Narrate2Nav: Real-Time Visual Navigation with Implicit Language Reasoning in Human-Centric Environments

Amireza Payandeh, Anuj Pokhrel, Daeun Song, Marco Zampieri, and Xuesu Xiao

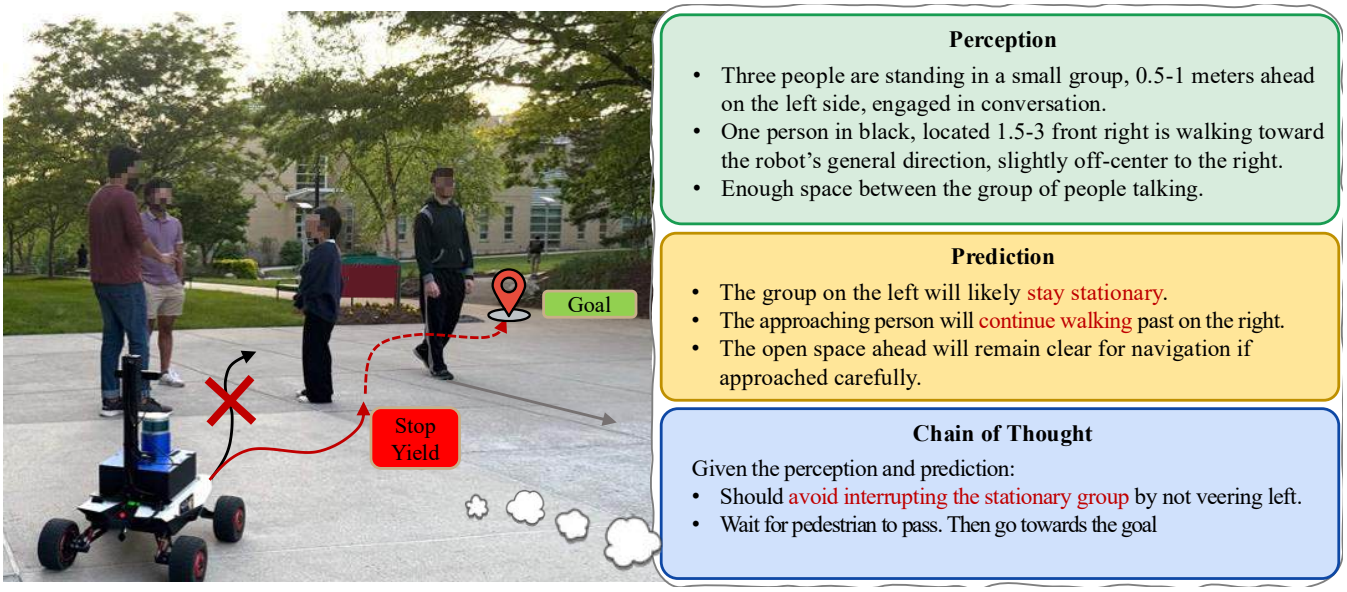

Fig. 1: NARRATE2NAV: Implicit natural language reasoning to navigate in human-centered environments while understanding the social cues and contextual information.

***Abstract*— Large Vision-Language Models (VLMs) have demonstrated potential in enhancing mobile robot navigation in human-centric environments by understanding contextual cues, human intentions, and social dynamics while exhibiting reasoning capabilities. However, their computational complexity and limited sensitivity to continuous numerical data impede real-time performance and precise motion control. To this end, we propose NARRATE2NAV, a novel real-time vision-action model that leverages a novel self-supervised learning framework based on the Barlow Twins redundancy reduction loss to embed implicit natural language reasoning, social cues, and human intentions within a visual encoder—enabling reasoning in the model's latent space rather than token space. The model combines RGB inputs, motion commands, and textual signals of scene context during training to bridge from robot observations to low-level motion commands for short-horizon point-goal navigation during deployment. Extensive evaluation of NARRATE2NAV across various challenging scenarios in both offline unseen dataset and real-world experiments demonstrates an overall improvement of 52.94% and 41.67%, respectively, over the next best baseline. Additionally, qualitative comparative analysis of NARRATE2NAV's visual encoder attention map against four other baselines demonstrates enhanced attention to navigation-critical scene elements, underscoring its effectiveness in human-centric navigation tasks.**

I. INTRODUCTION

A delivery robot moves through a festival, navigating a dynamic environment where a group of children crosses its path chasing a drifting balloon, while a couple ahead stops unexpectedly to take a selfie, and a few people are chatting with enough space between them for the robot to pass. Instead of colliding with the children, freezing in place, or using the path between the people having a conversation, the robot smoothly adjusts its path, yielding to the children and navigating around the photo-taking couple and the chatting group. This scenario highlights the need to enhance the navigation stack with higher-level reasoning capabilities that adhere to complex and subtle social rules, such as understanding why a human is moving in a certain way, when to yield, and who will cross the path. To enable such reasoning capabilities, robots need to accurately predict human intentions, interpret social cues, and employ human-like language reasoning to take socially aware actions.

Traditional methods often rely on rule-based systems or predefined path-planning algorithms, which struggle to adapt to dynamic, human-centric environments and typically lack the flexibility to incorporate social norms or interpret subtle contextual cues [1]–[3]. Learning-based approaches, such as imitation learning and reinforcement learning [4], [5], which learn from demonstration [6], [7] or trial and error [8], have shown promising results. However, these methods require collecting substantial amount of data. Additionally, they often fail to understand social cues and environmental context and do not achieve human-like language reasoning through pixel values.

