

# Decoding Pedestrian Crossing Intentions from Egocentric Vision via Vision-Language Models

Danya Li<sup>1</sup>, Xiang Su<sup>2</sup>, Yan Feng<sup>3</sup>, and Rico Krueger<sup>1</sup>

**Abstract**—Egocentric vision captures pedestrian visual perception and behavior from a first-person perspective, offering fine-grained insights crucial for pedestrian intention modeling. However, its application in traffic safety remains largely unexplored. We reformulate pedestrian crossing intention prediction as a Visual Question Answering (VQA) task, leveraging the pre-trained knowledge and reasoning abilities of Vision-Language Models (VLMs). We first benchmark two state-of-the-art VLMs in a zero-shot setting, finding that they achieve moderate gains over random guessing in accuracy but exhibit limited higher-level reasoning. Built upon this observation, we employ parameter-efficient fine-tuning to adapt VLMs to the task. Our results show that the adapted models significantly outperform zero-shot approaches and achieve a 9% relative improvement in accuracy over a specialized transformer-based baseline. Finally, we demonstrate that integrating dynamic eye-gaze signals and personal attributes further boosts predictive accuracy, establishing a new state-of-the-art for egocentric intent decoding.

## I. INTRODUCTION

Ensuring pedestrian safety in increasingly complex urban environments critically depends on accurate pedestrian behavior anticipation. Traditional approaches have predominantly relied on exocentric perspectives, such as vehicle-mounted or surveillance cameras, which offer a stable and global view of the scene [1], [2]. However, these external viewpoints often fail to capture the pedestrian’s first-person perception and fine-grained behavioral cues. In contrast, egocentric vision offers direct access to the pedestrian’s line of sight, providing rich information that is crucial for understanding intention and action execution [3], [4].

The growing interest in head-mounted display technologies like smart glasses has opened up opportunities for egocentric sensing. Although still in the early stages of adoption, their immense potential has catalyzed research into egocentric vision in indoor settings, enabling the development of general-purpose ego-based assistants [5], [6], [7]. Extending these capabilities to dynamic urban environments remains an important yet underexplored challenge. Addressing this gap is key to enabling proactive AI systems for pedestrian assistance, including navigation support [8], [9], [3] and specialized aids for visually impaired users [10].

Existing work on egocentric prediction has focused on motion forecasting. Prior studies have explored egocentric

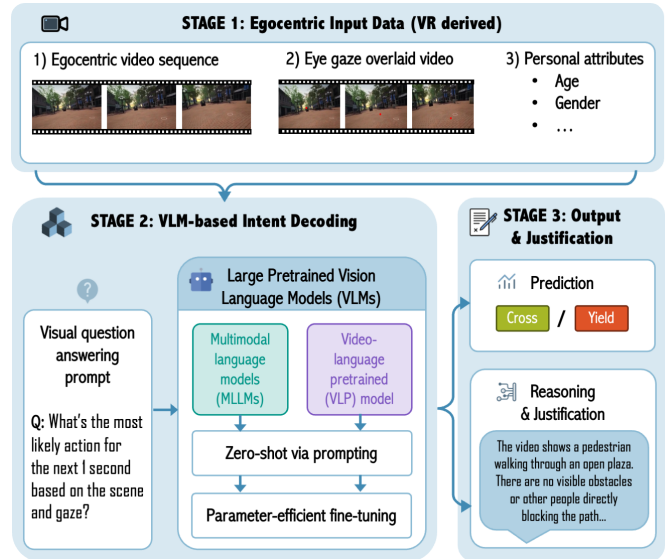


Fig. 1: A pipeline of our approach.

trajectory prediction in crowded spaces [11] or sidewalks [12], [9], typically relying on specialized deep learning architectures such as convolutional neural networks [12], diffusion models [3], and transformer-based approaches [8]. These models also depend on diverse inputs, ranging from standard views [9], [8], stereo images [13], to panoramic videos [3]. This reliance on specialized pipelines limits their ability to generalize across complex real-world urban scenarios.

