CP-VLM: Causal Prompting for Human Intention Inference with Vision–Language Models

Kazuki Osamura Hidetsugu Uchida Narishige Abe
FUJITSU LIMITED, Japan
{osamura.kazuki, u.hidetsugu, abe.narishige}@fujitsu.com

Abstract

Understanding human intentions from visual observations is essential for proactive human–robot collaboration. While recent vision–language models excel at scene description, they struggle with goal-directed reasoning beyond surface correlations. We propose CP-VLM, a framework that enhances intention inference through causally inspired structured prompting. Using efficient low-rank adaptation, we fine-tune the language decoder without architectural changes. Experiments on the JRDB-Social dataset demonstrate that CP-VLM improves F1 by +27.8% over baselines with comparable inference time, showing that causal structure enables deeper and more robust intention inference with minimal overhead.

1. Introduction

Understanding human intentions [12] is essential for autonomous systems in real-world environments, including service robots [14, 29] and human-centered AI [24]. Recent vision–language models (VLMs) [1, 2, 16, 22] achieve strong performance on descriptive tasks such as visual question answering (VQA) [7, 10]. However, practical deployment requires not only recognizing what occurs but also inferring why it happens, demanding causal reasoning. This is particularly important in domains such as eldercare [5, 17], customer service [4, 34], and collaborative work [28, 33], where anticipating human goals [21, 27] enables proactive assistance [6]. While current VLMs capture observable correlations, they struggle to uncover the underlying causes of behavior [10]. Bridging this gap requires integrating visual perception with causal reasoning [3, 23, 25, 26].

Scene graphs [13, 15, 19, 35] provide structured relational representations but remain observational and lack causal semantics. Consequently, surface actions (e.g., *standing in a queue*) are often misinterpreted as intentions, whereas causal reasoning reveals deeper goals (e.g., *ordering food*). As illustrated in Fig. 1, conventional approaches rely on observational mappings within VLMs, capturing surface correlations.

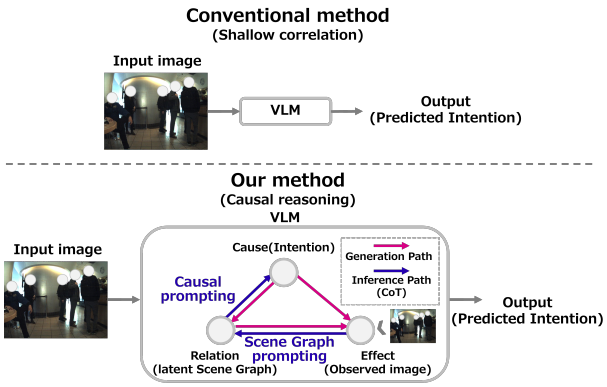

Figure 1. Comparison between conventional and proposed methods. Conventional methods rely on surface correlations in VLMs, whereas our method enables intention inference via structured prompting with latent scene-level and causal structures.

In contrast, our method employs structured prompting to induce latent scene and causal structures for goal-directed intention inference. We formulate intention understanding as reverse causal inference, reconstructing latent goals from observable behaviors through causally inspired prompting. This process is conceptually explained by a *Generative Path* inducing latent structures and an *Inference Path* performing causal Chain-of-Thought reasoning [30, 31].

Based on this formulation, we propose CP-VLM, which integrates causal prompting into VLMs for human intention inference. CP-VLM induces latent relational and causal structures through structured prompting, without explicit scene graph detectors, and is efficiently adapted via LoRA fine-tuning of the language decoder. Experiments on JRDB-Social [12, 18] show consistent F1 improvements over baseline prompting methods with comparable inference time.

The main contributions are summarized as follows:

- We introduce CP-VLM, a framework that enhances VLMs with structured and causally inspired prompting for human intention inference.

- We present an efficient LoRA-based adaptation strategy that preserves the base architecture.
- CP-VLM achieves a +27.8% F1 improvement on JRDB-Social over baseline prompting methods with minimal overhead.

