Title: Context-Aware Multimodal Navigation through Causal Crowd Shaping with Vision Language Models

Author: Dr. Arman Ibrayeva, Information Science, Cornell University (Cornell Tech)

Abstract. The primary goal of this work is to build a next-generation autonomous system that is not merely adaptive to environment dynamics, but shapes and controls it. The project will address two Amazon Robotics Spring 2026 focus areas: "AI for Robotics and Human-Robot Interaction (HRI)", and "Autonomous Navigation". In Amazon fulfillment centers, autonomous mobile robots (AMRs) lose throughput in mixed human–robot zones as they wait for workers who block their path. Scaled across million AMRs and many encounters per shift, even a second saved per yield-event compounds into massive recovered work hours. This bottleneck reflects a broader scientific limitation. Most navigation methods model humans as independent agents whose motion is unaffected by the robot. A smaller class accounts for humans' cooperation with the robot, modeling their reaction dependence on the robot's current state, not a causal effect of the robot's next actions. Human responses are accounted for in safety constraints (that shape robot's action), not as outcomes to be optimized (in robot's planning). **Conflicting** objectives of balancing (c1) *safety* and (c2) *time efficiency* in dense human environment navigation is an open challenge, especially (c3) *under computational constraints* of real-time deployment and (c4) accounting for *real world complexity*. Instead of treating navigation as pure obstacle avoidance, we study passage creation: the robot advances in the environment by influencing local human motion in the same way people do when passing pedestrians on a sidewalk or when changing lanes in traffic, which **existing autonomous systems do not achieve** [1]. We will address the problem (c1) through a vision-language model (VLM) based context awareness that evaluates not only collision risk, but also whether inducing a person to yield could expose them to surrounding hazards; (c2) through "passage negotiation": first communicating its intent via multimodal cues then careful "motion nudges", while *accounting for humans' comfort***.** We are proposing a novel framework in which the robot optimizes not only its own actions but also **treats humans behavior as optimizable parameters**; (c3) via decoupled architecture: slower VLM that tunes parameters of faster causal-interaction model and controller; (c4) incorporating VLM also provides robust semantic understanding across diverse real-world scenes and human contexts. More broadly, the framework can be extended from fulfillment-center AMRs and service robotics to Zoox urban driving in the future.

Keywords: context-aware navigation, human-robot interaction, multimodal perception, transparency.

Introduction. Amazon expects automation to generate approximately $12.6B in savings over 2025–2027, as part of its long-term objective to automate roughly 75% of its operations [2]. This is reflected in measurable productivity gains over the past decade: the number of packages shipped per employee annually has increased from 175 to nearly 3,870 [3]. The company frames the goal of automation as "making warehouse work safer and more efficient," as robots significantly outperform human labor in throughput [4].

Amazon deploys advanced coordination systems such as DeepFleet, built on Amazon SageMaker. This robot traffic controller demonstrated a visible impact, improving **travel efficiency by 10%** and **reducing delivery costs** [5]. This highlights that traffic regulation is a critical factor for productivity, although current solutions focus on robot–robot coordination, not human-robot traffic. Most of the company's AMRs operate in caged zones on fiducial marked floors with no human entry. The introduction of the Amazon Proteus robot reflects the company's interest in extending AMRs into human-shared environments. Amazon acknowledges the difficulty of safely incorporating robots into human-shared environments and expresses a vision to expand such deployments across the network [6]. Proteus uses a **reactive safety policy**, **stopping** when humans enter its path, and is deployed in low-density outbound GoCart staging [6]. To our knowledge, Amazon does not currently deploy AMRs designed to operate efficiently in **tight or crowded environments**, **where humans block passage**. Stopping policy (and broader reactive avoidance principle) is inherently inefficient and becomes particularly limiting in dense settings [7].

Amazon's scale converts small per-encounter delays into substantial aggregate losses**.** It operates over 1M robots across more than 300 fulfillment and logistics facilities and employs ~1.6M workers, of which ~1.2M are in operations and fulfillment roles [8, 9, 5]; and spent ~$100B on fulfillment and ~$90B on shipping in 2024 [8]. Financial analyses of Amazon's business model states that given its "*high revenue, large cost base, and relatively low margins, even small improvements in operational efficiency translate into substantial bottom-line impact*" [10].