To address these limitations, we explore the potential of pre-trained Vision-Language Models (VLMs) for egocentric intention prediction by reformulating the problem as a visual question answering (VQA) task. We investigate a fundamental question: **Can we construct a generalist egocentric model for pedestrian safety from urban monocular first-person videos?** As a proof-of-concept for egocentric urban intelligence, we focus on pedestrian crossing intention prediction. More specifically, we first evaluate the zero-shot reasoning capabilities of two state-of-the-art VLMs, assessing whether they inherently possess the perceptual and reasoning skills required for urban navigation. We then analyze the impact of parameter-efficient fine-tuning in adapting these models for traffic-specific prediction tasks. Finally, we examine how integrating eye-gaze signals and personal attributes further enhances predictive performance. An overview of our framework is presented in Fig. 1.

<sup>1</sup>Danya Li and Rico Krueger are with Department of Technology, Management, and Economics, Technical University of Denmark, Denmark danli@dtu.dk, rickr@dtu.dk

<sup>2</sup>Xiang Su is with Department of Computer Science and Department of Agricultural Sciences, University of Helsinki, Finland xiang.su@helsinki.fi

<sup>3</sup>Yan Feng is with Department of Transport and Planning, Delft University of Technology, Netherlands y.feng@tudelft.nl

## II. METHODOLOGY

Our objective is to predict the pedestrian’s crossing intention over a 1-second future horizon, given a 2-second egocentric video observation. We formulate this task as a VQA problem to leverage the rich world knowledge embedded in large pre-trained models, enabling reasoning beyond purely data-driven supervision.

### A. Model selection

We consider models from two complementary VLM families:

- 1) **Multi-modal Large Language Models (MLLMs):** These models process video inputs and are pre-trained on massive diverse datasets to acquire broad world knowledge. To balance performance and computational efficiency, we adopt the lightweight Qwen3-VL-2B-Instruct model [14] (hereafter referred to as *Qwen*), which has demonstrated promising performance in egocentric understanding [15].
- 2) **Video-Language pre-trained Models (VLPs):** VLPs represent a de-facto approach for video–text tasks, with explicit pretraining on egocentric videos and question answering data to achieve cross-modal alignment [16]. However, their training data is predominantly biased toward indoor activities [17], [5]. We select Ground-VQA [18] (hereafter referred to as *VLP*), a state-of-the-art model with dual video and language encoders and cross-modal fusion.

### B. Prompt design for VLMs

We evaluate both standard prompting and Chain-of-Thought (CoT) prompting [19]:

- **Standard prompt:** Our preliminary experiments show that off-the-shelf VLMs often produce biased or inconsistent outputs. To enforce a structured response and fairly assess model capabilities, we design a constrained prompt: “What is your most likely action in the next 1 second based on what you saw in the egocentric video for the past 2 seconds? Choose one option: (A) cross (B) yield.”
- **CoT prompt:** First, we use “Let’s think step by step” [20] to activate the chain-of-thought process. To further encourage a multi-step reasoning, we use “Analyze the egocentric video. First, describe the visual elements related to the crossing task. Second, Evaluate the attention presence, perceived proximity, and perceived risk. Thirdly, explain the logic connecting these elements. Finally, provide the final answer. Output format: Reasoning: [maximum 5 sentences about your reasoning]. Answer: [just the letter and option].”

### C. Fine-tuning strategies

We employ supervised fine-tuning (SFT) using annotated triplets (video, question  $\rightarrow$  answer) to align models with egocentric traffic reasoning. To maintain computational efficiency while preserving pre-trained knowledge, we adopt

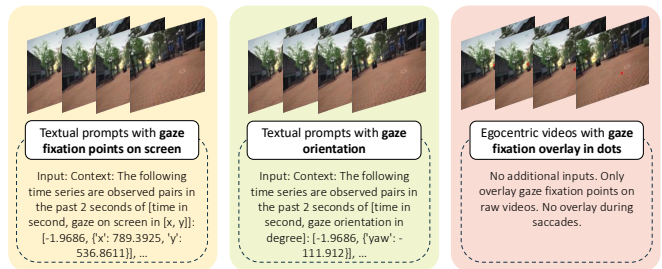


Fig. 2: Eye gaze-guided prompts.

Low-Rank Adaptation (LoRA) [21], which injects trainable low-rank matrices into transformer layers.

