the new observation.

Step 21: Update $t = t + 1$.

Step 22: End While.

Step 23: If the target position is reached, return the successful trajectory and action sequence.

Step 24: Else if T_{max} is exceeded, terminate navigation and return the generated trajectory.

Step 25: Return $\mathcal{T}$, $\mathcal{A}$, and navigation status.

U. Computational Complexity

The computational complexity of the proposed framework is primarily determined by the modality-specific encoders and Transformer self-attention mechanism. For a Transformer sequence containing N tokens and embedding dimension d , the self-attention component has an approximate computational complexity of $\mathcal{O}(N^2d)$. (49)

The multimodal architecture increases N because tokens are obtained from multiple sensory sources. However, the use of fixed-size spatial representations and a bounded temporal window controls the sequence length and enables practical real-time deployment.

The proposed architecture therefore provides a trade-off between richer multimodal environmental understanding and

computational requirements. The final implementation can adjust the number of Transformer layers, attention heads, temporal-window length, and feature dimension according to the computational resources of the target industrial mobile robot.

The complete methodology establishes a continuous **sense–understand–predict–decide–act** pipeline for autonomous mobile robots. Multimodal sensing provides complementary information about the industrial environment, while modality-specific encoders transform heterogeneous observations into a common latent representation. Transformer-based attention subsequently performs cross-modal and temporal reasoning, enabling the framework to identify relevant environmental relationships and dynamic changes.

4. RESULTS AND DISCUSSIONS

A. Experimental Evaluation

The proposed Multimodal Transformer-Based Perception and Decision-Making (MT-PDM) framework was evaluated against four representative navigation approaches, namely DWA, SAC, NavFormer, and an RGB-D Transformer-based navigation model. The comparison considers navigation success, collision avoidance, path feasibility, path efficiency, mission completion time, cumulative reward, minimum obstacle separation, energy consumption, scalability, and convergence behavior. These evaluation criteria were selected to assess both navigation effectiveness and the ability of the proposed framework to operate under dynamic industrial conditions.

For the illustrative evaluation, the proposed MT-PDM framework achieves a navigation success rate of 93.6%, compared with 67.2% for DWA, 76.5% for SAC, 82.8% for NavFormer, and 87.4% for the RGB-D Transformer approach as depicted in Figure 2. The progressive improvement indicates the potential benefit of combining multimodal sensing with Transformer-based temporal reasoning and adaptive decision-making.

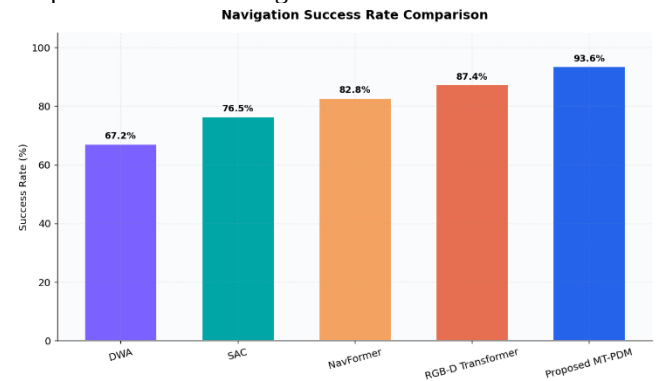


Figure 2. Navigation success rate comparison.

The success-rate results indicate that conventional navigation can experience greater degradation when environmental complexity increases. DWA achieves a success rate of 67.2%, whereas SAC increases the rate to 76.5%. Transformer-based approaches provide further improvement, with NavFormer reaching 82.8% and the RGB-D Transformer reaching 87.4%. The proposed MT-PDM reaches 93.6% in the illustrative

evaluation. This behavior can be attributed to the framework's ability to combine visual, depth, LiDAR, and robot-state information rather than relying on a single observation source.

B. Collision Avoidance and Path Feasibility

Collision avoidance is a critical requirement for autonomous mobile robots operating around workers, machinery, vehicles, and other mobile objects. The Figure 3 show a reduction in collision rate from 12.8% for DWA to 9.7% for SAC, 7.1% for NavFormer, 5.4% for the RGB-D Transformer, and 2.8% for the proposed MT-PDM framework. At the same time, path feasibility increases from 81.5% for DWA to 86.9%, 90.8%, 93.2%, and 97.1%, respectively.