2. Related Work

2.1. Vision-Language Models for Human Intention

Recent large-scale VLMs [1, 2, 16, 32, 36] achieve strong performance on descriptive tasks such as VQA, but remain limited in inferential reasoning, particularly for human intention understanding. This limitation stems from their reliance on surface correlations rather than causal dependencies. While prior work has explored intention recognition [12, 18] and action planning with VLMs [8, 9, 28, 33], our approach enhances reasoning purely through structured prompting without modifying model architectures.

2.2. Scene Understanding with Scene Graphs

Scene graphs [13, 15, 19, 20, 35] model entities and relations in structured scenes and have been widely applied to visual reasoning tasks. However, they are inherently observational, capturing spatial relations and interactions without causal or temporal semantics required for intention inference. Consequently, they describe visible actions but fail to represent the underlying goals driving behavior.

Instead of proposing a new graph representation, we leverage VLM reasoning to reinterpret latent relational structures as causal dependencies through structured prompting.

2.3. Causal Reasoning

Causal reasoning [3, 23, 25, 26] provides principled frameworks for understanding event dynamics and human behavior. Although causal modeling has improved robustness and interpretability in vision, integrating causal semantics into VLM-based intention inference remains challenging.

We address this by converting structured relational information into causally inspired prompts, enabling VLMs to perform cause–effect reasoning for deeper intention inference.

3. Method

Fig. 2 presents CP-VLM, which enhances intention inference by embedding causal structure into prompts and training the VLM with LoRA. Our method is organized around two complementary paths that reflect the causal semantics of the task: (i) a *Generative Path* that specifies how intentions give rise to observations through a latent scene graph, and (ii) an *Inference Path* that inverts this process to recover intentions from observations via a two-stage prompting design. In practice, the Inference Path is instantiated by *scene graph prompting* and *causal prompting*. The resulting prompt–answer pairs are then used for *LoRA-based Fine-tuning* of the language decoder.

3.1. Generative Path

The generative process proceeds in the forward causal direction, from human intention C to observable image I via the latent scene graph R:

$$p(I, R, C) = p(I \mid R, C)\, p(R \mid C)\, p(C), \tag{1}$$

Here, $p(C)$ denotes the prior distribution over intentions, $p(R \mid C)$ represents the conditional generation of scene graphs given the intention, and $p(I \mid R, C)$ models the likelihood of the image conditioned on both the intention and the scene graph. In this formulation, the scene graph R serves as a structural bottleneck connecting between latent intentions and observed visual evidence. This factorization corresponds to the *Generative Path* illustrated in Fig. 1, where intentions cause actions that produce the observable scene. In particular, the observable image I corresponds to the effect, generated as a consequence of the intention C through the scene graph R. By explicitly modeling this direction, we provide a causal interpretation of how human goals become observable behavior. *These equations are conceptual abstractions used to motivate prompt design, not models for causal identification or intervention.*

3.2. Inference Path

The goal of inference is to recover the underlying human intention C given an observed image I. This is achieved by integrating out the latent scene graph R:

$$p(C \mid I) = \int_R p(C \mid I, R)\, p(R \mid I), \tag{2}$$

Here, R acts as an intermediate representation that mediates between the observation and the intention. Operationally, this posterior is instantiated as a two-stage reasoning procedure: (1) inducing multiple plausible latent scene graphs conditioned on the observation (*scene graph prompting*, approximating $p(R \mid I)$), and (2) inferring the intention given both the observation and the causally structured scene graph (*causal prompting*, modeling $p(C \mid I, R)$). Importantly, the marginalization over R reflects that the model does not commit to a single explicit scene graph. Instead, through implicit Chain-of-Thought reasoning, it integrates over multiple latent relational structures, thereby avoiding the need to deterministically output one graph while still benefiting from causal structure in intention inference. This formulation aligns with the Inference Path illustrated in Fig. 1, where the model leverages Chain-of-Thought reasoning to invert the generative process from observation back to intention.