We investigate which components require domain adaptation by selectively tuning: 1) language encoders, 2) cross-modal fusion modules, and 3) both language and cross-modal component jointly. The vision encoder remains frozen, as it serves as a pre-trained backbone for visual feature extraction in VLPs [18].

### D. Representing contextual information

We incorporate fine-grained eye-gaze signals to precisely reveal pedestrians’ line of sight [22]. Following [15], [23], we design three gaze-guided representations, including 1) textual prompts with gaze fixation coordinates on screen, 2) textual prompts with gaze orientation (in degrees), and 3) visual overlay of gaze fixation points in red dots on video frames. An illustration is shown in Fig. 2. No overlay is provided during saccades. In addition to gaze information, we also consider personal attributes. To model this, we include demographic information in our prompts. For example: “You are a female of 27 years old, with Master degree or equivalent education level. Your dominant hand in daily life is left hand. Your walk frequency in daily life is every day...”.

## III. EXPERIMENTAL SETUP

Due to the limited availability of real-world egocentric data for pedestrian intention prediction in urban environments, we utilize a VR-based dataset from [23]. This dataset captures egocentric pedestrian navigation alongside automated shuttles in a shared space. It provides synchronized egocentric videos, eye gaze tracking, demographic profiles, and movement trajectories.

To ensure data quality, we retain only critical interactions occurring prior to the intersection crossing point. The data is segmented into 2-second observation windows and 1-second prediction horizons with a 0.5-second stride, yielding 6,047 QA samples. To prevent data leakage and evaluate robust generalization, we partition the dataset at the participant-level into training, validation, and test sets using a 6:1:3 split.

**Labeling** We define the binary intention labels—“cross” and “yield”—based on pedestrian kinematic behavior within the 1-second future horizon. Labels are determined by the duration for which the pedestrian speed exceeds a predefined crossing threshold [23]. Specifically, if the speed remains above the threshold for the majority of the horizon, the sample is labeled “cross”; otherwise, it is labeled “yield”.

This results in 2,486 crossing and 3,561 yielding samples. To mitigate the resulting mild class imbalance, we apply random under-sampling to the majority class during subsequent training.

**Baselines** We establish a video-only deep learning baseline to forecast crossing intention. Frame-level visual features are extracted using a pre-trained CLIP backbone [24], and the resulting feature sequence is processed by a transformer encoder followed by a classification head. This baseline is denoted as *CLIP+Transformer* hereafter. The model is trained for up to 100 epochs using the Adam optimizer (learning rate 0.001, batch size 64) with early stopping to prevent overfitting. In addition, we consider two other simple baselines: (1) always predicting the majority class, and (2) random guessing. All baseline results are reported in Tab. I.

**Metrics** We evaluate performance using accuracy, following standard practice in close-ended question answering (CloseQA) tasks [25], [15]. To account for class imbalance, we additionally report the macro F1 score. To mitigate choice-order bias, we randomize the associate between answer option and intention labels. Furthermore, we set the sampling temperature to 0 for deterministic generation, thereby isolating the model reasoning capabilities from stochastic variation.

## IV. RESULTS

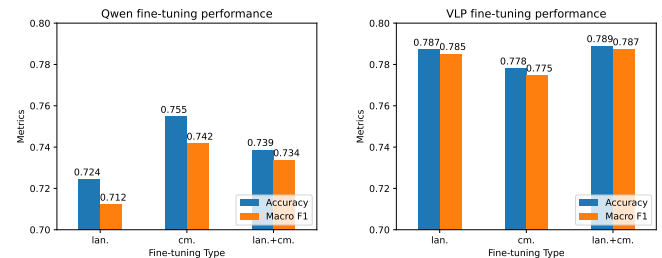
### A. Zero-shot and fine-tuning performance when using solely egocentric videos

We first evaluate the zero-shot capabilities of the Qwen and VLP models against the baseline in Tab. I. Under standard prompting, both models exhibit severe class bias and fail to outperform the baseline. Notably, the Qwen model shows a tendency toward the “cross” class (accuracy: 0.436), whereas VLP predicts almost exclusively “yield” (accuracy: 0.567). Introducing CoT prompting for Qwen model mitigates this bias and enhances generalization (accuracy: 0.558), albeit at a substantially higher computational cost, with latency increasing from under 0.1 seconds to approximately 1.5 seconds. We further present examples of Qwen’s reasoning outputs on the left side of Fig. 4. While the model demonstrates correct recognition of visual elements, it can misinterpret higher-level dynamic aspects. For instance, in Fig. 4a, the model correctly identifies the vehicle and the pedestrian’s goal, but incorrectly infers that the shuttle is moving forward and has already passed the pedestrian, whereas the automated shuttle is in fact stopped in front of the pedestrian. Overall, the zero-shot results indicate that models pre-trained on general world knowledge or predominantly indoor egocentric data has limited capability in understanding dynamic urban traffic contexts.