Time optimality can be significantly increased if robots avoid freezing, and passively following slow walkers, or executing large detours when encountering people; and instead, create passage.

Beyond classical dynamic-environment navigation approaches, (e.g. dynamic window approach [11], model predictive control (MPC) [12]), many studies addressed navigation in the presence of humans. Most

treat humans as independent agents leading to detours around them or waits and this produces the freezing-robot problem in dense scenes [7]. A smaller class of studies models coupled robot–human motion, meaning robot's controller accounts for human's possible reaction (such as ORCA [13], Social Force [14], SICNav and SICNav-Diffusion [15, 16]), but relies on hand-specified interaction assumptions that fail to capture the diversity and unpredictability of real human behavior. For example, ORCA assumes agents reciprocally share collision-avoidance responsibility, Social Force assumes pedestrian motion can be captured through fixed attractive and repulsive forces, and SICNav inherits ORCA-style assumptions by embedding ORCA-based human prediction constraints into the robot planner. These assumptions can overlook differences in attention, urgency, cooperation, group behavior, and willingness to yield. Learning-based methods improve adaptability (e.g. SoNIC [17], Falcon [18] ICP [19]), but share following limitations: (a) "avoiding-only" policy with no passage negotiation and active communication with other agents; (b) safety is defined as **non-collision with the robot only**, without evaluating whether a robot's action may induce a human to move into harm (e.g. into restricted zones) or discomfort, which requires semantic scene understanding they lack. In addition, their state representations remain kinematic, limiting their ability to account for context-dependent human factors (e.g. attention, workload) [18, 19].

Recent work incorporates vision-language models (VLMs) into navigation to provide **semantic scene understanding** and **instructional grounding**, for tasks such as object search, instruction following, and directional guidance. The closest works that combine VLMs with planning are designed for different problems, and are not applicable to the problem we target: efficient and safe navigation in crowds, and restricted to coarse action spaces, high inference latency of several seconds per query (e.g., LISN [20]).

Autonomous vehicles (AVs) have long exhibited a fundamental bottleneck in dense traffic: an agent programmed to always yield makes no progress in dense pedestrian environments, since pedestrians who recognize the policy will exploit it [21]. Even state-of-the-art deployed systems exhibit the same constraint in vehicle–vehicle interaction: empirical analysis [22] shows that AVs **wait** for significant lag gaps to appear rather than evaluating risks and **taking actions for gap formation**. This conservative behavior translates to documented failure modes in deployment, including multi-vehicle deadlocks at intersections [23] and inability to merge in dense traffic [1]. Solving navigation in dense human-shared environments therefore requires a fundamentally different formulation.

2. Methods. We will build an interventional crowd navigator: the robot predicts how nearby humans would respond to each candidate action, and optimizes its own progress towards goal, as well as humans' progress towards passage clearing, under VLM-enforced safety and comfort in open-world scenes.

Semantic context from VLM. Safety in human-robot navigation is often formulated as preventing collision along the robot's own trajectory. In our approach, the robot must also avoid inducing human motion toward unsafe or task-critical regions. Our system detects open-ended hazards near people and infers human-state attributes, such as attentiveness and work engagement. These outputs help the robot avoid inducing people toward unsafe regions and improve human behavior prediction. We use a foundation model (FM), LoRA-tuned Qwen3-VL-3B [24], whose open-vocabulary reasoning covers diverse hazards, ranging from painted floor markings to active work zones, as well as diverse human-state attributes. However, FMs do not output reliable metric quantities and coordinates for control. We therefore fuse VLM reasoning with robot-frame geometry through three steps. First, a lightweight perception module detects and tracks humans using YOLO and ByteTrack, estimates their 3D positions and velocities from depth inputs, and extracts local occupancy from LiDAR. Second, we input these metrics to the VLM as visual prompts: each tracked person is annotated with kinematic information, and the local region around the person is divided into numbered sectors. The prompted RGB image and **session-level task description** are passed to the VLM. Third, after the VLM returns structured JSON with per-sector hazard scores, human-context attributes, and communication cues, a semantic-spatial grounding module maps these outputs back to the corresponding robot-frame sectors and tracked humans. The resulting grounded semantic state combines what the VLM recognizes with where it is located, and is passed to the human-motion predictor and planner.

