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

Danya Li[1], Xiang Su[2], Yan Feng[3], and Rico Krueger[1]

[1]Department of Technology, Management, and Economics, Technical University of Denmark, Denmark
[2]Department of Computer Science and Department of Agricultural Sciences, University of Helsinki, Finland
[3]Department of Transport and Planning, Delft University of Technology, Netherlands

Short summary

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 (e.g., vehicle dynamics). 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.

Keywords: Pedestrian-vehicle interaction; egocentric vision; intention prediction; vision-language model; urban mobility.

1 Introduction

Ensuring pedestrian safety in increasingly complex urban environments is critically dependent on accurate intent anticipation. Traditional approaches to this challenge have predominantly relied on exocentric perspectives, such as vehicle-mounted or surveillance cameras, which offer a stable and global view of the scene (Zhang & Berger, 2023; Landry & Akhloufi, 2025). However, these external viewpoints often fail to capture the pedestrian's first-person perception and fine-grained behavioral cues. In contrast, egocentric (first-person) vision offers direct access to the pedestrian's line of sight, providing rich information that is crucial for understanding intention and action execution (Wang et al., 2024; He et al., 2026).

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 (Grauman et al., 2022; Plizzari et al., 2024; Yang et al., 2025). 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 (Z. Qiu et al., 2025; Pan et al., 2025; Wang et al., 2024) and specialized aids for visually impaired users (Haghighi et al., 2024).

Existing work on egocentric prediction has primarily focused on motion forecasting. Prior studies have explored egocentric trajectory prediction in crowded spaces (J. Qiu et al., 2022) or sidewalks (Singh et al., 2016; Pan et al., 2025), typically relying on specialized deep learning architectures such as convolutional neural networks (Singh et al., 2016), diffusion models (Wang et al., 2024), and transformer-based approaches (Z. Qiu et al., 2025). These models also depend on diverse inputs, ranging from standard views (Pan et al., 2025; Z. Qiu et al., 2025), stereo images (Park et

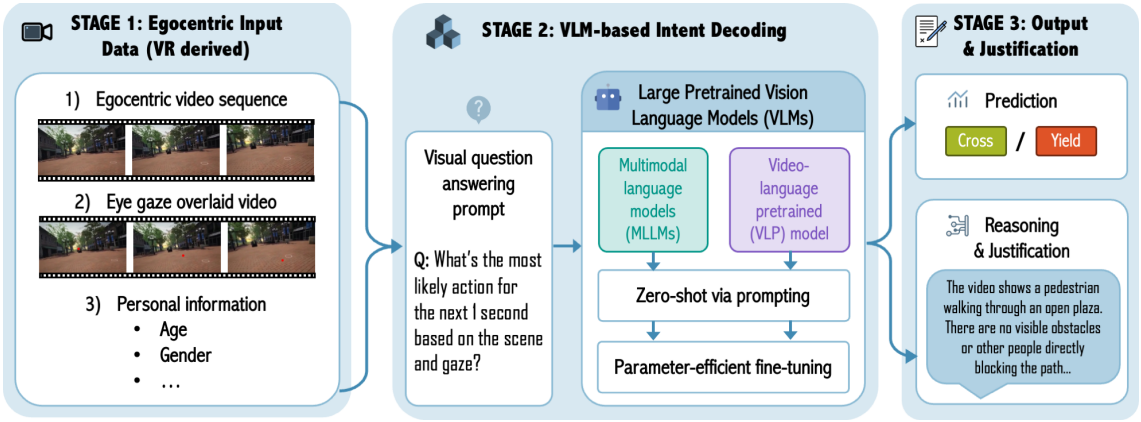


Figure 1: A pipeline of our approach.

al., 2016), to panoramic videos (Wang et al., 2024). 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.

2 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. By doing so, we leverage the rich world knowledge embedded in large pretrained models, enabling reasoning beyond purely data-driven supervision.

Model selection

We consider models from two complementary VLM families:

1) **Multi-modal Large Language Models (MLLMs):** These models process video inputs and are pretrained on massive diverse datasets to acquire broad world knowledge. To balance performance and computational efficiency, we adopt the lightweight Qwen3-VL-2B-Instruct model (Team, 2025) (hereafter referred to as *Qwen*), which has demonstrated promising performance in egocentric understanding Peng et al. (2025).
2) **Video-Language Pretrained 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 (Pramanick et al., 2023). However, their training data is predominantly biased toward indoor activities Damen et al. (2018); Grauman et al. (2022). We select GroundVQA (Di & Xie, 2024) (hereafter referred to as *VLP*), a state-of-the-art model with dual video and language encoders and cross-modal fusion.

