GraphCoT-VLA: A 3D Spatial-Aware Reasoning Vision-Language-Action Model for Robotic Manipulation with Ambiguous Instructions

Helong Huang[1*], Min Cen[2*†], Kai Tan[1], Xingyue Quan[1], Guowei Huang[1], Hong Zhang[3†]
[1]Noah's Ark Lab, Huawei
[2]School of Artificial Intelligence and Data Science, University of Science and Technology of China
[3]School of Management, University of Science and Technology of China
{huanghelong1, tankai, quanxingyue, huangguowei}@huawei.com, cenmin0127@mail.ustc.edu.cn, zhangh@ustc.edu.cn

Abstract

Vision-language-action models have emerged as a crucial paradigm in robotic manipulation. However, existing VLA models exhibit notable limitations in handling ambiguous language instructions and unknown environmental states. Furthermore, their perception is largely constrained to static two-dimensional observations, lacking the capability to model three-dimensional interactions between the robot and its environment. To address these challenges, this paper proposes GraphCoT-VLA, an efficient end-to-end model. To enhance the model's ability to interpret ambiguous instructions and improve task planning, we design a structured Chain-of-Thought reasoning module that integrates high-level task understanding and planning, failed task feedback, and low-level imaginative reasoning about future object positions and robot actions. Additionally, we construct a real-time updatable 3D Pose-Object graph, which captures the spatial configuration of robot joints and the topological relationships between objects in 3D space, enabling the model to better understand and manipulate their interactions. We further integrates a dropout hybrid reasoning strategy to achieve efficient control outputs. Experimental results across multiple real-world robotic tasks demonstrate that GraphCoT-VLA significantly outperforms existing methods in terms of task success rate and response speed, exhibiting strong generalization and robustness in open environments and under uncertain instructions.

Introduction

To achieve natural and efficient human-robot interaction in the real world, end-to-end vision-language-action (VLA) models are increasingly becoming a central paradigm in robotics (Black et al. 2024; Liu et al. 2024; Wen et al. 2024, 2025a; Zhang and Yan 2023; Team et al. 2024). VLA models interpret and execute human instructions expressed in natural language within a closed-loop system that integrates perception, understanding, and action (Din et al. 2025). Many existing approaches build upon pre-trained Vision-Language Models (VLMs) to form Vision-Language Action (VLA) frameworks, leveraging VLMs' strong visual grounding and language understanding capabilities to jointly encode multimodal inputs and directly predict control actions for robotic manipulation (Black et al. 2024; Kim et al. 2024).

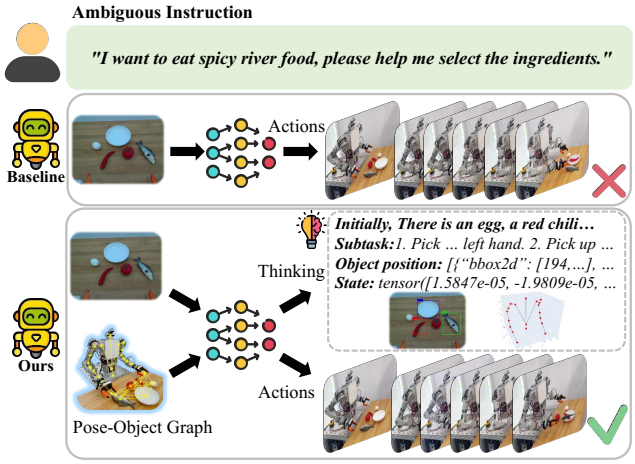


Figure 1: Comparison between our method and baseline.