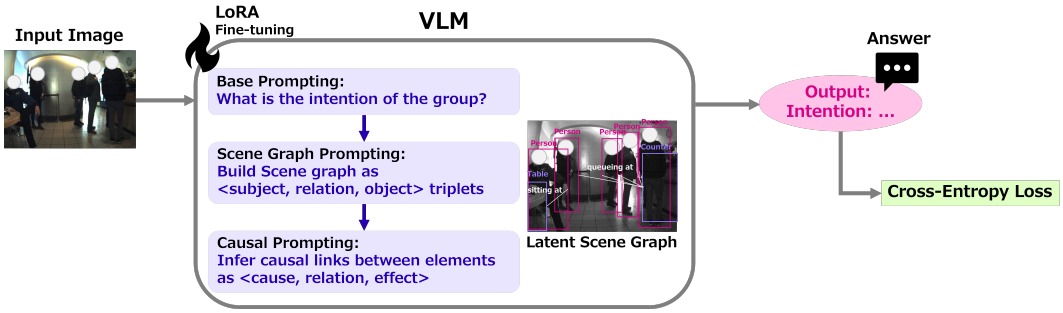

Figure 2. Overview of the proposed CP-VLM framework. An input image and causal prompting template are fed into the LoRA-tuned language decoder to infer human intentions. The model is trained with cross-entropy loss, encouraging structured, goal-oriented reasoning beyond surface-level correlations.

3.3. LoRA-based Fine-tuning

The VLM is fine-tuned using LoRA [11], injecting trainable low-rank matrices into the cross-attention layers of the language decoder (query and value projections) while keeping the original weights θ frozen. The objective is to maximize the likelihood of the ground-truth intention given the causally structured prompt. Let $\mathbf{y} = (y_1, \ldots, y_L)$ denote the target intention tokens, $\mathbf{z}$ the visual embedding, and $\mathbf{x}(p)$ the tokenized prompt. The cross-entropy loss is defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{t \in \mathcal{T}_{\text{ans}}} \log P_\theta\big(y_t \mid y_{<t}, \mathbf{z}, \mathbf{x}(p)\big), \tag{3}$$

where $\mathcal{T}_{\text{ans}}$ indexes the answer tokens on the assistant side. Supervision is applied only to these tokens, ensuring that the model learns intention predictions consistent with the causal structure encoded in the prompts. Intention inference is formulated as conditional language modeling.

4. Experiments and Results

4.1. Dataset and Evaluation Protocol

We evaluate CP-VLM on the JRDB-Social dataset [12], a large-scale benchmark for human intention understanding in social environments. The dataset consists of video sequences annotated with multiple intention labels across 11 categories: *wandering; discussing an object or matter; studying, writing, reading, or working; waiting for someone or something; socializing; excursion; attending a class, lecture, or seminar; eating or ordering food; commuting; navigating; and unknown* (Fig. 3). Following the official split [12], we use 2,589 training and 841 validation sequences, and report results on the validation set. We adopt mean F1 score as the primary metric and measure average inference time per frame to evaluate efficiency under the official protocol.

4.2. Implementation Details

We adopt Qwen2.5-VL-7B [1] as the base vision–language model and fine-tune it using LoRA [11] with all base parameters frozen. LoRA adapters are applied to the query and value projections of the decoder's cross-attention layers, while the vision encoder remains fixed. We use a LoRA rank of $r = 8$, scaling factor $\alpha = 8$, and dropout rate of 0.2. Training is performed for 5 epochs on a single NVIDIA RTX 2080 Ti GPU using AdamW with a learning rate of 2×10^{-5} and a batch size of 32, with gradient accumulation over 4 steps, cosine learning rate scheduling with warmup, and gradient clipping at 1.0. All experiments use a fixed random seed. During inference, we apply deterministic greedy decoding (num beams = 1, temperature = 0) with a maximum output length of 128 tokens. The prompt templates are provided in Fig. 2.

4.3. Ablation Study on Prompting Strategies

To validate the effectiveness of each component, an ablation study comparing three prompting strategies is conducted; the results are summarized in Table 2.