With recent advancements in large Vision-Language Models (VLMs) demonstrating an understanding of contextual information, human intentions, and social cues, several research efforts have explored the integration of natural language to enhance robots' navigation performance using human-like language reasoning [9]–[12]. However, current approaches rely on large VLMs, whose heavy computational demands restrict their applicability in real-time scenarios. Additionally, due to their insensitivity to continuous numerical data, many existing VLM-based navigation frameworks prompt VLMs to generate high-level linguistic macro-actions, which are difficult to ground into real-world, low-level motion commands [9]. To this end, we ask the question: ***How can we leverage human-like language reasoning in mobile robot navigation while meeting real-time constraints?***

In this work, we present NARRATE2NAV a novel real-time vision-action model that implicitly integrates human-like language reasoning to bridge robot visual observations and low-level motion commands for short-horizon point-goal navigation in dynamic, crowded environments. Our approach embeds language-based reasoning, social cues, and contextual awareness into a visual encoder through a novel pre-training method using the Barlow Twins redundancy reduction loss [13]. Specifically, NARRATE2NAV enriches the current state representation with a multi-modal future state embedding that incorporates RGB visual inputs, low-level motion commands, and textual descriptions such as scene context, human intentions, trajectory summaries, and chain-of-thought (CoT) reasoning. This enables the model to explicitly leverage multi-modal signals during training, while relying only on RGB inputs with implicit language reasoning occurring in latent space rather than token space during navigation inference.

The summary of our contributions is as follows:

1. NARRATE2NAV, a novel real-time vision-action model that bridges the gap between visual observations and low-level motion commands, implicitly considering human-like language reasoning in human-centric environments.
2. A new Self-Supervised Learning (SSL) framework that refines the visual encoder's attention maps using textual signals to focus on image regions relevant to the navigation task and enhance its ability to operate effectively in dynamic environments.
3. Extensive analysis of state-of-the-art learning-based mobile robot navigation models across various challenging scenarios in human-centered environments, including both real-world experiments and offline-based evaluations, demonstrating a 52.94% and 41.67% improvement of NARRATE2NAV over the next best baseline, respectively.

II. RELATED WORK

Visual Navigation: GNM [14] combined various datasets with different robot embodiments to learn an omni-policy for visual navigation, mapping RGB inputs to low-level motion commands. ViNT [15] replaced the GNM model by introducing a Transformer-based, end-to-end behavioral cloning model trained on diverse real-world robot datasets to enable zero-shot generalization across different robot embodiments. NoMAD [16] introduced a unified diffusion policy that supported both goal-conditioned navigation and task-agnostic exploration through goal masking. CityWalker [17] proposed a pipeline to learn a navigation policy in dynamic urban environments using in-the-wild city walking video data. However, all these approaches focus solely on mapping visual observations to actions and fall short of embedding social context, human intentions, and human-like language reasoning essential for navigation in dynamic, human-centric environments, which is the gap NARRATE2NAV fills.

Representation Learning for Navigation: Eftekhar et al. [18] introduced a task-conditioned bottleneck using a small learnable codebook module to selectively filter point of interest in visual observations. VANP [19] and Vi-LAD [20] refined visual attention maps using action signals. CAHSOR [21] explored human preference learning and competence awareness for off-road navigation through the use of SSL. ViPlanner [22] introduced a semantic-aware, end-to-end local path planning framework trained entirely in simulation by fusing semantic segmentation with depth information to better assess terrain traversability and enable zero-shot transfer to real-world environments. NavFormer [23] presented a transformer-based policy with dual visual encoders for target-driven navigation in unfamiliar, dynamic settings, trained through cross-task learning for exploration and collision avoidance. NARRATE2NAV is based on the hypothesis that textual descriptions of navigation scenarios are another useful modality to inform navigation decisions, which none of the mentioned approaches utilized to train the visual encoder.

Language for Navigation: LeLaN [24] provides a dataset with language annotations derived from human walking and YouTube videos for the task of language-based object-goal navigation. In contrast, NARRATE2NAV uses textual descriptions as a representation learning signal rather than as explicit navigation goals. RING [25] trained a transformer-based model on large-scale simulation data for generalization to unseen embodiments for object goal navigation. Social-LLaVA [10] proposed using natural language to bring human-like language reasoning to mobile robot navigation in crowds, generating high-level navigation actions through CoT reasoning expressed in natural language. OLiViaNav [11] is the closest work to NARRATE2NAV and used GPT-4o to label the MuSoHu [6] dataset with textual social context for social robot navigation using a LiDAR and an RGB camera. In contrast, to our knowledge, NARRATE2NAV is the first RGB-only visual navigation model to learn 3D-to-2D structural features from weak textual supervision (e.g., "doorway on the right 2 m ahead") and to use chain-of-thought-based textual action descriptions for training in human-centric settings.