Most existing VLA models rely on clear and structured language instructions to perform specific tasks (Black et al. 2024; Kim et al. 2024; Liu et al. 2024). However, in real-world scenarios, users often lack full awareness of the environment and may issue vague commands like "I want to eat spicy river food." without knowing whether the necessary ingredients are available. When encountering such ambiguous or infeasible instructions, these models often fail or produce hallucinated actions, revealing a fundamental flaw in existing approaches: their inability to contextualize multimodal perceptual inputs and adapt action planning accordingly, which prevents reasoning and decision-making from being grounded in the real physical environment. Some works have introduced Chain-of-Thought (CoT) (Wei et al. 2022) with VLA models to improve reasoning ability. ECoT (Zawalski et al. 2024) decomposes tasks into multi-step subtasks for planning, but struggles with ambiguous commands and unmet user expectations due to limited situational awareness. CoT-VLA (Zhao et al. 2025) imagines observations of subtasks, yet lacks the ability to plan or give feedback based on a full understanding of the scene.

*These authors contributed equally.
†Corresponding authors.

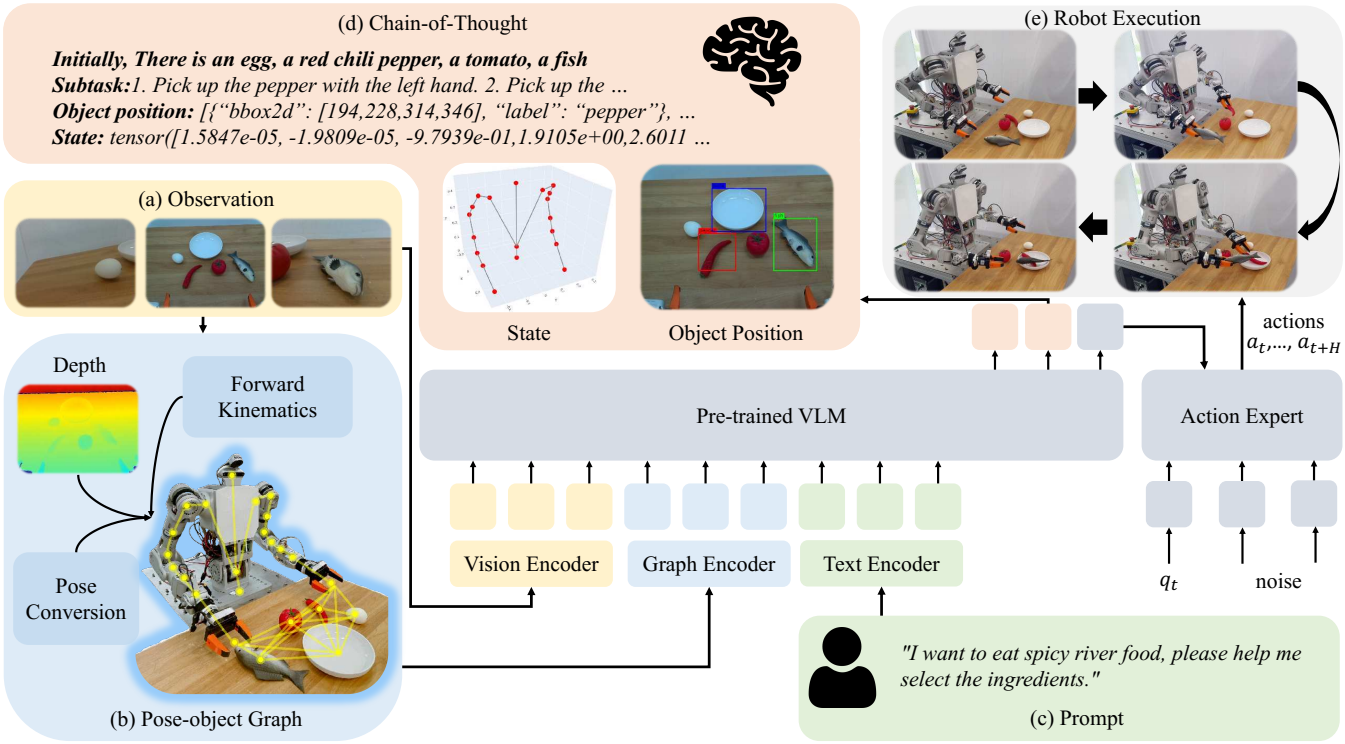


Figure 2: Overview of GraphCoT-VLA. GraphCoT-VLA takes multi-view visual observations, an ambiguous language instruction, and a 3D Pose-Object Graph as input. The graph is constructed using robot kinematics and depth data to represent 3D spatial relationships in the scene. A structured Chain-of-Thought reasoning process enables scene understanding, subtask decomposition, and imagination of object states and positions, which are autoregressively generated to guide the action expert.