- **Base Prompting** [12]: A standard VQA-style prompt for direct intention querying, following the JRDB-Social baseline.
- **Scene Graph Prompting**: A prompt that encourages the model to reason about subject–relation–object triplets, capturing the structure of a scene.
- **Causal Prompting**: A prompt that further guides the model to infer latent cause–effect relationships underlying observed actions.

As shown in Table 2, LoRA-based fine-tuning consistently improves performance across all prompting strategies. Starting from base prompting with LoRA (F1: 31.0%), incorporating structured scene reasoning yields a substantial gain (scene graph prompting with LoRA: 55.4%, +24.4). Further introducing causal reasoning provides an additional improvement (causal prompting with LoRA: 58.8%, +3.4). Overall, CP-VLM achieves a +27.8% im-

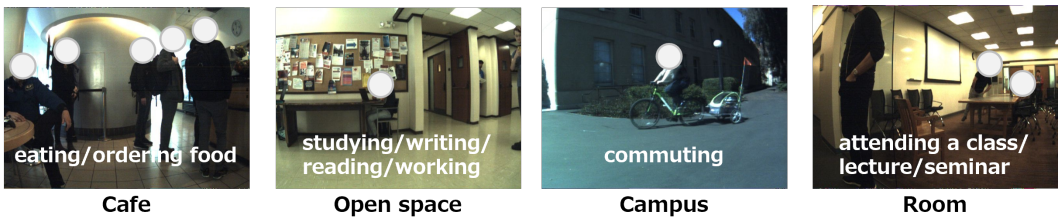

Figure 3. Example images from the JRDB-Social dataset [12], illustrating the ground-truth human intentions in diverse scenes.

Table 1. Qualitative comparison of outputs under different prompting strategies in diverse scenes. Inferences matching the ground-truth human intentions are highlighted in bold.

Scene	Prompting Strategies	Output	Inferred Intention
Café	Base Prompting [12]	–	waiting for something
	Scene Graph Prompting	people, standing in line, queue	waiting for something
	Causal Prompting	waiting for someone/something → standing in line	**eating/ordering food**
Open Space	Base Prompting [12]	–	**working**
	Scene Graph Prompting	person, sitting, on chair	**working**
	Causal Prompting	using computer at desk → sitting, on chair	**working**
Campus	Base Prompting [12]	–	**commuting**
	Scene Graph Prompting	people, riding, bicycle	**commuting**
	Causal Prompting	riding, bicycle → moving forward	**commuting**
Room	Base Prompting [12]	–	waiting for something
	Scene Graph Prompting	people, standing, in a line	waiting for something
	Causal Prompting	facing each other in pairs → standing in a line	discussing a matter

Table 2. Performance comparison of baseline and proposed prompting methods on the JRDB-Social dataset [12].

Method	F1 (%)	Time (s)
Base Prompting [12]	19.2	1.84
Base Prompting with LoRA	31.0	1.84
Scene Graph Prompting	22.6	1.90
Scene Graph Prompting with LoRA	55.4	1.90
Causal Prompting	24.3	1.92
Causal Prompting with LoRA(Ours, Full)	**58.8**	**1.92**

Table 3. Effect of causal prompting across different open-source VLMs (F1 score [%] in a zero-shot setting).

Model	Base	Scene Graph	Causal
LLaVA-1.5 [16]	1.5	6.1	13.7
Qwen2.5-VL-7B [1]	19.2	22.6	24.3
InternVL2.5-8B [22]	30.7	35.4	37.6

provement in F1 score over the base prompting with LoRA, while maintaining comparable inference time. These results highlight the importance of embedding a causal structure into prompts for enhancing intention inference in VLMs.

The generality of causal prompting is further validated in different VLMs. As shown in Table 3, both LLaVA-1.5 [16] and InternVL2.5-8B [22] achieve improvements with causal prompting, even in a zero-shot setting. These results indicate that causal prompting benefits diverse architectures. The main experiments focus on Qwen2.5-VL-7B as a widely adopted baseline, with LoRA fine-tuning highlighting the added benefits of lightweight adaptation.