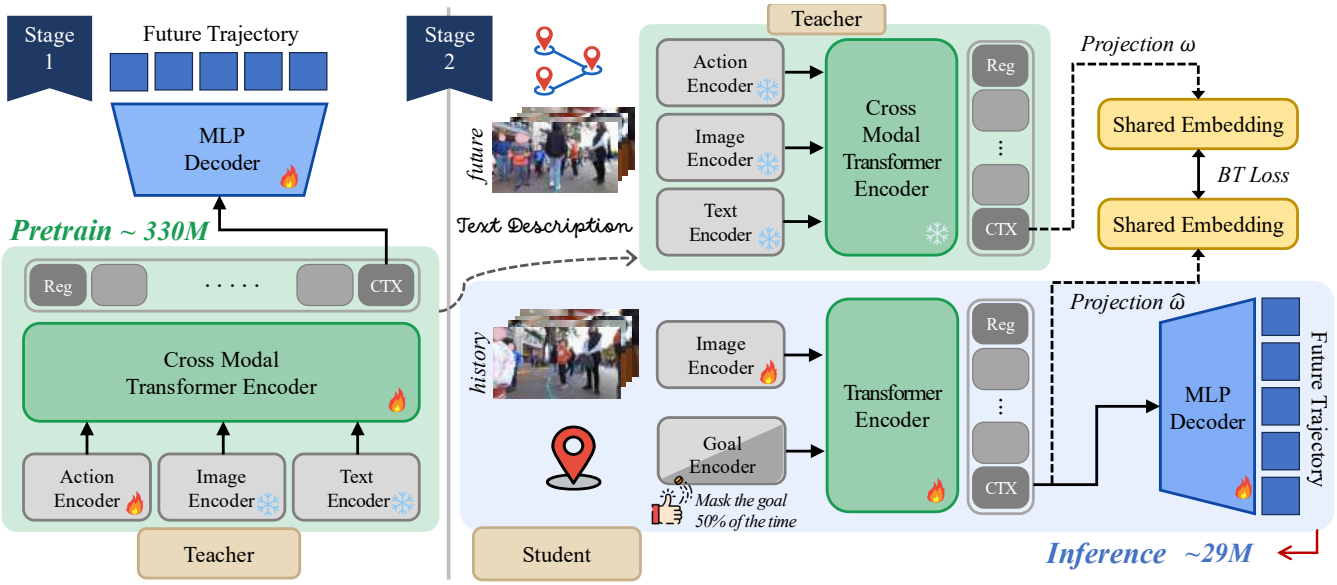

Fig. 2: NARRATE2NAV In Stage 1, a large Teacher Model (∼330M parameters) is pre-trained to predict trajectories by leveraging multi-modal inputs—low-level actions, observations, and textual signals—to embed a comprehensive understanding for each. In Stage 2, a lightweight Student Model (∼29M parameters) is trained using a redundancy-reducing Barlow Twins loss (BT Loss) to embed future state representations and implicitly incorporate human-like language reasoning into current state visual representations, utilizing only RGB history and goal information. Finally, the downstream task trains the Student Decoder to generate low-level motion commands.

III. NARRATE2NAV

We introduce NARRATE2NAV, a real-time vision-action model for robot navigation. Our approach leverages a novel training framework that integrates a unified multi-modal latent representation, learnable goal prediction, and implicit human-like language reasoning, enabling efficient navigation in dynamic, human-centric environments.

A. Problem Formulation

We formulate visual navigation as the task of driving a robot through a dynamic, human-centric environment using only RGB images from a front-facing camera, as explored in prior work [14], [19]. At each time step t, the robot receives a sequence of recent observations $o_t = \{I_{t-\tau}\}_{\tau=0}^{N-1} \in \mathcal{O}$ comprising RGB images I from the past N frames, and a goal position g specified in the robot's current 2D coordinate frame. The objective is to produce navigation action a defined as a sequence of future low-level motion commands (e.g., trajectory waypoints) for short-horizon navigation. Our objective is to train a policy $\pi_\theta(a \mid o, g) = P(a \mid o, g; \theta)$, parameterized by θ, that produces actions by conditioning on both observations and goals, generating low-level motion commands for visual navigation at each time step.

B. Architecture

The NARRATE2NAV architecture comprises a two-stage training pipeline: a Teacher Model and a Student Model (see Fig. 2).

In Stage 1, a large Teacher Model (∼330M parameters) is pre-trained to predict future trajectories by leveraging multi-modal inputs—visual history, low-level actions, and trajectory description from CoT textual reasoning. In Stage 2, a lightweight Student Model (∼29M parameters) is trained via a redundancy reduction Barlow Twin loss (BT Loss) to implicitly embed human-like language reasoning into visual representations, using only RGB history and goal information. At inference, NARRATE2NAV operates in real time using RGB inputs, bridging visual observations to low-level motion commands while implicitly incorporating language-driven reasoning.

1) Language as a modality: We employ natural language to endow NARRATE2NAV with human-like reasoning capabilities. To integrate this as a real-time module within a vision-language-action model, we treat language as a distinct modality for each state. Specifically, we generate natural language narrations describing the robot's egocentric view for each interaction, capturing its perception and predictions of surrounding human movements. These narrations serve as inputs for CoT reasoning to produce actions. Additionally, we use language as a weak signal to estimate distance measurements, replacing traditional LiDAR and 3D visual inputs with lightweight textual descriptions of spatial relationships. We also incorporate a general description of the goal position and the robot's future navigation plan. Specifically, At each step, we overlay waypoints for the next 2.5 seconds (5 frames at 2 Hz) onto the RGB image and prompt a VLM to describe the trajectory in relation to human positions within the scene. As highlighted by prior research [26], current SoTA VLMs exhibit limitations in spatial reasoning, a critical capability for navigation tasks. To address this, we evaluate multiple VLMs and select Qwen2-VL-72B-Instruct [27] for its superior performance in our context.