Building on this observation, we adapt the pre-trained models using parameter-efficient fine-tuning while preserving their general knowledge. After adaptation, both models outperform the baseline: Qwen achieves an accuracy of 0.755 while VLP reaches 0.789, compared to the CLIP+Transformer baseline accuracy of 0.727. This suggests

TABLE I: Performance of baselines, zero-shot, and fine-tuning using only egocentric videos. **Bold** number denotes the best performance in the group.

Group	Model	Prompt	Acc.	M-F1	Time (s)
Baselines	Majority	—	0.567	0.362	—
	Random	—	0.500	0.497	—
	Transformer	—	<b>0.727</b>	<b>0.724</b>	—
Zero-shot	Qwen	Standard	0.436	0.323	0.08
	VLP	Standard	<b>0.567</b>	0.378	0.007
	Qwen	CoT (simple)	0.550	0.547	1.32
	Qwen	CoT (multi)	0.558	<b>0.554</b>	1.61
Fine-tuned	Qwen	Standard	0.755	0.742	0.08
	VLP	Standard	<b>0.789</b>	<b>0.787</b>	0.007



(a) Qwen fine-tuning result.

(b) VLP fine-tuning result.

Fig. 3: Evaluation of fine-tuning strategies. “lan.” denotes language modules, “cm.” denotes cross-modality ones, “lan.+cm.” denotes both.

that language pretraining contributes to improved cross-modality understanding. However, a distributional gap remains between the models’ pretraining data and the target urban domain. Furthermore, the optimal fine-tuning strategy varies across architectures, as shown in Fig. 3. Updating only cross-modal modules is sufficient for Qwen, whereas the VLP model performs best when both language and cross-modal components are fine-tuned. Interestingly, VLP consistently shows greater improvement than Qwen in egocentric intention prediction, indicating higher adaptability to the target setting, potentially due to its egocentric pretraining.

### B. Impact of contextual information

We first evaluate the influence of contextual information via zero-shot prompting. For the VLP model, incorporating additional context induces hallucinated outputs in approximately 5% of cases; therefore, it is excluded from this comparison. In contrast, context-guided prompts with Qwen yields marginal improvements across different prompting methods (Tab. II), and more complex multi-step reasoning prompts do not provide further gains. Qualitative results in Fig. 4 suggest that gaze information may enhance situational awareness to some extent, though indirectly. Specifically, the model’s generated reasoning does not explicitly reference the gaze data; however, as seen in Fig. 4b, the model utilizing the gaze overlay (i.e., the middle panel) can more accurately project future states and potential events. It correctly infers that an ongoing interaction with others might be occurring, making it prudent to wait. This implicit utilization contrasts



**Model 1: Egocentric only, CoT (simple)**

**Q:** You are a pedestrian in a shared space environment where pedestrians and vehicles share the same area without strict traffic rules. You need to get to the white circle safely. Based on what you saw in the egocentric video for the past 2 seconds, what is your most likely action in the next 1 second? Choose one option: (A) cross (B) yield. Let's think step by step. Output format: Reasoning: [maximum 5 sentences about your reasoning]. Answer: [just the letter and action].

**A:** cross ✓

**Reasoning:** In the previous frame, the vehicle was approaching from behind and moving forward, while I am standing still. The vehicle has now moved further away, indicating it may have completed its approach or passed me. Given that there is no visible obstacle in my path and the vehicle has already left the immediate vicinity, the safest course of action would be to continue walking forward as long as it remains clear. Since the white circle represents a destination point and I'm not currently obstructed, crossing seems appropriate.