Action-outcome predictor. The predictor estimates how each nearby human will move with and without the candidate action so the controller can select actions whose induced response opens passage. The world state (robot kinematics, per-person position, velocity and VLM-derived features, the static-obstacle positions, and the active communication cue) is mapped through small MLPs to shared token dimension, with persons kept as separate tokens, and processed by a Transformer encoder. Self-attention captures interactions among humans, the robot, and the candidate action, while remaining invariant to arbitrary indexing of the detected humans (insensitive to the order of detected persons), which is the appropriate inductive bias for variable-size

crowds. The encoder output is cached once per control cycle, and only the action token is replaced for each candidate, enabling efficient batched evaluation of candidate actions at the control rate. Two MLP heads consume the cached encoding: one predicts trajectories under no intervention, second – under the candidate action. We will train them jointly as a doubly-robust learner with cross-fitting, which corrects for confounding by the data-collecting policy. A deep ensemble estimates prediction uncertainty.

Context-Aware Interventional Controller (CAIC). We define the feasible action set (Afeas) by three constraints: a control-barrier-function condition enforcing minimum clearance between the robot and every person; a lower bound on time-to-collision; the probability threshold that any person enters a VLM-flagged hazard region at any lookahead step under the candidate action. The optimization runs at 20–30 Hz in a receding-horizon loop, with the VLM-derived semantic state refreshed at 2–5 Hz. The policy:

$$a^{*} = \arg\min_{ak \in \mathcal{A}feas} \mathbb{E}\xi \left[\sum_{\ell=0}^{T-1} c\Big(R_{t+\ell}, ak, H^{(1:N)}_{t+\ell+1}(ak) \Big) \,\Big|\, s_t \right],$$

the selected action a∗ minimizes over candidates ak∈Afeas⊆A the expected cumulative cost over a short horizon of T steps, conditioned on the current full state st (concatenated robot state R, states of N near humans H, and scene context). The cost c(·) consists of 3 terms. Goal deviation minimizes the robot's progress towards target, and each person's deviation from *desired set DH* (that create passage for robot); comfort is quantified as a deviation of H(ak) from their nominal, unaffected motion, weighted by the VLM-derived distraction gain; and smoothness of robot's motion.

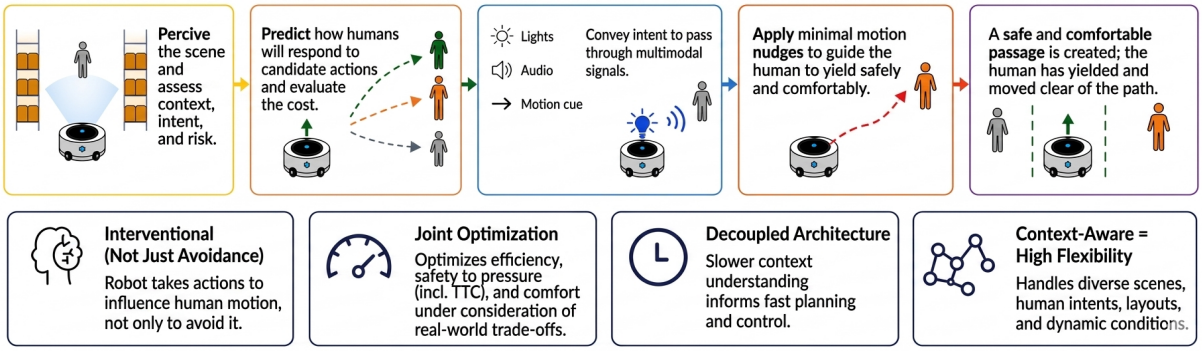


Figure 1. Robot passage negotiation sequence using perception, prediction, communication, and motion nudging