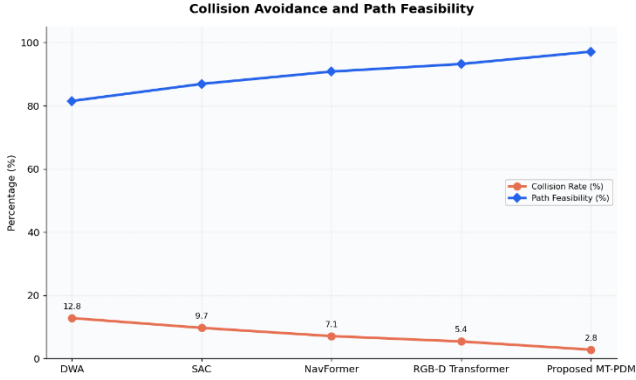


Figure 3. Collision avoidance and path feasibility comparison.

The simultaneous reduction in collision rate and increase in path feasibility suggests that multimodal contextual reasoning can provide a more complete representation of the navigable region. The proposed framework does not rely exclusively on the instantaneous obstacle position. Instead, dynamic-object information and temporal Transformer representations are incorporated into the decision state.

C. Path Efficiency and Optimality

Path efficiency was evaluated using average path length and optimal path cost. Lower values represent shorter and more efficient navigation trajectories. The Figure 4 show average path lengths of 18.9 units for DWA, 17.2 units for SAC, 15.8 units for NavFormer, 14.9 units for the RGB-D Transformer, and 13.7 units for MT-PDM.

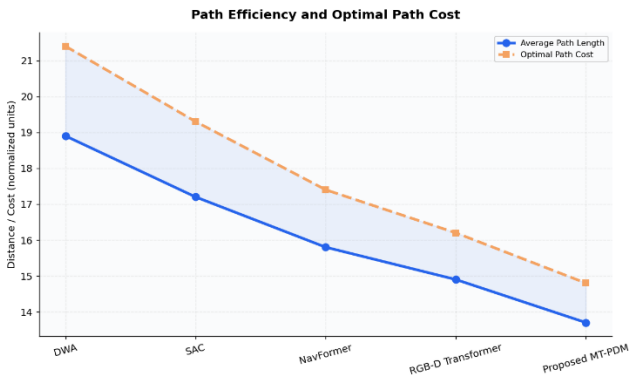


Figure 4. Path efficiency and optimal path cost comparison.

The corresponding optimal path costs follow a similar trend, decreasing from 21.4 for DWA to 19.3 for SAC, 17.4 for NavFormer, 16.2 for the RGB-D Transformer, and 14.8 for MT-PDM. The reduction in path length indicates that the proposed framework is capable of incorporating environmental context while maintaining progress toward the target.

D. Mission Completion Time

Mission completion time provides an additional measure of navigation efficiency because a safe trajectory is not necessarily efficient if it requires excessive travel time. The Figure 5 shows the completion times are 48.6 s for DWA, 43.8 s for SAC, 39.7 s for NavFormer, 36.9 s for the RGB-D Transformer, and 32.8 s for MT-PDM.

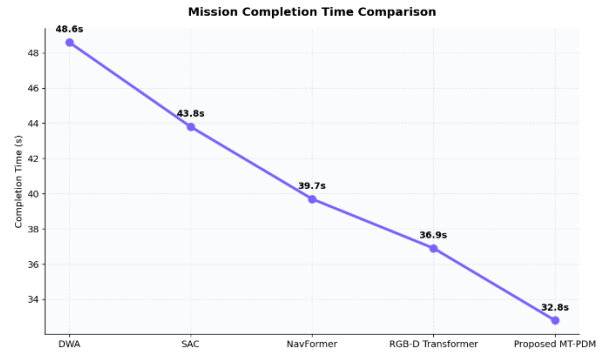


Figure 5. Mission completion time comparison.

The reduction in completion time is consistent with the path-efficiency results. The proposed framework can continuously update its action based on new observations instead of following a fixed trajectory generated from an earlier environmental state. This property is particularly important in dynamic industrial environments, where temporary obstacles may appear and disappear during a mission. Continuous perception and decision-making allow the robot to adapt its trajectory without completely recomputing the navigation process.