Prompt design for VLMs

We evaluate both standard prompting and Chain-of-Thought (CoT) prompting (Wei et al., 2022):

- **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 capabil-

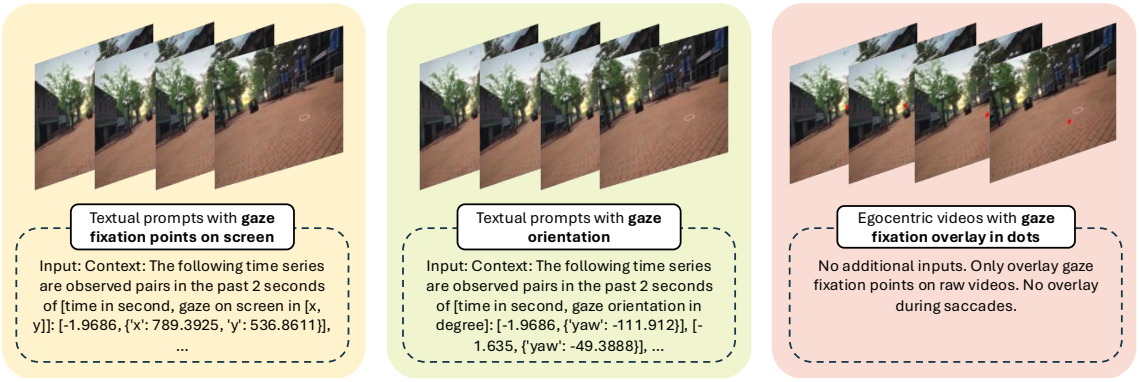


Figure 2: Eye gaze-guided prompts.

ities, 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" (Kojima et al., 2022) 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]."

Fine-tuning strategies

We employ supervised fine-tuning (SFT) using annotated triplets (video, question → answer) to align the pre-trained models with egocentric traffic reasoning. To maintain computational efficiency while preserving pretrained knowledge, we adopt Low-Rank Adaptation (LoRA) (Hu et al., 2022), 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 pretrained backbone for visual feature extraction in VLPs (Di & Xie, 2024).

Representing contextual information

We incorporate fine-grained eye-gaze signals to precisely reveal pedestrians' line of sight (X. Li et al., 2026). Following Peng et al. (2025); D. Li et al. (2026), 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...".

3 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 D. Li et al. (2026). This dataset captures egocentric pedestrian navigation alongside automated shuttles in a shared space. It provides synchronized egocentric videos, eye gaze tracking, demographics, 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

Table 1: Performance of baselines, zero-shot, and fine-tuning using only egocentric videos. **Bold** number denotes the best performance in the group.

Model group	Model	Prompt	Acc. (↑)	M-F1 (↑)	Inf. time (s)
Baselines	Majority class	—	0.567	0.362	—
	Random guess	—	0.500	0.497	—
	CLIP+Transformer	—	**0.727**	**0.724**	—
Zero-shot	Qwen	Standard	0.436	0.323	0.08
	VLP	Standard	**0.567**	0.378	0.007
	Qwen	CoT (simple)	0.550	0.547	1.32
	Qwen	CoT (multi)	0.558	**0.554**	1.61
Fine-tuned	Qwen	Standard	0.755	0.742	0.08
	VLP	Standard	**0.789**	**0.787**	0.007

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 movements within the 1-second future horizon. Labels are determined by the relative duration for which the pedestrian speed exceeds a predefined crossing threshold (D. Li et al., 2026). 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.