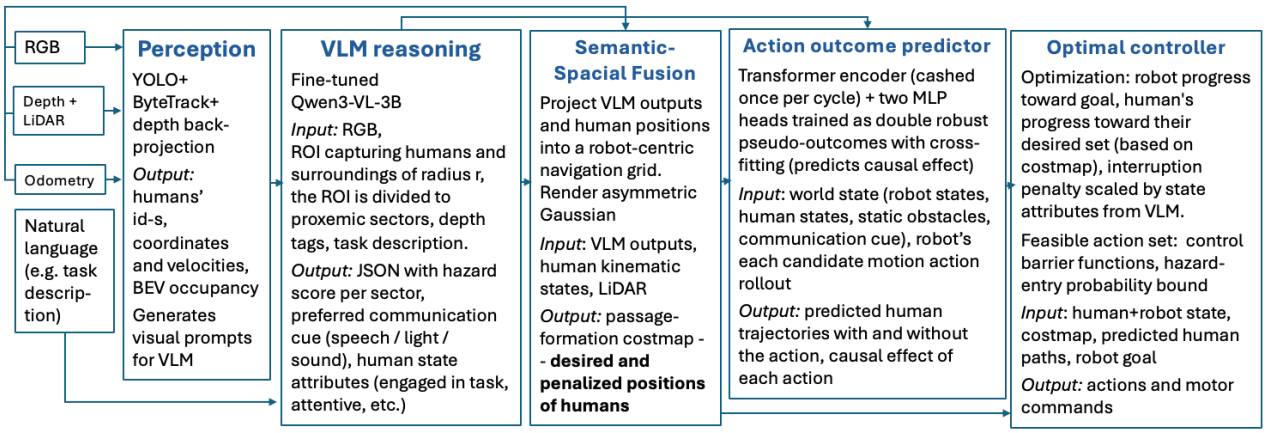


Figure 2. Methodology outline

Dataset and training. We will fine-tune Qwen3-VL-3B [24] with LoRA on the MUSON dataset [25], a recent dataset for socially compliant navigation (superseding SNEI / Social-LLaVA [26]). MUSON provides human-annotated scene-level reasoning: perception, prediction, chain-of-thought, action, and explanation over crowded human-robot interactions, which grounds the VLM in social navigation semantics. Pedestrian causal response will be trained on PeRoI (three robot conditions × pedestrian response labels) and extended with simulation data from an Arena-3.0 / HuNavSim [28, 29] under randomized assertiveness. W will collect additional social navigation data and closed-loop evaluation on a mobile robot (AgilexRobotics Limo Pro,

RGB-D + 2D lidar + Jetson Orin) on the Cornell Tech campus (corridors, café entryways, laboratory, where controlled crowd conditions will be created);

Evaluation. We will compare against well-known social-navigation methods: ORCA [13], Social Force Model (SFM) [14], SARL [30] (Socially Attentive RL); and SICNav [15] (bilevel MPC with embedded ORCA, the current SOTA interactive planner). *Navigation performance:* time to goal, motion smoothness; *influence quality*: passage creation rate, freezing rate; *safety*: collision rate, induced-hazard incidence; response prediction accuracy.

Contributions. We introduce a new navigation method for space-constrained human-shared environments by optimizing not only the robot's progress toward its goal but also each person's progress toward a *desired* position (that creates passage), and the human's goal is induced by the robot's kinematic motion and 'negotiated' through robot's multi-channel intent expression (language, sound, light). Grounding this formulation in real-world perception, we fuse multimodal sensing (RGB, depth, language, odometry) into VLM-based context awareness that evaluates not only collision risk but whether influencing a person introduces secondary hazards or disrupts task-engaged workers, yielding situationally grounded risk and comfort assessment; and a decoupled real-time architecture that enables fast control informed by slower semantic reasoning. To support this, we release new multimodal, physically embodied data and a benchmark capturing human–robot interactions in dense, real-world settings.