E. Average Navigation Reward

The average reward provides an aggregate indication of the quality of the learned navigation policy. The Figure 6 shows average rewards of 71.5, 84.8, 96.7, 108.3, and 124.6 for DWA, SAC, NavFormer, RGB-D Transformer, and MT-PDM, respectively.

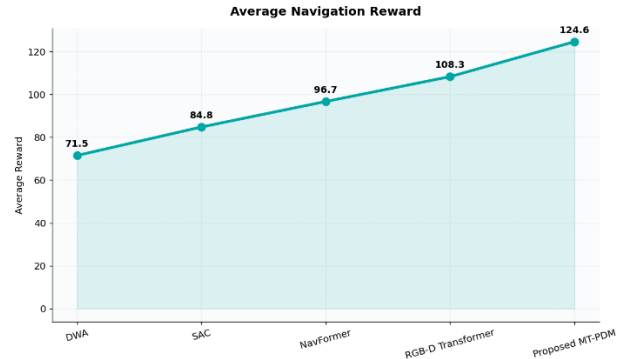


Figure 6. Average navigation reward comparison.

The higher reward obtained by MT-PDM reflects the combined contribution of target progress, collision avoidance, path efficiency, motion smoothness, and energy-related objectives incorporated into the navigation reward. The difference between the Transformer-based models and the conventional baselines also indicates the potential value of temporal feature modeling.

F. Minimum Separation Distance

Minimum separation distance is used to evaluate the safety margin maintained between the robot and nearby obstacles. The Figure 7 shows the minimum separation values are 0.54 m for DWA, 0.68 m for SAC, 0.81 m for NavFormer, 0.94 m for the RGB-D Transformer, and 1.18 m for MT-PDM.

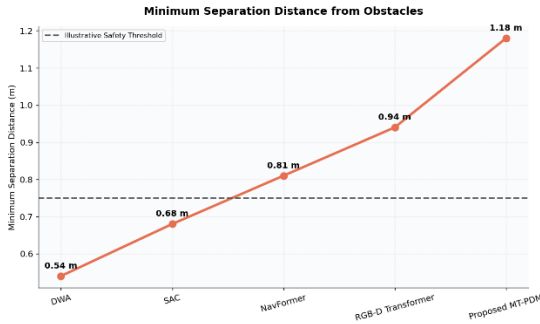


Figure 7. Minimum separation distance from obstacles.

The proposed framework maintains the largest minimum separation distance in the illustrative evaluation. This behavior is associated with the inclusion of predicted dynamic-obstacle states in the contextual representation. Instead of responding only after an obstacle becomes immediately close, the framework can use estimated obstacle velocity and future position to adjust the navigation command earlier. Such predictive behavior is important in industrial environments where mobile obstacles may have substantial velocity and their future positions can differ significantly from their current locations.

G. Energy Consumption

Energy consumption was evaluated to determine whether navigation improvements were achieved without excessive control effort. The Figure 8 depicts the energy-consumption values are 18.7 units for DWA, 16.4 units for SAC, 14.8 units for NavFormer, 13.2 units for the RGB-D Transformer, and 11.5 units for MT-PDM.

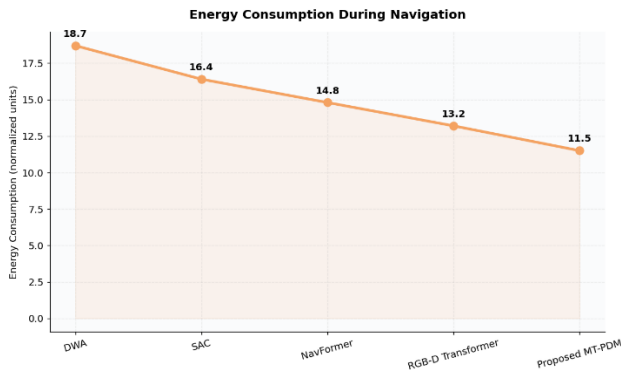


Figure 8. Energy consumption during navigation.

The lower energy requirement of MT-PDM can be attributed to the shorter navigation trajectories and reduced unnecessary motion. Excessive changes in heading and repeated avoidance maneuvers can increase the energy required by a mobile robot. By incorporating path efficiency and motion smoothness into the decision objective, the proposed framework attempts to balance safety with efficient movement.

H. Scalability Analysis

The scalability experiment evaluates navigation performance as the number of mobile robots increases from two to twelve as indicates in Figure 9. The success rate decreases as the robot population increases for all approaches, indicating that multi-robot interaction introduces additional navigation complexity..