**Model 2: Egocentric with gaze overlay, CoT (simple)**

**Q:** ...

**A:** yield x

**Reasoning:** In the previous frame, a vehicle was approaching from behind, and it appears that there is no immediate obstacle or danger in the path of the vehicle. The pedestrian has not yet crossed the street, and the vehicle seems to be moving at a safe distance. Given that the pedestrian needs to reach the white circle, which is located ahead, the safest course would be to wait until the vehicle passes before crossing. However, since the vehicle is still present and could potentially obstruct the path, yielding is necessary to ensure safety.

**Model 3: Egocentric + Demographics, CoT (simple)**

**Q:** Context: You are a female of 33 years old, with Master degree or equivalent education level. Your dominant hand in daily life is right hand. Your walk frequency in daily life is every day. Your VR frequency is sometimes. Your familiarity with automated shuttles is moderate familiar, and your previous experience with automated shuttles is little experience....

**A:** cross ✓

**Reasoning:** The automated shuttle has just passed through the scene, and it appears that there is no immediate obstacle or danger in its path. Since I am a pedestrian in a shared space, my primary concern is ensuring safety while navigating around potential hazards. Given that the shuttle has already moved out of the way, I should proceed cautiously but not stop or wait unnecessarily. In this context, crossing the street would be the safest choice as long as I maintain awareness of any possible movement from other vehicles or pedestrians.

(a) Failure case: Misinterpretation of vehicle dynamics. Both models fail to recognize the automated shuttle has stopped for the pedestrian. Without a clear understanding of the vehicle's intent, the models incorrectly estimate the safety of the path.



**Model 1: Egocentric only, CoT (simple)**

**Q:** You are a pedestrian in a shared space environment where pedestrians and vehicles share the same area without strict traffic rules. You need to get to the white circle safely. Based on what you saw in the egocentric video for the past 2 seconds, what is your most likely action in the next 1 second? Choose one option: (A) cross (B) yield. Let's think step by step. Output format: Reasoning: [maximum 5 sentences about your reasoning]. Answer: [just the letter and action].

**A:** cross x

**Reasoning:** In the previous two seconds, I observed that there was no immediate threat of collision with any vehicle or obstacle. The path ahead appears clear, and I am positioned at a safe distance from any potential hazards. Since my goal is to reach the white circle, which is located further down the street, I should proceed cautiously but not stop or turn back. Given the lack of visible obstacles or other pedestrians blocking my path, it would be reasonable to continue moving forward without yielding or stopping.

**Model 2: Egocentric with gaze overlay, CoT (simple)**

**Q:** ...

**A:** yield ✓

**Reasoning:** In the previous frame, I was standing at a red circle, which indicates a stop sign or a point of no entry. The presence of other pedestrians suggests that there might be an ongoing interaction with others who may also be moving through this area. Since I am currently in a shared space and have not yet crossed the street, it would be prudent to wait until all nearby individuals have passed before proceeding. Given the context of the scene, my immediate action should be to yield to any approaching pedestrians or vehicles, ensuring safety and compliance with the shared space rules.

**Model 3: Egocentric + Demographics, CoT (simple)**

**Q:** Context: You are a male of 29 years old, with Master degree or equivalent education level. Your dominant hand in daily life is right hand. Your walk frequency in daily life is every day. Your VR frequency is sometimes. Your familiarity with automated shuttles is moderate familiar, and your previous experience with automated shuttles is some experience....

**A:** yield ✓

**Reasoning:** I am observing a pedestrian walking across a brick-paved plaza in front of a row of shops. The person is moving from left to right, approaching the white circle at the center of the plaza. Given that this is a shared space with other pedestrians and no visible traffic signals or barriers, it is important to be cautious when crossing. Since there is no indication of any immediate danger or obstacle, and considering my own movement pattern as a pedestrian who walks frequently, I would typically proceed cautiously but not stop unless necessary. However, since the white circle appears to be an intended destination or point of interest, and given the context of needing to reach it, I might consider yielding if another pedestrian is present or if there's potential conflict.

(b) Success case: Gaze-informed hazard perception. The gaze-overlaid model correctly anticipates potential hazards. This shows gaze signals helps mimic human-like situational awareness.