2) Stage 1: Teacher Model Training: NARRATE2NAV embeds a comprehensive understanding of each state, including

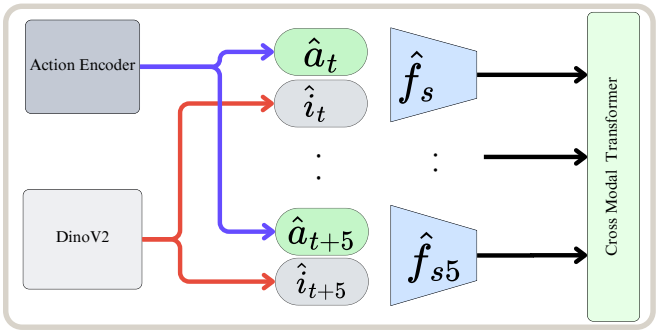

Fig. 3: Early fusion of action and observation embeddings, resulting in an effective embedding of each state.

the contextual information of the scene and the potential future movements of nearby humans, into a large Teacher Model (∼330M parameters), then to be infused into a smaller Student Model. The Teacher Model is trained in a supervised, end-to-end manner. Given n frames of context (referred to as future context in the stage 2), the n trajectories, and textual information about the scene, the model is tasked with generating the next n trajectories. The main goal is to learn an informative embedding for each state. Our objective is to learn a unified, modality-invariant state embedding that is maximally predictive of the future trajectory when conditioned on all available modalities (RGB, actions, and text). This embedding is then used as the distillation target for a lightweight student that operates using only RGB history and a goal at inference.

The Teacher Model comprises text, vision, and action encoders, whose output embeddings are used as input to a cross-modal transformer consisting of 6 layers, a hidden dimension of 256, and 8 attention heads.

Unified Multi-Modal Latent Representation: Inspired by Nazeri et al. [28], we employ early fusion of modalities (action, vision) for each state, subsequently passing them as unified tokens to the cross-modal transformer. Specifically, at each time step, given encoded RGB observations $o_t = \text{Seq}(I_{t+\tau_f} \rightarrow I_t)$ and corresponding encoded waypoint trajectories $a_t = \text{Seq}(a_{t+\tau_f} \rightarrow a_t)$ where $\tau_f \in [0, n]$, we compute a unified representation of the state as: $z_t = f_s(\hat{a}_t, \hat{i}_t)$ where f_s maps all embeddings to a shared latent space, mitigating statistical variance across modalities. As shown in Fig. 3, this alignment yields a uniform representation across modalities, streamlining multi-modal integration in the subsequent encoding layers. Our empirical analysis supports this architecture over interleaved (a_1, o_1, a_2, o_2, ..., a_n, o_n) or grouped (a_1, ..., a_n, o_1, ..., o_n) representation, which either do not converge or take significant computation to understand the relation of each observation to its corresponding action.

The Teacher Model uses a pretrained and frozen DINOv2 [29] as the vision encoder to extract visual features, while the 2D (X, Y) coordinates of the waypoints are projected into the embedding space using a linear transformation $\hat{a}_k = g(a_k) = W_a \cdot [X_k, Y_k]^T + b_a$, where a_k denotes the k-th waypoint in the trajectory sequence. For the text input, we use the pretrained CLIP [30] text encoder $\hat{t} = f_{\text{CLIP}}(T)$, $\hat{t} \in \mathbb{R}^{d_t}$, where d_t is the dimension of the text embedding. All embeddings are projected to a 256-dimensional space. As the focus is on learning non-fixed, relative relationship patterns among RGB, text, and waypoints in short-length sequences (non-fixed as text does not always refer to a specific frame), we choose learned positional encodings to adaptively capture these dynamic, task-specific interactions, rather than the well-established sinusoidal encodings used in natural language processing for absolute positional information. Therefore, we encode temporal state using a learnable positional encoding $P \in \mathbb{R}^{(1+\tau_f+2)\times 256}$ in the input sequence. The final transformer input, concatenated with a register token and context token, is:

$$Z_{\text{final}} = [\text{reg}, z_t, \ldots, z_{t+\tau_f}, \hat{t}, \text{ctx}] + P,$$
$$Z_{\text{final}} \in \mathbb{R}^{(1+\tau_f+2)\times 256}.$$

Finally, the Teacher Model (Fig. 2, left model), with a fully connected MLP decoder, is trained end-to-end to predict $\text{Seq}(a_t \rightarrow a_{t+\tau_f})$.

3) Stage 2: Main (Student) Model Training: As we define our task on RGB-only input, for the main model (only the main model is used at inference time), we pass only the sequence of RGB $o_t = \text{Seq}(I_{t+\tau_p} \rightarrow I_t)$ where $\tau_p \in [-n, 0]$ and one single point as the goal in 2D (X, Y) coordinates. Similar to the Teacher Model the learnable positional encoding is attached to the input. The final input to the transformer is:

$$Z_{\text{final}} = [\text{reg}, z_{t-\tau_P}, \ldots, z_t, \hat{g}, \text{ctx}] + P,$$
$$Z_{\text{final}} \in \mathbb{R}^{(1+\tau_P+2)\times 256}.$$

The Student Model uses Resnet50 [31] architecture as the vision encoder and encodes the 2D (X, Y) goal using a linear transformation.

Goal Infilling: To develop a unified policy supporting both goal-directed navigation and undirected exploration, we use a stochastic, learnable masked modeling approach (Fig. 4). In this framework, the goal is replaced with a learnable embedding for 50% of the training instances. Therefore, there are two learned embeddings for the goal: one for when the goal is explicitly provided, guiding task-oriented behavior, and another for when the goal is masked, promoting task-agnostic exploration. This method enables joint training of task-agnostic and task-oriented behaviors, effectively mitigating challenges posed by the absence of explicit goals [15], [28]. Once the supervised end-to-end training of the Teacher Model is complete, we freeze all its modules and remove its decoder. We then distill its embeddings into the main (student) model. Specifically, at each step of training, the inputs passed to the Teacher and Student Models are $\text{Seq}(I_{t+1} \rightarrow I_{t+\tau_f})$ and $\text{Seq}(I_{t-\tau_P} \rightarrow I_t)$ respectively, where τ_P denotes the past frames (e.g., $[I_{-5}, I_{-4}, \ldots, I_0]$) and τ_f denotes the future frames (e.g., $[I_1, I_2, \ldots, I_5]$). Typically, we set $\tau_P = \tau_f = 5$ and predict the corresponding sequence of actions for the next five time steps. NARRATE2NAV maximizes the shared information between the past-to-current state and future states by applying

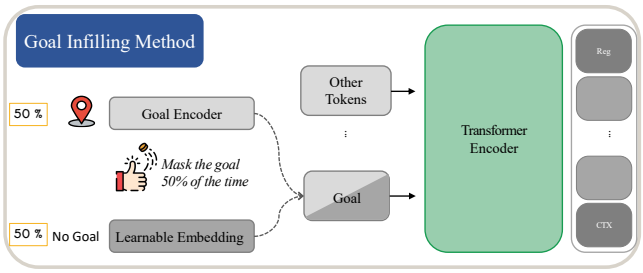

Fig. 4: Goal infilling strategy in NARRATE2NAV.