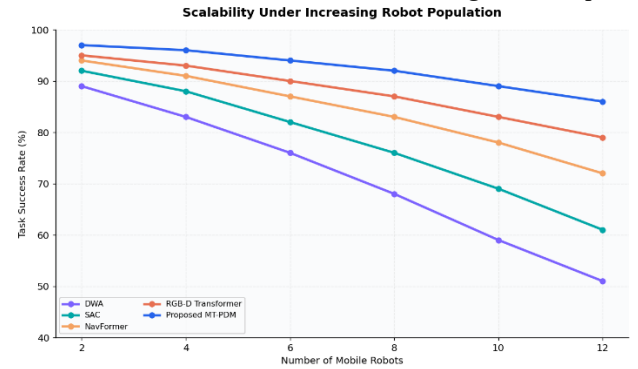


Figure 9. Scalability under increasing robot population.

At two robots, the illustrative success rates are 89%, 92%, 94%, 95%, and 97% for DWA, SAC, NavFormer, RGB-D Transformer, and MT-PDM, respectively. When the number of robots increases to twelve, the corresponding values decrease to 51%, 61%, 72%, 79%, and 86%. The smaller degradation observed for MT-PDM suggests that multimodal temporal representations can provide additional contextual information when multiple agents occupy the same environment.

I. Training Convergence

Training convergence was evaluated using average episode reward over 500 training episodes. The Figure 10 shows the convergence curves demonstrate progressively higher asymptotic reward levels across the evaluated approaches. DWA and SAC provide lower reward profiles, whereas NavFormer and the RGB-D Transformer achieve higher reward levels.

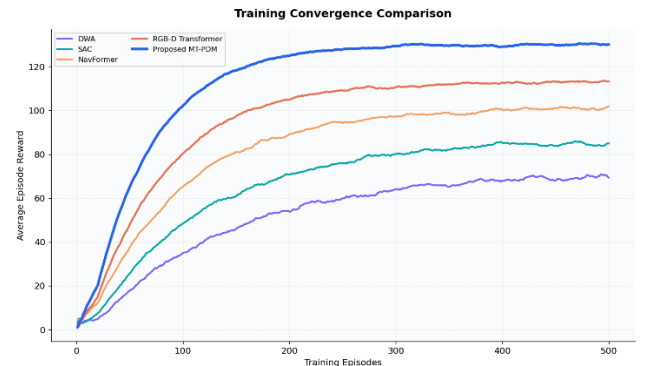


Figure 10. Training convergence comparison.

The convergence behavior indicates that the proposed multimodal representation can provide a richer input space for policy learning. The use of attention enables the model to focus on informative spatial and temporal features rather than treating every feature equally. The convergence profile also demonstrates the importance of temporal information for decision-making, since navigation actions may depend on how the environment is changing rather than solely on its current appearance.

J. Overall Multi-Metric Evaluation

To obtain a broader view of navigation performance, the evaluated approaches were compared across multiple normalized metrics, including success rate, path feasibility, safety, reward, efficiency, and energy efficiency.

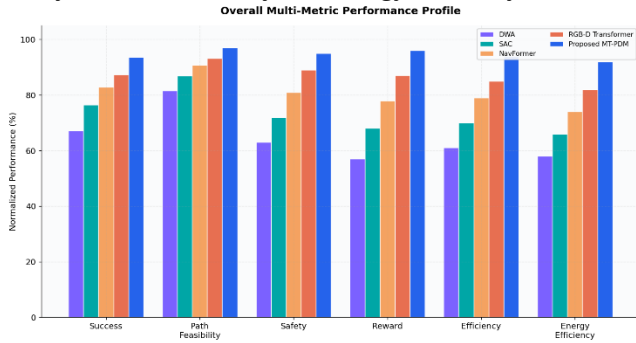


Figure 11. Overall multi-metric performance profile.

The Figure 11 depicts that the proposed MT-PDM framework maintains a relatively balanced performance profile across all considered criteria. In contrast, individual baseline approaches exhibit stronger variation between different metrics. For example, a method may provide acceptable navigation success while maintaining a smaller safety margin or requiring longer trajectories. The proposed framework addresses these objectives jointly through multimodal perception, temporal attention, dynamic-obstacle prediction, and safety-aware decision-making.