In addition, existing VLA models remain limited in observation modalities, typically relying only on RGB images from multiple viewpoints (Black et al. 2024; Zhao et al. 2025). This restricted perception scope hinders model's ability to develop a deep understanding of the environment. To improve this, recent studies have explored the integration of additional modalities, such as positional encodings (Qu et al. 2025) and depth information (Wang 2025) to enhance spatial perception . Although these methods expand the input space, they primarily rely on static 2D views of individual objects and lack the capacity to model complex 3D interactions between the robot and its environment.

To address the above challenges, we propose GraphCoT-VLA, an efficient end-to-end model for robotic manipulation. As shown in Fig. 1. the core of our model is a novel CoT architecture that performs dynamic observation analysis, interprets ambiguous instructions, generates failure feedback, and predicts future object states and robot actions, enabling robust reasoning and temporal forecasting in uncertain environments. We further introduce a real-time Pose-Object Graph to explicitly model 3D spatial relations between the robot and objects, enhancing perception and understanding of 3D space. To ensure real-time performance, GraphCoT-VLA integrates a dropout-based joint training strategy (Chen et al. 2025) and a hybrid CoT reasoning mechanism, balancing fast inference with deep reasoning. Our key contributions are as follows:

- A novel end-to-end model, GraphCoT-VLA, is proposed for robotic manipulation under ambiguous instructions and open-world conditions.
- A structured CoT module is proposed to support scene understanding, feedback, and future imagination.
- A real-time Pose-Object Graph is designed to model 3D interactions between the robot and surrounding objects.
- A hybrid CoT reasoning strategy using dropout-based training enable both fast inference and iterative refinement for real-time control.
- GraphCoT-VLA is validated on real robots, demonstrating superior performance in success rate, action fluency, temporal modeling, and task generalization.

Related Work

Vision-Language-Action Models. VLA models are increasingly emerging as a unified framework that enables robots to perform complex manipulation tasks (Black et al. 2024; Kim et al. 2024; Wen et al. 2025a; Zhang and Yan 2023; Team et al. 2024). Some approaches adopt Transformer-based (Vaswani et al. 2017) architectures to generate control commands (Brohan et al. 2022; Wu et al. 2023; Cheang et al. 2024), while others leverage the enhanced capabilities of VLMs (Zitkovich et al. 2023; Wen et al. 2025b). For instance, π_0 (Black et al. 2024) fine-tunes Gemma to generate continuous actions via flow match-

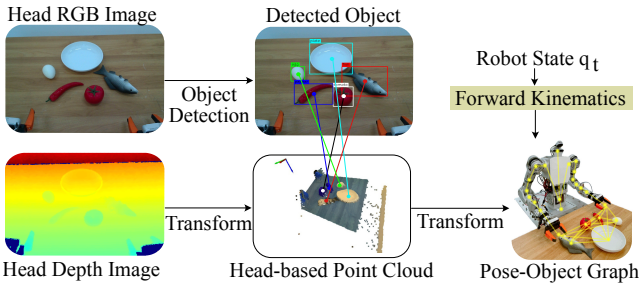


Figure 3: Pose-Object Graph generation.

ing (Lipman et al. 2022). Diffusion-based models have also emerged as promising solutions: RDT-1B (Liu et al. 2024) employs a diffusion Transformer to generate dual-arm control vectors. Other methods aim to improve cross-signal generalization. For example, DexVLA (Wen et al. 2025a) and Crossformer (Zhang and Yan 2023) are designed to handle diverse robotic platforms. In contrast to prior methods that are limited to executing explicit commands, our goal is to enhance task comprehension and feedback responsiveness in open-world scenarios with ambiguous instructions.