Broader impact. Our approach enables generalizable navigation across human-centered environments by combining causal reasoning that captures relationships that are invariant across environments (confounders), (in contrast to correlational learning methods) with VLMs for semantic understanding, ensuring robustness and reliability under real-world variability. The same formulation extends beyond fulfillment centers to outdoor deployments such as delivery robots and self-driving vehicles (Zoox).

Appendix A — References

[1] Yang Wang, et al. Traffic safety evaluation of emerging mixed traffic flow at freeway merging area considering driving behavior. Scientific Reports, 15, 2025. doi: 10.1038/s41598-025-94658-y.

[2] Karen Weise and Peter Eavis. Amazon plans to replace more than half a million jobs with robots. The New York Times, October 2025. URL https://www.nytimes.com/2025/10/21/business/economy/amazon-robots-jobs.html.

[3] Sebastian Herrera. Amazon is on the cusp of using more robots than humans in its warehouses. The Wall Street Journal, 2025. URL https://www.wsj.com/tech/ amazon-warehouse-robots-automation-942b814f.

[4] Amazon. Amazon introduces new robotics solutions. About Amazon (corporate news), 2022. URL https://www.aboutamazon.com/news/operations/amazon-introduces-new-robotics-solutions.

[5] Scott Dresser. Amazon launches a new AI foundation model to power its robotic fleet and deploys its 1 millionth robot. About Amazon (corporate news), July 2025. URL https://www.aboutamazon.com/news/operations/amazon-million-robots-ai-foundation-model.

[6] Amazon. 10 years of Amazon robotics: How robots help sort packages, move product, and improve safety. About Amazon (corporate news), 2022. URL https://www.aboutamazon.com/news/operations/10-years-of-amazon-robotics-how-robots-help-sort-packages-move-product-and-improve-safety.

[7] Pete Trautman and Andreas Krause. Unfreezing the robot: Navigation in dense, interacting crowds. In 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 797–803, 2010. doi: 10.1109/IROS.2010.5654369.

[8] Amazon.com, Inc. Annual report on Form 10-K for the fiscal year ended December 31, 2024. U.S. Securities and Exchange Commission filing, 2025. URL https://www.sec.gov/Archives/

edgar/data/1018724/000101872425000004/amzn-20241231.htm. Reports $98.5B fulfillment and $95.8B shipping costs in 2024.

[9] Daniel Bortz. Amazon aims to replace 600,000 workers with robots. Entrepreneur, 2025. URL https://www.entrepreneur.com/business-news/amazon-aims-to-replace-600000-workers-with-robots/498636.

[10] Summit Stocks. Amazon's robotics push is here: The bull case for operational leverage. Summit Stocks (Substack), 2025. URL https://summitstocks.substack.com/p/amazons-robotics-push-is-here-the.

[11] Dieter Fox, Wolfram Burgard, and Sebastian Thrun. The dynamic window approach to collision avoidance. IEEE Robotics & Automation Magazine, 4(1):23–33, 1997. doi: 10.1109/100.580977.

[12] James B. Rawlings, David Q. Mayne, and Moritz M. Diehl. Model Predictive Control: Theory, Computation, and Design. Nob Hill Publishing, 2nd edition, 2017.

[13] Jur van den Berg, et al. Reciprocal n-body collision avoidance. In Robotics Research: The 14th International Symposium ISRR, volume 70 of Springer

Tracts in Advanced Robotics, pages 3–19. Springer, 2011.

[14] Social force model for pedestrian dynamics. Physical Review E, 51(5):4282–4286, 1995.

[15] Sepehr Samavi, et al. SICNav: Safe and interactive crowd navigation using model predictive control and bilevel optimization. IEEE Transactions on Robotics, 41:801–818, 2024. doi: 10.1109/TRO.2024.3484634.

[16] Sepehr Samavi, Anthony Lem, Fumiaki Sato, Sirui Chen, Qiao Gu, Keijiro Yano, Angela P. Schoellig, and Florian Shkurti. SICNav-diffusion: Safe and interactive crowd navigation with diffusion trajectory predictions. IEEE Robotics and Automation Letters, 2025. arXiv:2503.08858.