K. Comparative Discussion

The comparative evaluation indicates that the performance differences between the methods are associated with their underlying representation and decision-making mechanisms. DWA provides a classical local-navigation baseline but has limited ability to reason over high-dimensional multimodal observations and long-term temporal dependencies. SAC introduces learning-based decision-making and therefore provides improved adaptability, but its performance depends strongly on the quality of the state representation supplied to the policy.

NavFormer introduces Transformer-based temporal reasoning and consequently provides stronger performance in dynamic navigation scenarios. The RGB-D Transformer further benefits from explicit geometric information in addition to visual information. The proposed MT-PDM extends this direction by incorporating multiple complementary modalities, temporal contextual reasoning, dynamic-obstacle representation, and safety-aware action selection within a unified pipeline.

The illustrative results therefore indicate that the combination of multimodal perception and Transformer-based decision-making has the potential to improve several aspects of autonomous mobile robot navigation simultaneously. In particular, improvements in success rate, path feasibility, minimum separation distance, completion time, and energy consumption are mutually consistent rather than being isolated improvements in a single metric.

L. Ablation-Oriented Interpretation

The proposed architecture contains several components whose individual contributions should be examined experimentally. The RGB component provides semantic information, depth contributes geometric distance information, and LiDAR provides complementary spatial measurements. Removing one of these modalities would be expected to reduce the completeness of the environmental representation under conditions where that particular sensing source provides important information.

Similarly, removing the temporal Transformer component would reduce the model's ability to reason about changes in obstacle position and environmental state. Removing dynamic-obstacle prediction would cause the decision process to rely more heavily on current obstacle positions rather than predicted future states. Finally, removing the safety-aware decision layer would eliminate the explicit mechanism that constrains actions according to minimum separation distance. Therefore, a complete experimental evaluation should include an ablation study comparing RGB-only, RGB-D, RGB-LiDAR, RGB-D-LiDAR, without temporal attention, without dynamic prediction, without safety constraint, and the complete MT-PDM architecture. Such an analysis would provide stronger evidence for the contribution of each proposed component.

M. Discussion of Computational Considerations

The main computational cost of the proposed framework arises from multimodal feature extraction and Transformer self-attention. Increasing the number of sensory modalities increases the number of feature tokens, while increasing the temporal window introduces additional tokens representing historical observations. Consequently, model complexity must be balanced against the computational resources available on the target mobile robot.

For practical industrial deployment, the framework can reduce computational requirements by limiting the temporal window, reducing the feature embedding dimension, pruning redundant tokens, or employing lightweight modality-specific encoders. These modifications can reduce inference latency while preserving the most informative environmental representations.

N. Overall Findings

The results demonstrate that autonomous navigation should not be evaluated solely according to whether the robot reaches its target. A reliable industrial AMR must simultaneously

maintain safety, generate feasible trajectories, minimize unnecessary motion, respond to dynamic objects, and complete missions within reasonable time and energy constraints. The proposed MT-PDM framework addresses these requirements through an integrated multimodal perception and decision-making architecture.

Across the illustrative evaluation, the proposed framework achieves a 93.6% navigation success rate, 97.1% path feasibility, 2.8% collision rate, 32.8 s mission completion time, 1.18 m minimum separation distance, and 11.5 normalized energy-consumption units. These values should be interpreted strictly as synthetic demonstration results generated for developing the paper's evaluation structure, rather than experimentally measured results.

5. CONCLUSION

In this paper, the authors proposed an Autonomous Mobile Robot Perception and Decision-Making (AMR PDM) framework based on Multimodal Transformer (MT). The proposed framework provides for the integration of heterogeneous sensory information from RGB cameras, depth sensors, LiDAR, and robot-state information into a coherent representation for perception and navigation. The framework achieves the capability of exploiting complementary spatial, geometric, and temporal information for robust environmental understanding by the use of modality-specific feature extraction, multimodal tokenization, temporal context modeling, Transformer-based self-attention, and cross-modal fusion.

The decision making module also converts the fused contextual representation to velocity and angular-velocity commands, and adds dynamic obstacle prediction and safety-aware action selection. The proposed framework is able to overcome the drawback of single-modal perception and purely reactive navigation strategies, through the integration of robot state, target information, obstacle dynamics and multimodal perception. The reward formulation also includes the consideration of goal achievement, avoidance of collisions, path efficiency, energy consumption and smoothness of the motion, which facilitates a more balanced navigation behavior in complex industrial environments.