Barlow Twins [13] loss on the projected context tokens (ctx) of the two models' transformers, thereby aligning past observations with the future multi-modal feature space. In contrast to vision SSL models that rely on joint embeddings of augmented images, NARRATE2NAV aligns the future multi-modal feature space—encompassing action latent space $\mathcal{A}$, text latent space $\mathcal{T}$, and pixel latent space $\mathcal{O}$—with the latent space of past pixels. This alignment is enforced through the Barlow Twins loss, defined as:

$$\mathcal{L}_{\text{BT}} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2,$$

where $\mathcal{C} = \frac{Z_{\text{past}}^\top Z_{\text{future}}}{\|Z_{\text{past}}\| \|Z_{\text{future}}\|}$ is the cross-correlation matrix computed between the normalized embeddings from past pixel features Z_{past} and future multi-modal features Z_{future}, and λ is a weighting factor that balances the invariance and redundancy reduction terms. After completing the pretraining, we add the decoder and train the model end-to-end using Mean Squared Error (MSE) between the model's predictions and the corresponding ground truth value.

C. Implementation

We utilize a selected subset of the SCAND [32] dataset, gathered from varied, densely populated public environments with intricate human-robot interaction scenarios. Additionally, we collect another dataset to ensure a robust evaluation on out-of-distribution, unseen data, providing a fair comparison across all baselines and our model. We collect data using an AgileX Scout Mini robot equipped with a ZED2 stereo camera, leveraging visual odometry to estimate the robot's position and orientation relative to its own body frame. The robot is driven at linear velocities ranging from 0 to 1.6 m/s and angular velocities ranging from -1.5 to 1.5 rad/s, consistent with the velocity ranges of the SCAND [32] dataset. The collected dataset encompasses a variety of complex social scenarios, including Navigation in Crowds, Frontal Approach, Human Following, Narrow Passageway, and Intersection. [33]. We train our model on 8 × A100 GPUs (40 GB memory per GPU) using the AdamW optimizer with a learning rate of 2e-4 for 267 epochs in stage one, and a total (pretraining + downstream task) of 747 epochs in stage two. During deployment, the model takes as input the goal coordinates relative to the robot's frame, along with five past images encompassing the most recent 2.5 seconds of robot history. The output trajectory predicted by the model is tracked using a Pure Pursuit controller with a dynamic lookahead distance of at-least 0.2m in front of the robot.

IV. ANALYSIS

In this section, we discuss comparative evaluations and ablation studies of our approach. The experiments are designed to highlight the advantages introduced by our proposed innovations.

A. Quantitative Analysis

We evaluate NARRATE2NAV in five distinct, challenging social scenarios as described by Francis et al. [1], encompassing both indoor and outdoor environments. We compare our approach with previous SoTA baseline models, including GNM [14], ViNT [15], NoMaD [16], and CityWalker [17]. We acknowledge OLiVia-Nav [9] as closest approach to our work; however, due to the unavailability of its open-source code, implementation details, and prompt design, we are unable to reproduce its results for a direct comparison. Table I presents the comparative performance of NARRATE2NAV against other methods, both when a goal is provided and when it is not provided in the unseen offline dataset. Notably, GNM relies on a goal and is therefore excluded from this experiment.

Note that the ALL column does not represent an average of the other columns; instead, it is an evaluation of a broader set of unlabeled samples along with the scenario-specific samples, providing a holistic assessment of prediction accuracy throughout the sequence. *Final Displacement Error* (FDE) focuses on the spatial deviation at the final time step. Additionally, we use the *Average Orientation Error* (AOE) as defined by Liu et al. [17], which measures the angular difference between the predicted and ground truth trajectory vectors.

$$\text{AOE}(k) = \frac{1}{n} \sum_{i=1}^{n} \theta_i^k = \frac{1}{n} \sum_{i=1}^{n} \arccos \left(\frac{\langle \hat{a}_i^k, a_i^k \rangle}{\|\hat{a}_i^k\| \|a_i^k\|} \right)$$

where $\hat{a}_i^k$ and a_i^k denote the predicted and ground truth orientation vectors at time step i for the k-th sample, respectively. Our analysis shows that NARRATE2NAV outperforms the baselines by 52.94% over all the scenarios, thanks to the enriched visual embedding with social context, implicit human-like language reasoning, and pretext training method.

B. Qualitative Analysis

We present a qualitative analysis of the learned activation maps and predicted trajectories. To visualize the attention maps, we apply the Jet colormap, with red denoting areas of highest attention, on the last layer of each model's vision encoder. Since CityWalker uses a pretrained, frozen vision encoder, we do not show it here.

As shown in Fig. 5, the NARRATE2NAV attention map highlights the two people crossing its path, focusing on features relevant to navigation, while the attention maps of NoMAD [16], GNM [14], and ViNT [15] (tagged as 2, 3, and 4, respectively) are less interpretable, often highlighting

TABLE I: Comparison of NARRATE2NAV with four SoTA Methods on the Unseen Dataset.