[17] Jianpeng Yao, Xiaopan Zhang, Yu Xia, Zejin Wang, Amit K. Roy-Chowdhury, and Jiachen Li. SoNIC: Safe social navigation with adaptive conformal inference and constrained reinforcement learning. arXiv preprint arXiv:2407.17460, 2024.

[18] Zeying Gong, et al. From cognition to precognition: A future-aware framework for social navigation. In 2025 IEEE International Conference on Robotics and Automation (ICRA), 2025. arXiv:2409.13244.

Corl [19] Zhe Huang, et all. Interaction-aware conformal prediction for crowd navigation. arXiv preprint arXiv:2502.06221, 2025.

[20] Junting Chen, et al. LISN: Language-instructed social navigation with VLM-based controller modulating. arXiv preprint arXiv:2512.09920, 2025.

[21] Adam Millard-Ball. Pedestrians, autonomous vehicles, and cities. Journal of Planning Education and Research, 38(1):6–12, 2018. doi: 10.1177/0739456X16675674.

[22] Tarek Sayed, et al. Automated vehicles at unsignalized intersections: Safety and efficiency implications of mixed human and automated traffic. arXiv preprint arXiv:2410.12538, 2024.

[23] Sean O'Kane. Waymo explains why its robotaxis got stuck during the SF blackout. TechCrunch, 2025. URL https://techcrunch.com/2025/12/24/waymo-explains-why-its-robotaxis-got-stuck-during-the-sf-blackout/

[24] Shuai Bai, et al. Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923, 2025.

[25] Zhuonan Liu, et al. MUSON: A reasoning-oriented multimodal dataset for socially compliant navigation in urban environments. arXiv:2512.22867, 2025.

[26] Amirreza Payandeh, et al. Social-LLaVA: Enhancing social robot navigation through human–language reasoning. In 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2025. arXiv:2501.09024.

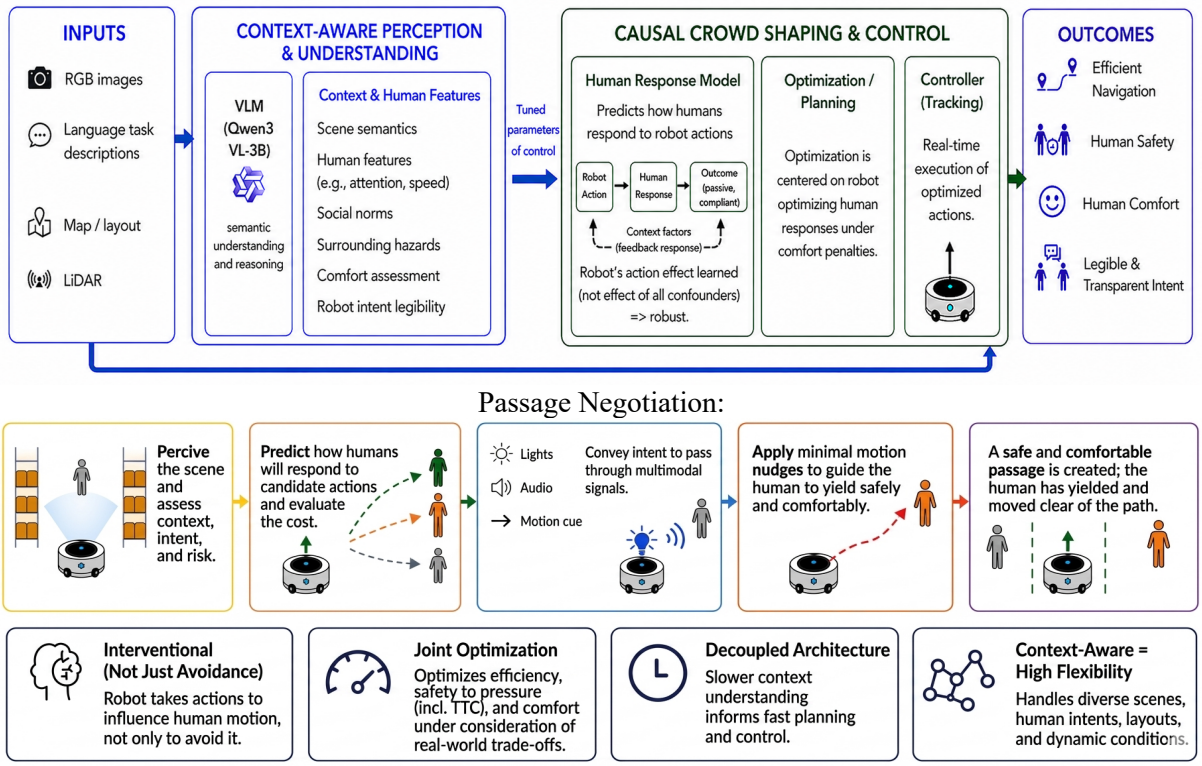
[27] Subham Agrawal, et al. PeRoI: A pedestrian-robot interaction dataset for learning avoidance, neutrality, and attraction behaviors in social navigation. arXiv preprint arXiv:2503.16481, 2025.