4.4. Qualitative Analysis

To qualitatively evaluate reasoning, Table 1 compares predictions across prompting strategies in diverse scenes (Fig. 3). As prompting evolves from base and scene graph to causal prompting, outputs shift from shallow interpretations (e.g., *waiting*) to structured reasoning paths (e.g., *waiting → standing in line*) that support deeper intentions such as *ordering food*. Even when final intentions remain unchanged (e.g., *working* or *commuting*), causal prompting promotes reasoning beyond surface correlations, highlighting the benefit of causal structure. However, it is not always accurate: in the *Room* scene, it yields a plausible but incorrect intention (*discussing a matter*) instead of *attending class*, indicating sensitivity to ambiguous cues. Causal prompting typically follows a two-stage process from action relations to intention inference.

5. Conclusion

We propose CP-VLM, a framework that improves human intention inference in VLMs via causally inspired prompting with LoRA-based fine-tuning. CP-VLM moves beyond surface correlations by inducing interpretation of intentions grounded in scene-level relations, achieving a +27.8% F1 improvement on JRDB-Social. These results demonstrate that structured causal prompting is an effective strategy for human-centered AI.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl: Advanced multimodal large language model. In *arXiv preprint arXiv:2409.00001*, 2024. 1, 2, 3, 4

[2] Yuxiao Bai, Aixin Sun, Tingting Huang, Yichi Zhang, Lijun Wang, Qingxiao Guan, Xin Huang, Han Zhang, Xiaoguang Hu, and Jianjian Sun. Qwen-vl: A versatile vision-language model. In *arXiv preprint arXiv:2308.12966*, 2023. 1, 2

[3] Chris Baker, Joshua B. Tenenbaum, and Brian S. Gershman. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. In *Cognition*, 2017. 1, 2

[4] Hrishav Bakul Barua, Chayan Sarkar, Achanna Anil Kumar, Arpan Pal, and Balamuralidhar P. I can attend a meeting too! towards a human-like telepresence avatar robot to attend meeting on your behalf. In *arXiv preprint arXiv:2006.15647*, 2020. 1

[5] Patrick Benavidez, Mohan Kumar, and Sos Agaian. Design of a home multi-robot system for the elderly and disabled. In *10th System of Systems Engineering Conference (SoSE)*, 2015. 1

[6] Sera Buyukgoz, Jasmin Grosinger, Mohamed Chetouani, and Alessandro Saffiotti. Two ways to make your robot proactive: reasoning about human intentions, or reasoning about possible futures. In *arXiv preprint arXiv:2210.11584*, 2022. 1

[7] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali-x: On scaling up multilingual multimodal models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[8] Yongchao Chen, Jacob Arkin, Charles Dawson, Yang Zhang, Nicholas Roy, and Chuchu Fan. Autotamp: Autoregressive task and motion planning with llms as translators and checkers. In *arXiv preprint arXiv:2306.06531*, 2024. 2

[9] Yan Ding, Xiaohan Zhang, Chris Paxton, and Shiqi Zhang. Task and motion planning with large language models for object rearrangement. In *arXiv preprint arXiv:2303.06247*, 2023. 2

[10] Qiyue Gao, Xinyu Pi, Kevin Liu, Junrong Chen, Ruolan Yang, Xinqi Huang, Xinyu Fang, Lu Sun, Gautham Kishore, Bo Ai, Stone Tao, Mengyang Liu, Jiaxi Yang, Chao Jung Lai, Chuanyang Jin, Jiannan Xiang, Benhao Huang, Zeming Chen, David Danks, Hao Su, Tianmin Shu, Ziqiao Ma, Lianhui Qin, and Zhiting Hu. Do vision-language models have internal world models? towards an atomic evaluation. In *arXiv preprint arXiv:2506.21876*, 2025. 1

[11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *arXiv preprint arXiv:2106.09685*, 2021. 3

[12] Simindokht Jahangard, Zhixi Cai, Shiki Wen, and Hamid Rezatofighi. Jrdb-social: A multifaceted robotic dataset for understanding of context and dynamics of human interactions within social groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 7031–7040, 2023. 1, 2, 3, 4