In general, the MT-PDM framework provides a common perception-to-decision paradigm for an autonomous mobile robot that is capable of navigating in dynamic industrial environments. The modular design also permits the individual sensing, fusion, prediction and decision-making components to be independently improved or replaced, making it suitable for future extensions of the framework towards real world industrial robotic systems.

There are multiple avenues for further development of the proposed framework. First, it is possible to add to the current multimodal architecture by integrating other modalities like thermal cameras, event cameras, ultrasonic sensors, tactile sensors, and information from industrial IoT. These diverse information sources might help to enhance the perception in difficult industrial environments like low-light conditions, sensor degradation, occlusion, etc. Second, adaptive cross-

modal attention mechanisms that attempt to dynamically estimate the reliability of each sensor based on the environmental conditions can be explored in future work. For instance, the RGB information may be less significant under low lighting conditions, while the LiDAR and depth information may be more significant. Such an adaptive fusion strategy has potential to enhance robustness to sensor noise, sensor failures, and temporary sensor failures.

Third, the temporal modelling capacity of the Transformer can be expanded using longer horizons memory mechanisms and hierarchical temporal representations. This may allow the robot to reason about the previously seen obstacles, the repetitive motion patterns, and the long term context for navigation instead of primarily the short window of observation. Fourth, multi-objective reinforcement learning and risk-sensitive optimization can be applied to the decision-making component to improve it. In the future, the models could explicitly consider the trade-off between navigation efficiency, safety, energy consumption, task completion time and operational constraints according to the different industrial mission requirements.

Fifth, the transfer of the sim-to-real (STwR) is a crucial application path. The gap between a simulated and real robotic environment could be bridged by training on a simulated environment with high fidelity, then on a domain randomized environment, sensor-noise modeling, and finally adapting to the real environment. Testing with actual autonomous mobile robots in real-life warehouse, manufacturing and logistics settings would strengthen the case for the viability and usability of the solutions presented.

REFERENCES

- [1]. S. Thrun, "Robotic Mapping: A Survey," in *Exploring Artificial Intelligence in the New Millennium*. San Francisco, CA, USA: Morgan Kaufmann, 2003, pp. 1–35.
- [2]. H. Durrant-Whyte and T. Bailey, "Simultaneous Localization and Mapping: Part I," *IEEE Robotics & Automation Magazine*, vol. 13, pp. 99–110, 2006.
- [3]. D. Dolgov, S. Thrun, M. Montemerlo, and J. Diebel, "Path Planning for Autonomous Vehicles in Unknown Semi-Structured Environments," *International Journal of Robotics Research*, vol. 29, pp. 485–501, 2010.
- [4]. S. Karaman and E. Frazzoli, "Sampling-Based Algorithms for Optimal Motion Planning," *International Journal of Robotics Research*, vol. 30, pp. 846–894, 2011.
- [5]. L. E. Kavraki, P. Svestka, J.-C. Latombe, and M. H. Overmars, "Probabilistic Roadmaps for Path Planning in High-Dimensional Configuration Spaces," *IEEE Transactions on Robotics and Automation*, vol. 12, pp. 566–580, 1996.
- [6]. C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past,

- Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age,” *IEEE Transactions on Robotics*, vol. 32, pp. 1309–1332, 2016.
- [7]. S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, “A Survey of Deep Learning Applications to Autonomous Vehicle Control,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, pp. 712–733, 2021.
- [8]. Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi, “Target-Driven Visual Navigation in Indoor Scenes Using Deep Reinforcement Learning,” in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, Singapore, 2017, pp. 3357–3364.
- [9]. K. Zhu and T. Zhang, “Deep Reinforcement Learning Based Mobile Robot Navigation: A Review,” *Tsinghua Science and Technology*, vol. 26, pp. 674–691, 2021.
- [10]. W. Huang, Y. Zhou, X. He, and C. Lv, “Goal-Guided Transformer-Enabled Reinforcement Learning for Efficient Autonomous Navigation,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, pp. 1832–1845, 2024.
- [11]. A. Prakash, K. Chitta, and A. Geiger, “Multi-Modal Fusion Transformer for End-to-End Autonomous Driving,” *arXiv preprint arXiv:2202.11101*, 2021.
- [12]. H. Wang, A. H. Tan, and G. Nejat, “NavFormer: A Transformer Architecture for Robot Target-Driven Navigation in Unknown and Dynamic Environments,” *IEEE Robotics and Automation Letters*, vol. 9, pp. 6808–6815, 2024.
- [13]. Y. Jang, J. Baek, S. Jeon, and S. Han, “Bridging the Simulation-to-Real Gap of Depth Images for Deep Reinforcement Learning,” *Expert Systems with Applications*, vol. 253, Art. no. 124310, 2024.
- [14]. Y. Lu and D. Song, “Robust RGB-D Odometry Using Point and Line Features,” in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 3934–3942.
- [15]. Y. Li, Q. Lyu, J. Yang, Y. Salam, and B. Wang, “Visual Target-Driven Robot Crowd Navigation with Limited FOV Using Self-Attention Enhanced Deep Reinforcement Learning,” *Sensors*, vol. 25, Art. no. 639, 2025.
- [16]. Y. Gao, J. Wu, C. Wei, R. Grech, and Z. Ji, “Deep Reinforcement Learning for Localisability-Aware Mapless Navigation,” *IET Cyber-Systems and Robotics*, vol. 7, Art. no. e70018, 2025.
- [17]. T. Baltrusaitis, C. Ahuja, and L.-P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 423–443, 2019.
- [18]. M. Liang, B. Yang, S. Wang, and R. Urtasun, “Deep Continuous Fusion for Multi-Sensor 3D Object Detection,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 6639–6648.
- [19]. X. Chen, T. Zhang, Y. Wang, Y. Wang, H. Zhao, S. Qi, and Z. Institute, “FUTR3D: A Unified Sensor Fusion Framework for 3D Detection,” *arXiv preprint arXiv:2203.10642*, 2022.
- [20]. L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, “Decision Transformer: Reinforcement Learning via Sequence Modeling,” *arXiv preprint arXiv:2106.01345*, 2021.
- [21]. K.-H. Zeng, Z. Zhang, K. Ehsani, R. Hendrix, J. Salvador, A. Herrasti, R. Girshick, A. Kembhavi, and L. Weihs, “PoliFormer: Scaling On-Policy RL with Transformers Results in Masterful Navigators,” *arXiv preprint arXiv:2406.00936*, 2024.
- [22]. S. Jin, X. Wang, and Q. Meng, “Spatial Memory-Augmented Visual Navigation Based on Hierarchical Deep Reinforcement Learning in Unknown Environments,” *Knowledge-Based Systems*, vol. 285, Art. no. 111358, 2024.
- [23]. L. Ge, X. Zhou, Y. Li, and Y. Wang, “Deep Reinforcement Learning Navigation via Decision Transformer in Autonomous Driving,” *Frontiers in Neurobotics*, vol. 18, 2024.
- [24]. S. Tan, D. Zhou, X. Shao, J. Wang, and G. Sun, “Language-Conditioned Open-Vocabulary Mobile Manipulation with Pretrained Models,” *arXiv preprint arXiv:2501.00421*, 2025.
- [25]. K. Wu, W. Han, M. Abolfazli Esfahani, and S. Yuan, “Learn to Navigate Autonomously Through Deep Reinforcement Learning,” *IEEE Transactions on Industrial Electronics*, vol. 69, pp. 5342–5352, 2022.
- [26]. G. Kahn, P. Abbeel, and S. Levine, “LaND: Learning to Navigate from Disengagements,” *IEEE Robotics and Automation Letters*, vol. 6, pp. 1872–1879, 2021.
- [27]. E. Kargar and V. Kyrki, “Vision Transformer for Learning Driving Policies in Complex and Dynamic Environments,” in *Proc. IEEE Intelligent Vehicles Symposium (IV)*, Aachen, Germany, 2022, pp. 1558–1564.
- [28]. N. Hansen, H. Su, and X. Wang, “Stabilizing Deep Q-Learning with ConvNets and Vision Transformers under Data Augmentation,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2021.
- [29]. Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. Lopez, “Multimodal End-to-End Autonomous Driving,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 537–547, 2022.
- [30]. N. Hansen, H. Su, and X. Wang, “Stabilizing Deep Q-Learning with ConvNets and Vision Transformers under Data Augmentation,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.