	Method	Navigation in Crowds			Frontal Approach			Human Following			Narrow Passageway			Intersection			All		
		↓AOE	↓ADE	↓FDE	↓AOE	↓ADE	↓FDE	↓AOE	↓ADE	↓FDE	↓AOE	↓ADE	↓FDE	↓AOE	↓ADE	↓FDE	↓AOE	↓ADE	↓FDE
With Goal	GNM	0.20	0.42	0.84	0.17	0.36	0.72	0.13	0.39	0.79	0.23	0.61	1.20	0.19	0.36	0.74	0.13	0.38	0.71
	ViNT	0.18	0.54	1.00	0.19	0.51	0.95	0.17	0.55	1.03	0.30	0.79	1.47	0.20	0.48	0.94	0.17	0.50	0.93
	NoMaD	0.13	0.40	0.73	0.07	0.29	0.55	0.04	0.34	0.61	0.09	0.53	0.93	0.10	0.30	0.57	0.09	0.34	0.61
	CityWalker	0.84	0.95	1.75	0.80	0.90	1.67	0.84	0.98	1.83	0.75	1.08	2.02	0.83	0.90	1.66	0.87	0.93	1.71
	Narrate2Nav	**0.05**	**0.16**	**0.23**	**0.01**	**0.10**	**0.13**	**0.01**	**0.14**	**0.20**	**0.04**	**0.17**	**0.24**	**0.02**	**0.13**	**0.18**	**0.04**	**0.16**	**0.24**
Without Goal	ViNT	0.29	0.57	1.10	0.38	0.60	1.17	0.26	0.62	1.16	0.33	0.85	1.57	0.27	0.50	0.98	0.34	0.60	1.15
	NoMaD	0.14	0.41	0.76	0.09	0.31	0.58	0.05	0.34	0.61	0.08	0.52	0.91	0.10	0.31	0.58	0.10	0.36	0.65
	CityWalker	0.84	0.95	1.75	0.80	0.90	1.67	0.84	0.98	1.83	0.75	1.08	2.02	0.83	0.90	1.66	0.87	0.93	1.71
	Narrate2Nav	**0.07**	**0.26**	**0.45**	**0.03**	**0.16**	**0.27**	**0.01**	**0.21**	**0.35**	**0.06**	**0.23**	**0.38**	**0.05**	**0.20**	**0.35**	**0.05**	**0.23**	**0.4**

Narrate2Nav NoMAD GNM ViNT

Fig. 5: Qualitative comparative analysis of activation maps and predicted trajectories (projected in 2D) in two scenarios (outdoor and indoor) in human-populated areas between NARRATE2NAV and four SoTA models. In both scenarios, the robot is expected to stop and yield to people, while all SoTA models generate trajectories that collide with them. NARRATE2NAV's attention map clearly highlights the people obstructing its path.

entire or irrelevant parts of the scene. This results in inaccurately predicted trajectories that collide with humans. We attribute this refined attention map to the textual signals that directly point to regions of interest for navigation tasks.

C. Real-World Analysis

We evaluate the success rate of goal reaching and collision avoidance for our method, NARRATE2NAV, across four challenging real-world scenarios, comparing its performance against baselines GNM [14], ViNT [15], and NoMAD [16]. CityWalker [17] is excluded from real-world deployment due to its poor performance on the offline test set. To ensure a fair assessment, we omit one scenario, Navigation in Crowds, where all models, including NARRATE2NAV, exhibit degraded performance. As required by their methodologies, we pre-recorded trajectory images to construct topological maps for the baselines, consistent with GNM [14], ViNT [15], and NoMAD [16]. The performance results are summarized in Table II.

We define the success rate as reaching the goal and a collision as any instance where the human must yield to the robot (though reaching the goal may still be possible). During experiments, changes in scene dynamics, lighting, or available space significantly impact baseline performance, leading to lower success rates. In contrast, NARRATE2NAV demonstrate robustness to these variations, maintaining consistent performance across scenarios. To address challenges in large environments, where baselines have more room to deviate from the goal, we conduct human-following experiments in narrow passageways, but this is not feasible for the intersection scenario. Notably, the narrow passageways enable higher goal-reaching success for baselines compared to intersections and frontal approaches, as the constrained spaces limit deviations from the goal.

D. Ablation Study

Text Module: Table III examines the effectiveness of incorporating social context and textual signals in the model's pretext training. We ablate the textual input, following the same design shown in Fig. 2, and remove only the text encoder model. Our results indicate that the overall performance of our model on the same task drops 14.8%. This clearly show the effect of the text signals in our model.

Early Fusion vs. Late Fusion of the Goal Representation: We study the choice between early fusion and late fusion for goal-observation in the Student Model. We begin training with late fusion, where the history of observations is first passed to a transformer, and the goal is encoded using a linear

TABLE II: Comparison of NARRATE2NAV with three SoTA Methods in Real-World Scenarios. Collision and success rate are reported per 10 trials. 'FAIL' indicates that the method fails to complete any trial in that scenario.

Method	Intersection		Frontal Approach		Human Following		Narrow Passageway		All	
	↓Collision	↑Success Rate	↓Collision	↑Success Rate	↓Collision	↑Success Rate	↓Collision	↑Success Rate	↓Collision	↑Success Rate
GNM	**3/10**	0/10	FAIL	FAIL	5/10	9/10	**5/10**	9/10	4.3/10	6/10
ViNT	**3/10**	1/10	8/10	2/10	3/10	**10/10**	7/10	**10/10**	5.25/10	5.75/10
NoMaD	7/10	2/10	5/10	3/10	**1/10**	**10/10**	7/10	8/10	5/10	5.75/10
Narrate2Nav	**3/10**	**10/10**	**3/10**	**10/10**	2/10	**10/10**	**5/10**	4/10	**3.25/10**	**8.5/10**

TABLE III: Results on Ablating the Text Module.