[13] Justin Johnson, Ranjay Krishna, Michael Stark, Li Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2

[14] Kento Kawaharazuka, Tatsuya Matsushima, Andrew Gambardella, Jiaxian Guo, Chris Paxton, and Andy Zeng. Real-world robot applications of foundation models: A review. In *arXiv preprint arXiv:2402.05741*, 2025. 1

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li Jia Li, David A. Shamma, Michael Bernstein, and Li Fei Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *International Journal of Computer Vision (IJCV)*, 2017. 1, 2

[16] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *arXiv preprint arXiv:2310.03744*, 2024. 1, 2, 4

[17] Hanheide Marc, Hebesberger Denise, and Krajník Tomáš. The when, where, and how: An adaptive robotic infoterminal for care home residents. In *the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017. 1

[18] Roberto Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. In *IEEE transactions on pattern analysis and machine intelligence*, 2021. 1, 2

[19] Maëlic Neau, Paulo E. Santos, Anne Gwenn Bosser, Cédric Buche, and Akihiro Sugimoto. React: Real-time efficiency and accuracy compromise for tradeoffs in scene graph generation. In *The 36th British Machine Vision Conference (BMVC)*, 2025. 1, 2

[20] Trong-Thua Nguyen, Pha Nguyen, and Khoa Luu. Hig: Hierarchical interlacement graph approach to scene graph generation in video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[21] Maithili Patel and Sonia Chernova. Proactive robot assistance via spatio-temporal object modeling. In *arXiv preprint arXiv:2211.15501*, 2022. 1

[22] Maithili Patel, Aswin Prakash, and Sonia Chernova. Predicting routine object usage for proactive robot assistance. In *arXiv preprint arXiv:2412.05271*, 2024. 1, 4

[23] Judea Pearl. Causality: Models, reasoning, and inference. In *Cambridge University Press*, 2009. 1, 2

[24] Yao Rong, Tobias Leemann, Thai trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable ai: A survey of user studies for model explanations. In *arXiv preprint arXiv:2210.11584*, 2024. 1

[25] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. In *Proceedings of the IEEE - Advances in Machine Learning and Deep Neural Networks*, 2021. 1, 2

[26] Ivaxi Sheth, Bahare Fatemi, and Mario Fritz. Causalgraph2llm: Evaluating llms for causal queries. In *arXiv preprint arXiv:2410.15939*, 2025. 1, 2

[27] Maayan Shvo, Ruthrash Hari, Ziggy O'Reilly, Sophia Abolore, Sze-Yuh Nina Wang, and Sheila A. McIlraith. Proactive robotic assistance via theory of mind. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022. 1

[28] Weizheng Wang, Le Mao, Ruiqi Wang, and Byung Cheol Min. Multi-robot cooperative socially-aware navigation using multi-agent reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 1, 2

[29] Wei Wang, Weihe Ren, Mengjia Li, and Pingping Chu. A survey on the use of intelligent physical service robots for elderly or disabled bedridden people at home. In *International Journal of Crowd Science*, 2024. 1

[30] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS)*, 2021. 1

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS)*, 2022. 1

[32] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. In *arXiv preprint arXiv:2505.09388*, 2025. 2

[33] Yoshiki Yano, Kazuki Shibata, Maarten Kokshoorn, and Takamitsu Matsubara. Icco: Learning an instruction-conditioned coordinator for language-guided task-aligned multi-robot control. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025. 1, 2

[34] Chen Zhe, Alonso Mora Javier, Bai Xiaoshan, Harabor Daniel Damir, and Stuckey Peter James. Integrated task assignment and path planning for capacitated multi-agent pickup and delivery. In *IEEE Robotics and Automation Letters*, 2021. 1

[35] Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Mingtao Feng, Xia Zhao, Qiguang Miao, Syed Afaq Ali Shah, and Mohammed Bennamoun. Scene graph generation: A comprehensive survey. In *arXiv preprint arXiv:2201.00443*, 2024. 1, 2

[36] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. In *arXiv preprint arXiv:2504.10479*, 2025. 2