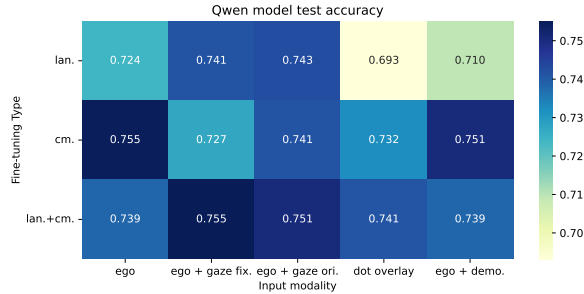
Fig. 4: Reasoning analysis of Qwen predictions with CoT. A comparison between egocentric-only input (left) and gaze-augmented egocentric input (right). Highlights indicate reasoning correctness: red (incorrect), green (correct), yellow (partially correct), and blue (context relevant).

with other modalities like personal context, where the model explicitly relates attributes like high walking frequency to more cautious behavior, leading to a correct prediction. Overall, these findings indicate that current pre-trained models struggle to effectively leverage context-guided prompts through zero-shot prompting alone.

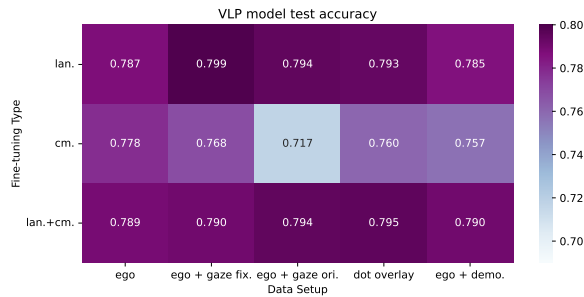
To assess whether models can explicitly learn to utilize this context, we fine-tune both models using context-guided QA pairs (Fig. 5). Incorporating eye gaze data during fine-tuning yields mixed results. While the VLP model show slight performance gains with certain gaze representations, overall gains relative to the egocentric-only baseline remain limited

TABLE II: Impact of contextual information on pedestrian intention prediction via zero-shot prompting using Qwen model. **Bold** number denotes the best performance in the group.

Input modality	Standard prompt		CoT prompt (simple)		CoT prompt (multi)	
	Acc.	M-F1	Acc.	M-F1	Acc.	M-F1
Egocentric only	0.436	0.323	0.550	0.547	0.558	0.554
Ego + Gaze fixation	<b>0.475</b>	<b>0.456</b>	0.560	0.557	0.547	0.545
Ego + Gaze orientation	0.457	0.438	0.550	0.549	0.564	0.562
Ego with dot overlay	0.453	0.375	<b>0.584</b>	<b>0.567</b>	<b>0.594</b>	<b>0.570</b>
Ego + Demographics	0.464	0.411	0.579	0.567	0.579	0.543



(a) Qwen model performance



(b) VLP model performance

Fig. 5: Fine-tuned models with different context-guided prompts using different fine-tuning strategies.

for both models. We speculate that this limited effect stems from insufficient behavioral diversity in the dataset, particularly the lack of scenarios involving distracted pedestrians, where gaze information would play a more decisive role.

## V. CONCLUSION

In this study, we explored the feasibility of VLMs to decode pedestrian crossing intentions from egocentric vision. Our findings highlight both current limitations and promise of VLMs: while zero-shot models remain constrained in reasoning about dynamic urban environments, parameter-efficient fine-tuning enables significant performance gains over specialized baselines. Integrating eye-gaze data and personal attributes demonstrates the value of multi-modal context in enhancing predictive accuracy.

Beyond demonstrating empirical improvements, our work underscores the broader potential of VLMs in traffic safety applications. However, qualitative analysis suggests that

current models still lag behind human-level understanding, particularly in complex or ambiguous scenarios. Future research should investigate large-scale traffic-specific pretraining, richer multi-modal representations, and more sophisticated reasoning mechanisms to advance VLM-driven perception for safe and interpretable urban mobility.