Method	All		
	↓AOE	↓ADE	↓FDE
No Text Module (Ablation)	0.06	0.19	0.24
Narrate2Nav	**0.04**	**0.16**	0.24

transformation. The output sequential embeddings from the

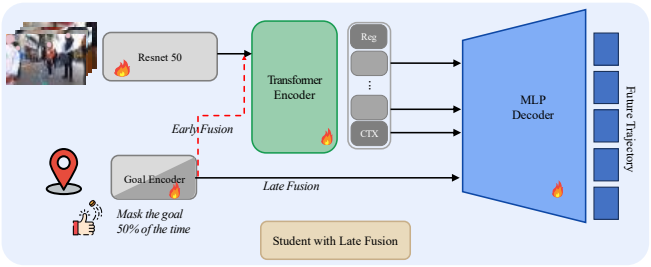

Fig. 6: Late fusion of goal ablation study

transformer and the goal encoder are then concatenated and processed by a fully connected neural network. We observe rapid convergence, but the model mainly learns goal distance and produces evenly segmented trajectories and straight line instead of reasoning about actions (e.g., turns, stops).

TABLE IV: Comparison of Fusion Methods.

Method	All		
	↓AOE	↓ADE	↓FDE
Late Fusion (Ablation)	0.89	0.98	**0.04**
Early Fusion	**0.04**	**0.16**	0.24

Table IV shows that the late fusion ablation model reaches the goal (almost zero FDE), but the path does not align with the ground truth trajectory (ADE is large), which may cause collisions with obstacles, as the majority of the predicted trajectories are straight lines toward the goal.

V. CONCLUSION

In this work, we propose NARRATE2NAV, a self-supervised learning-based approach to real-time vision-action modeling that integrates human-like language reasoning for context-aware navigation in human-centric environments. Using a novel pretraining method with Barlow Twins loss and multi-modal future state representations, it embeds social and contextual cues into a visual encoder, enabling efficient RGB-only navigation. NARRATE2NAV achieves a 52.94% improvement in ADE and a 41.67% improvement in real-world experiments over SoTA baselines across various challenging scenarios. Furthermore, our analysis of visual attention maps shows that NARRATE2NAV highlights key social cues and critical regions for navigation, showing the effectiveness of our framework in complex environments. Looking ahead, our approach opens promising new avenues for integrating richer forms of human-like language reasoning into real-time robotic systems.

VI. LIMITATIONS

Our primary motivation in this work is to demonstrate the potential of textual input as a distinct modality and to propose a novel Vision-Language-Action (VLA) architecture. This architecture not only addresses the computational challenges of VLMs but also achieves data efficiency and outperforms models trained on internet-scale data. We acknowledge the open challenges that limit our work. Despite extensive research on grounding language in pixel values, VLMs still struggle with basic spatial reasoning tasks, such as left-right differentiation, particularly in complex scenarios. Our analysis of a generated dataset reveals instances of inaccurate descriptions. These inaccuracies, combined with the challenge of determining which textual information provides a useful signal for generating improved trajectories, constrain the scalability of our approach. In this work, we primarily adopt the methodology proposed by Social-LLaVA [10] as a signal for the vision encoder. However, there remains significant scope for research into diverse information styles to analyze how they can enhance robotic observation, next-state prediction, and subsequent action prediction. Given our contribution to mitigating the need for large VLMs that are not optimized for real-time inference, it is valuable to compare the performance of NARRATE2NAVE architecture with frameworks that directly utilize off-the-shelf large VLMs for visual robot navigation in human-populated environments.

We identify a potential direction for future research is to benchmark our model's performance against such frameworks in real-time navigation tasks. Finally, we acknowledge that our attempt is to transfer reasoning-like features via latent embedding supervision. Further testing is required to fully verify emergent reasoning behaviors.

We recognize that our model is not yet ready for different robot embodiments or strongly robust across diverse sce-

narios. Engineering effort, hyperparameter tuning, scenario selection, context, and, most importantly, the training data play pivotal roles in this regard. In the real-world experiment, we present a proof of concept demonstrating our work's advantages over SoTA models, with the goal of motivating further research into using linguistic signals as a separate modality for improving short- and long-horizon path planning in visual navigation. The real-world experiment is subject to various factors, such as training data, human behavior, and test environment, which may affect outcomes. We strive to mitigate these limitations to deliver robust, reliable, and fair analysis, fostering further research in this field.

REFERENCES

[1] A. Francis, C. Pérez-d'Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra *et al.*, "Principles and guidelines for evaluating social robot navigation algorithms," *ACM Transactions on Human-Robot Interaction*, vol. 14, no. 2, pp. 1–65, 2025.

[2] R. Mirsky, X. Xiao, J. Hart, and P. Stone, "Conflict avoidance in social navigation—a survey," *ACM Transactions on Human-Robot Interaction*, vol. 13, no. 1, pp. 1–36, 2024.

[3] C. Mavrogiannis, F. Baldini, A. Wang, D. Zhao, P. Trautman, A. Steinfeld, and J. Oh, "Core challenges of social robot navigation: A survey," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 3, pp. 1–39, 2023.

[4] X. Xiao, B. Liu, G. Warnell, and P. Stone, "Motion planning and control for mobile robot navigation using machine learning: a survey," *Autonomous Robots*, vol. 46, no. 5, pp. 569–597, 2022.