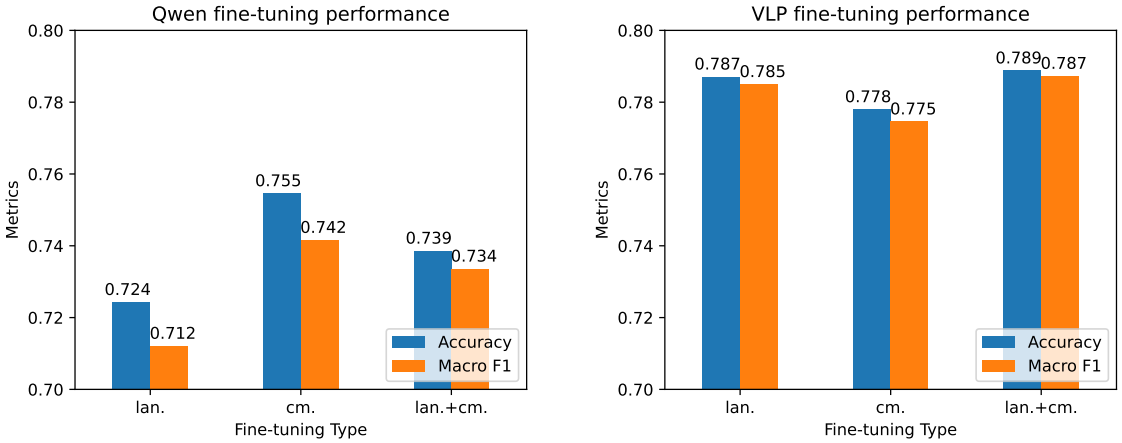
Baselines We establish a video-only deep learning baseline to forecast crossing intention. Frame-level visual features are extracted using a pretrained CLIP backbone (Radford et al., 2021), and the resulting feature sequence is processed by a transformer encoder followed by a classification head. This 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. 1.

Metrics We evaluate performance using accuracy, following standard practice in close-ended question answering (CloseQA) tasks (Patel et al., 2025; Peng et al., 2025). 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.

4 Results

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

We first evaluate the zero-shot capabilities of Qwen and VLP models against the baselines in Tab. 1. 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 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.



(a) Qwen model fine-tuning result.

(b) VLP model fine-tuning result.

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

Table 2: Impact of contextual information on pedestrian intention prediction via zero-shot prompting using Qwen model. **Bold** number denotes the best performance in the group.

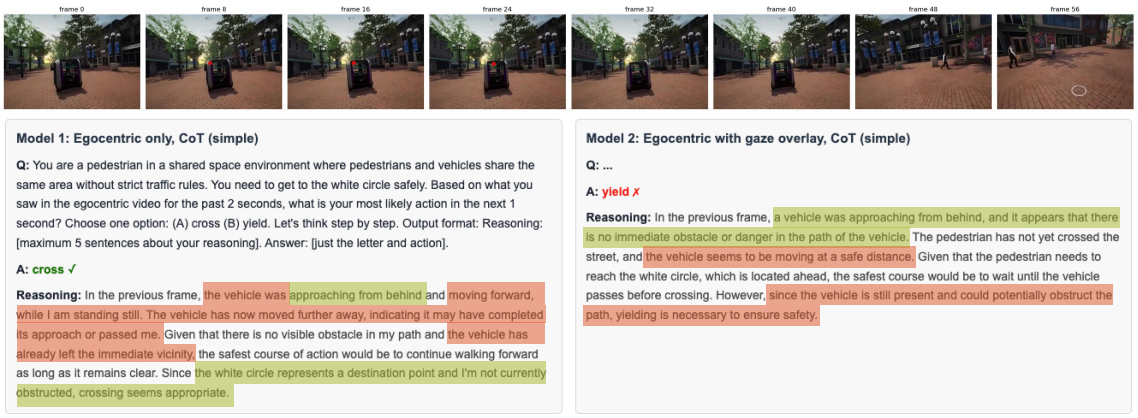
Input modality	**Standard prompt**		**CoT prompt (simple)**		**CoT prompt (multi)**	
	Acc. (↑)	M-F1 (↑)	Acc. (↑)	M-F1 (↑)	Acc. (↑)	M-F1 (↑)
Egocentric only	0.436	0.323	0.550	0.547	0.558	0.554
Ego + Gaze fixation	**0.475**	**0.456**	0.560	0.557	0.547	0.545
Ego + Gaze orientation	0.457	0.438	0.550	0.549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

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 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.