## REFERENCES

- [1] C. Zhang and C. Berger, "Pedestrian Behavior Prediction Using Deep Learning Methods for Urban Scenarios: A Review," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–23, 2023.
- [2] F.-G. Landry and M. A. Akhlofi, "Predicting Pedestrian Crossing Intention in Autonomous Vehicles: A Review," *Neurocomputing*, vol. 618, p. 129105, Feb. 2025.
- [3] W. Wang, C. K. Liu, and M. Kennedy, "EgoNav: Egocentric Scene-aware Human Trajectory Prediction," Aug. 2024, arXiv:2403.19026 [cs].
- [4] Y. He, Y. Huang, G. Chen, L. Lu, B. Pei, J. Xu, T. Lu, and Y. Sato, "Bridging Perspectives: A Survey on Cross-view Collaborative Intelligence with Egocentric-Exocentric Vision," *International Journal of Computer Vision*, vol. 134, no. 2, p. 62, Jan. 2026.
- [5] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18995–19012.
- [6] C. Plizzari, G. Goletto, A. Furnari, S. Bansal, F. Ragusa, G. M. Farinella, D. Damen, and T. Tommasi, "An outlook into the future of egocentric vision," *International Journal of Computer Vision*, vol. 132, no. 11, pp. 4880–4936, 2024.
- [7] J. Yang, S. Liu, H. Guo, Y. Dong, X. Zhang, S. Zhang, P. Wang, Z. Zhou, B. Xie, Z. Wang *et al.*, "Egolife: Towards egocentric life assistant," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 28885–28900.
- [8] Z. Qiu, Z. Liu, W. Niu, T. Bhattacharjee, and S. Kalantari, "EgoCogNav: Cognition-aware Human Egocentric Navigation," Nov. 2025, arXiv:2511.17581 [cs].
- [9] B. Pan, A. W. Harley, F. Engelmann, C. K. Liu, and L. J. Guibas, "Lookout: Real-world humanoid egocentric navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 24977–24988.
- [10] Y. Haghighi, C. Demonsant, P. Chalimourdas, M. T. Naeini, J. K. Munoz, B. Bacca, S. Suter, M. Gani, and A. Alahi, "Heads-up: Head-mounted egocentric dataset for trajectory prediction in blind assistance systems," *arXiv preprint arXiv:2409.20324*, 2024.
- [11] J. Qiu, L. Chen, X. Gu, F. P.-W. Lo, Y.-Y. Tsai, J. Sun, J. Liu, and B. Lo, "Egocentric human trajectory forecasting with a wearable camera and multi-modal fusion," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8799–8806, 2022.
- [12] K. K. Singh, K. Fatahalian, and A. A. Efros, "KrishnaCam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2016, pp. 1–9.
- [13] H. S. Park, J.-J. Hwang, Y. Niu, and J. Shi, "Egocentric future localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4697–4705.
- [14] Q. Team, "Qwen3 technical report," 2025. [Online]. Available: <https://arxiv.org/abs/2505.09388>
- [15] T. Peng, J. Hua, M. Liu, and F. Lu, "In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting," *arXiv preprint arXiv:2509.07447*, 2025.
- [16] S. Pramanick, Y. Song, S. Nag, K. Q. Lin, H. Shah, M. Z. Shou, R. Chellappa, and P. Zhang, "Egovlpv2: Egocentric video-language pre-training with fusion in the backbone," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 5285–5297.
- [17] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling Egocentric Vision: The EPIC-KITCHENS Dataset," Jul. 2018, arXiv:1804.02748 [cs].
- [18] S. Di and W. Xie, "Grounded question-answering in long egocentric videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12934–12943.

- [19] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
- [20] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.
- [21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, “Lora: Low-rank adaptation of large language models.” *Iclr*, vol. 1, no. 2, p. 3, 2022.
- [22] X. Li, H. Qiu, L. Wang, H. Zhang, C. Qi, L. Han, H. Xiong, and H. Li, “Challenges and trends in egocentric vision: A survey,” *Machine Intelligence Research*, vol. 23, no. 1, pp. 1–33, 2026.
- [23] D. Li, Y. Feng, and R. Krueger, “Eye Gaze-Informed and Context-Aware Pedestrian Trajectory Prediction in Shared Spaces with Automated Shuttles: A Virtual Reality Study,” Mar. 2026, arXiv:2603.19812 [cs].
- [24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 8748–8763.
- [25] A. Patel, V. Chitalia, and Y. Yang, “Advancing Egocentric Video Question Answering with Multimodal Large Language Models,” Apr. 2025, arXiv:2504.04550 [cs].