[5] A. Payandeh, K. T. Baghaei, P. Fayyazsanavi, S. B. Ramezani, Z. Chen, and S. Rahimi, "Deep representation learning: Fundamentals, technologies, applications, and open challenges," *IEEE Access*, vol. 11, pp. 137 621–137 659, 2023.

[6] D. M. Nguyen, M. Nazeri, A. Payandeh, A. Datar, and X. Xiao, "Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 7442–7447.

[7] J. Liang, A. Payandeh, D. Song, X. Xiao, and D. Manocha, "Dtg : Diffusion-based trajectory generation for mapless global navigation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 5340–5347.

[8] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning," *The International Journal of Robotics Research*, 2016.

[9] D. Song, J. Liang, A. Payandeh, A. H. Raj, X. Xiao, and D. Manocha, "Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 508–515, 2025.

[10] A. Payandeh, D. Song, M. Nazeri, J. Liang, P. Mukherjee, A. H. Raj, Y. Kong, D. Manocha, and X. Xiao, "Social-llava: Enhancing robot navigation through human-language reasoning in social spaces," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[11] S. Narasimhan, A. H. Tan, D. Choi, and G. Nejat, "Olivia-nav: An online lifelong vision language approach for mobile robot social navigation," 2025. [Online]. Available: https://arxiv.org/abs/2409.13675

[12] D. Song, J. Liang, X. Xiao, and D. Manocha, "Vl-tgs: Trajectory generation and selection using vision language models in mapless outdoor environments," *IEEE Robotics and Automation Letters*, vol. 10, no. 6, pp. 5791–5798, 2025.

[13] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International conference on machine learning*. PMLR, 2021, pp. 12 310–12 320.

[14] D. Shah, A. Sridhar, A. Bhorkar, N. Hirose, and S. Levine, "Gnm: A general navigation model to drive any robot," 2023. [Online]. Available: https://arxiv.org/abs/2210.03370

[15] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine, "Vint: A foundation model for visual navigation," 2023. [Online]. Available: https://arxiv.org/abs/2306.14846

[16] A. Sridhar, D. Shah, C. Glossop, and S. Levine, "Nomad: Goal masked diffusion policies for navigation and exploration," 2023. [Online]. Available: https://arxiv.org/abs/2310.07896

[17] X. Liu, J. Li, Y. Jiang, N. Sujay, Z. Yang, J. Zhang, J. Abanes, J. Zhang, and C. Feng, "Citywalker: Learning embodied urban navigation from web-scale videos," *arXiv preprint arXiv:2411.17820*, 2024.

[18] A. Eftekhar, K.-H. Zeng, J. Duan, A. Farhadi, A. Kembhavi, and R. Krishna, "Selective visual representations improve convergence and generalization for embodied ai," 2024. [Online]. Available: https://arxiv.org/abs/2311.04193

[19] M. Nazeri, J. Wang, A. Payandeh, and X. Xiao, "Vanp: Learning where to see for navigation with self-supervised vision-action pre-training," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 2741–2746.

[20] M. Elnoor, K. Weerakoon, G. Seneviratne, J. Liang, V. Rajagopal, and D. Manocha, "Vi-lad: Vision-language attention distillation for socially-aware robot navigation in dynamic environments," 2025. [Online]. Available: https://arxiv.org/abs/2503.09820

[21] A. Pokhrel, M. Nazeri, A. Datar, and X. Xiao, "Cahsor: Competence-aware high-speed off-road ground navigation in $\mathbb{SE}(3)$," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 9653–9660, 2024.

[22] P. Roth, J. Nubert, F. Yang, M. Mittal, and M. Hutter, "Viplanner: Visual semantic imperative learning for local navigation," 2024. [Online]. Available: https://arxiv.org/abs/2310.00982

[23] H. Wang, A. H. Tan, and G. Nejat, "NavFormer: A Transformer Architecture for Robot Target-Driven Navigation in Unknown and Dynamic Environments," *arXiv preprint arXiv:2402.06838*, 2024.

[24] N. Hirose, C. Glossop, A. Sridhar, D. Shah, O. Mees, and S. Levine, "Lelan: Learning a language-conditioned navigation policy from in-the-wild video," in *Conference on Robot Learning*, 2024.

[25] A. Eftekhar, L. Weihs, R. Hendrix, E. Caglar, J. Salvador, A. Herrasti, W. Han, E. VanderBil, A. Kembhavi, A. Farhadi, R. Krishna, K. Ehsani, and K.-H. Zeng, "The one ring: a robotic indoor navigation generalist," 2024. [Online]. Available: https://arxiv.org/abs/2412.14401

[26] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu, "Spatialrgpt: Grounded spatial reasoning in vision language models," 2024. [Online]. Available: https://arxiv.org/abs/2406.01584

[27] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, and Z. Fan, "Qwen2 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2407.10671

[28] M. Nazeri, A. Pokhrel, A. Card, A. Datar, G. Warnell, and X. Xiao, "Vertiformer: A data-efficient multi-task transformer for off-road robot mobility," 2025. [Online]. Available: https://arxiv.org/abs/2502.00543

[29] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2024. [Online]. Available: https://arxiv.org/abs/2304.07193

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[32] H. Karnan, A. Nair, X. Xiao, G. Warnell, S. Pirk, A. Toshev, J. Hart, J. Biswas, and P. Stone, "Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation," *IEEE Robotics and Automation Letters*, 2022.

[33] S. Pirk, E. Lee, X. Xiao, L. Takayama, A. Francis, and A. Toshev, "A protocol for validating social navigation policies," *arXiv preprint arXiv:2204.05443*, 2022.