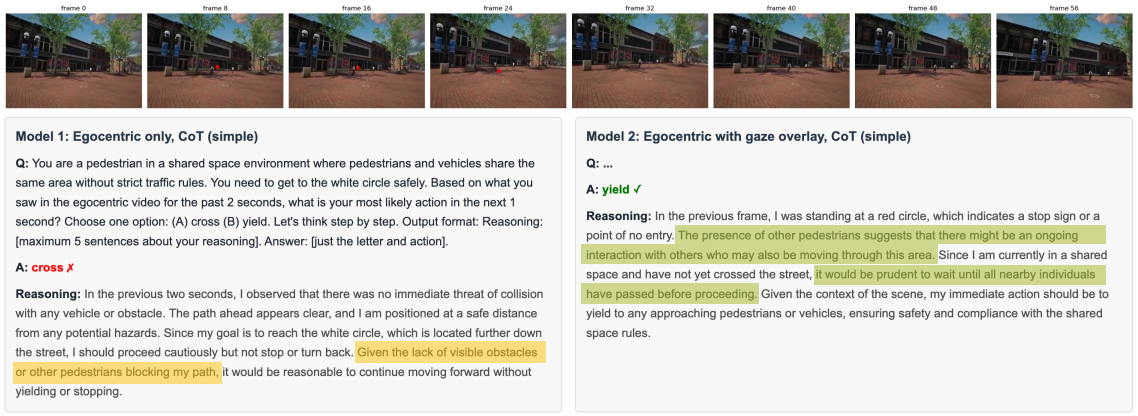
Impact of contextual information

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 only marginal improvements across different prompting methods (Tab. 2), 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. For example, in Fig. 4b, the model more accurately projects future states and potential events by inferring that an ongoing interaction with others might be occurring and thus it would be prudent to wait. 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, shown in Fig. 5. Incorporating eye gaze data during fine-tuning yields mixed results. While the VLP model show slight performance gains with certain gaze rep-



(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.



(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.

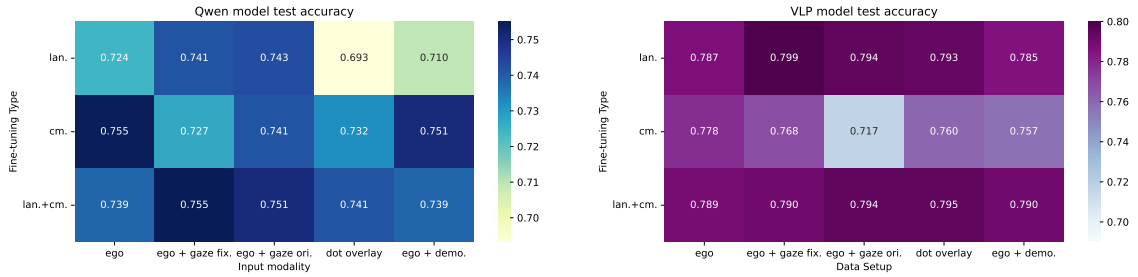
Figure 4: Reasoning analysis of Qwen predictions with CoT. A comparison between egocentric-only input (left) and gaze-augmented egocentric input (right). Highlights indicate reasoning correctness: red (incorrect), green (correct), and yellow (partially correct).

resentations, overall gains relative to the egocentric-only baseline remain limited for both models. We speculate that this limited effect stems from insufficient behavioral diversity in the dataset, particular the lack of scenarios involving distracted pedestrians, where gaze information would play a more decisive role.

5 Conclusion

In this work, we demonstrated the feasibility of leveraging VLMs to decode pedestrian crossing intentions formulated as a VQA task. Our zero-shot evaluations reveal that existing pretrained models has limited capabilities in understanding dynamic egocentric urban scenes. We further show that parameter-efficient fine-tuning effectively adapts these models to the traffic domain, enabling them to significantly outperform specialized baseline. Moreover, incorporating precise eye-gaze signals and personal attributes provides modest predictive enhancements, highlighting the potential of multi-modal contextual information for egocentric prediction.

Despite these results, this study represents an initial step toward VLM-driven egocentric urban perception. Our qualitative analysis indicates that current models still fall short of human-level scene understanding. Future work should focus on equipping VLMs with robust traffic reasoning capabilities through large-scale domain-specific pretraining on urban datasets. This will be essential for advancing toward general-purpose, proactive AI assistants for pedestrian safety.