[28] Linh Kästner, et al. Arena 3.0: Advancing social navigation in collaborative and highly dynamic environments. arXiv preprint arXiv:2406.00837, 2024.

[29] Noé Pérez-Higueras, et al. HuNavSim: A ROS 2 human navigation simulator for benchmarking human-aware robot navigation. arXiv preprint arXiv:2305.01303, 2023.

[30] Changan Chen, et al. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In 2019 IEEE International Conference on Robotics and Automation (ICRA), pages 6015–6022, 2019. doi: 10.1109/ICRA.2019. 8794134.

[31] Zhiyu Huang, etal. TIC-VLA: A think-in-control vision-language-action model for robot navigation in dynamic environments. arXiv preprint arXiv:2602.02459, 2026



Passage Negotiation:

Figure version 1

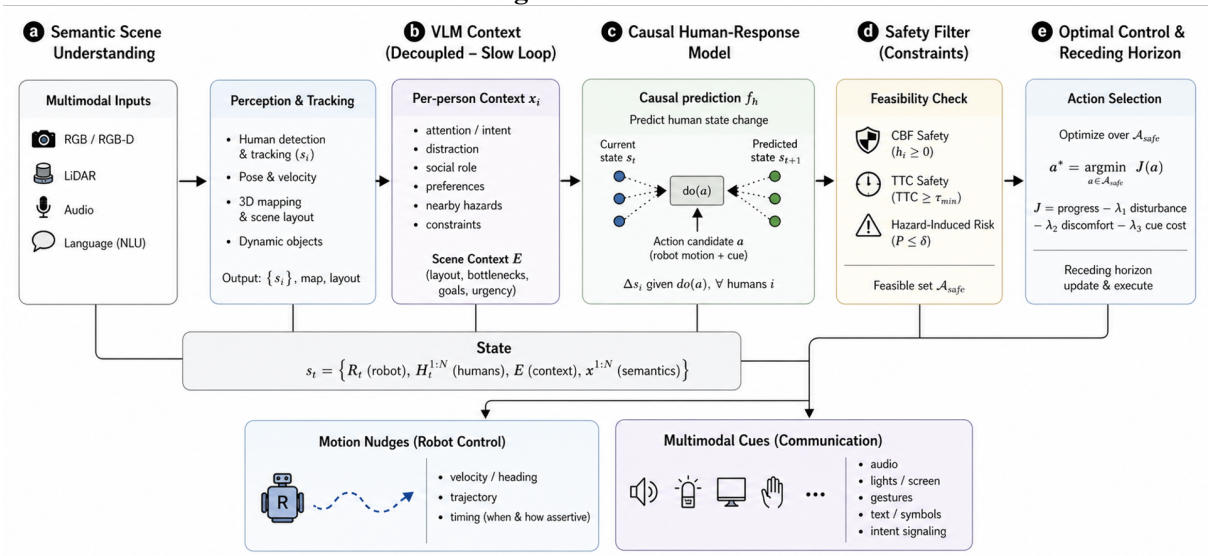


Figure version 2