(a) Qwen model performance

(b) VLP model performance

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

References

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., ... Wray, M. (2018, July). *Scaling Egocentric Vision: The EPIC-KITCHENS Dataset.* arXiv. (arXiv:1804.02748 [cs])

Di, S., & Xie, W. (2024). Grounded question-answering in long egocentric videos. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12934–12943).

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., ... others (2022). Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 18995–19012).

Haghighi, Y., Demonsant, C., Chalimourdas, P., Naeini, M. T., Munoz, J. K., Bacca, B., ... Alahi, A. (2024). Heads-up: Head-mounted egocentric dataset for trajectory prediction in blind assistance systems. *arXiv preprint arXiv:2409.20324*.

He, Y., Huang, Y., Chen, G., Lu, L., Pei, B., Xu, J., ... Sato, Y. (2026, January). Bridging Perspectives: A Survey on Cross-view Collaborative Intelligence with Egocentric-Exocentric Vision. *International Journal of Computer Vision*, *134*(2), 62.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... others (2022). Lora: Low-rank adaptation of large language models. *Iclr*, *1*(2), 3.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, 22199–22213.

Landry, F.-G., & Akhloufi, M. A. (2025, February). Predicting Pedestrian Crossing Intention in Autonomous Vehicles: A Review. *Neurocomputing*, *618*, 129105.

Li, D., Feng, Y., & Krueger, R. (2026, March). *Eye Gaze-Informed and Context-Aware Pedestrian Trajectory Prediction in Shared Spaces with Automated Shuttles: A Virtual Reality Study.* arXiv. (arXiv:2603.19812 [cs])

Li, X., Qiu, H., Wang, L., Zhang, H., Qi, C., Han, L., ... Li, H. (2026). Challenges and trends in egocentric vision: A survey. *Machine Intelligence Research*, *23*(1), 1–33.

Pan, B., Harley, A. W., Engelmann, F., Liu, C. K., & Guibas, L. J. (2025). Lookout: Real-world humanoid egocentric navigation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 24977–24988).

Park, H. S., Hwang, J.-J., Niu, Y., & Shi, J. (2016). Egocentric future localization. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4697–4705).

Patel, A., Chitalia, V., & Yang, Y. (2025, April). *Advancing Egocentric Video Question Answering with Multimodal Large Language Models.* arXiv. (arXiv:2504.04550 [cs])

Peng, T., Hua, J., Liu, M., & Lu, F. (2025). In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. *arXiv preprint arXiv:2509.07447*.

Plizzari, C., Goletto, G., Furnari, A., Bansal, S., Ragusa, F., Farinella, G. M., ... Tommasi, T. (2024). An outlook into the future of egocentric vision. *International Journal of Computer Vision*, *132*(11), 4880–4936.

Pramanick, S., Song, Y., Nag, S., Lin, K. Q., Shah, H., Shou, M. Z., ... Zhang, P. (2023). Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 5285–5297).

Qiu, J., Chen, L., Gu, X., Lo, F. P.-W., Tsai, Y.-Y., Sun, J., ... Lo, B. (2022). Egocentric human trajectory forecasting with a wearable camera and multi-modal fusion. *IEEE Robotics and Automation Letters*, *7*(4), 8799–8806.

Qiu, Z., Liu, Z., Niu, W., Bhattacharjee, T., & Kalantari, S. (2025, November). *EgoCogNav: Cognition-aware Human Egocentric Navigation.* arXiv. (arXiv:2511.17581 [cs])

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021, July). Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748–8763). PMLR.

Singh, K. K., Fatahalian, K., & Efros, A. A. (2016, March). KrishnaCam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1–9).

Team, Q. (2025). *Qwen3 technical report.* Retrieved from `https://arxiv.org/abs/2505.09388`

Wang, W., Liu, C. K., & Kennedy, M. (2024, August). *EgoNav: Egocentric Scene-aware Human Trajectory Prediction.* arXiv. (arXiv:2403.19026 [cs])

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... others (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824–24837.

Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., ... others (2025). Egolife: Towards egocentric life assistant. In *Proceedings of the computer vision and pattern recognition conference* (pp. 28885–28900).

Zhang, C., & Berger, C. (2023). Pedestrian Behavior Prediction Using Deep Learning Methods for Urban Scenarios: A Review. *IEEE Transactions on Intelligent Transportation Systems*, 1–23.