Enhance the Reasoning and Comprehension Abilities of VLA. Although some VLA models leverage VLMs, they still lack explicit temporal planning and reasoning capabilities. Recent approaches enhance reasoning by introducing CoT mechanisms. For example, CoT-VLA (Zhao et al. 2025) employs causal attention to generate subgoal images, ECoT (Zawalski et al. 2024) incorporates intermediate reasoning steps to understand task structure. For perception, some works integrate richer information to improve understanding. SpatialVLA (Qu et al. 2025) incorporates positional encodings to capture key spatial points, while RoboFlamingo-Plus (Wang 2025) fuses RGB and depth images to enhance perception. However, existing methods still lack dynamic planning and feedback mechanisms when faced with ambiguous instructions. Most are limited to static 2D observations of single objects, making it difficult to model 3D interactions between the robot and its environment. Therefore, this study proposes a novel structured CoT and 3D spatial graph framework to improve the model's understanding and reasoning of ambiguous scenes.

Methods

This section presents our overall approach, we first introduce the Pose-Object Graph, then describe the design of CoT, followed by the overall architecture and co-training strategy.

Pose-Object Graph

Real-time Construction. To model 3D spatial relationships between the robot and scene objects, we construct a real-time Pose-Object Graph G_i at each time step t_i, running at 50 Hz, using synchronized multimodal data $\mathcal{D} = \{(t_i, I_i^{\text{rgb}}, I_i^{\text{depth}}, \mathbf{q}_i)\}_{i=1}^{N}$, where t_i is the timestamp, I_i^{rgb} and I_i^{depth} are the RGB and depth images, and $\mathbf{q}_i$ is the J-DOF joint configuration. We apply YOLO-World (Cheng et al. 2024) to detect n object bounding boxes from I_i^{rgb}, extract their 2D centers and project them to 3D points in the head camera frame using depth values and intrinsics K, then transform them into the robot base frame using extrinsics T. The robot's end-effector positions are computed via forward kinematics from $\mathbf{q}_i$. All object and end-effector nodes are fully connected to form G_i, which is provided to the VLA model for enhanced 3D scene understanding. Fig. 3 illustrates the process and details are given in Algorithm 1.

Algorithm 1: Real-time Pose-Object Graph Construction.

Input: Demonstration $\mathcal{D} = \{(t_i, I_i^{\text{rgb}}, I_i^{\text{depth}}, \mathbf{q}_i)\}_{i=1}^{N}$, intrinsics matrix K, extrinsic matrix T
Output: The corresponding generated graphs $\{G_i\}_{i=1}^{N}$

```
for i = 1 to N do                          ▷ Same logic used during inference
    B_i ← DetectObj(I_i^rgb)
    {(x_j, y_j)}_{j=1}^{n} ← F_center(B_i)
    Initialize V_obj ← ∅
    for j = 1 to n do
        D_j ← F_depth(I_i^depth, (x_j, y_j))
        P_j^head ← D_j · K^{-1}(x_j, y_j, 1)^⊤
        P_j^base ← T · P_j^head            ▷ Convert to robot base frame
        V_obj ← V_obj ∪ {P_j^base}
    end for
    V_ee ← FK(q_i)
    E ← {(v_obj, v_ee) | v_obj ∈ V_obj, v_ee ∈ V_ee}
    G_i ← (V_obj ∪ V_ee, E)
end for
return {G_i}_{i=1}^{N}
```

Graph Encoder. We encode the Pose-Object Graph G_t using a two-layer graph neural network (GNN) (Scarselli et al. 2008), where each layer sequentially applies layer normalization, graph convolution, and ReLU activation. Let $\mathbf{H}^{(0)} \in \mathbb{R}^{(m+n)\times d}$ denote the initial embeddings of Pose-Object Graph, represented as 3D positions in a d-dimensional space. We update feature by:

$$\mathbf{H}^{(l+1)} = \sigma\big(\mathcal{G}(\mathcal{N}(\mathbf{H}^{(l)}), \mathbf{A})\big), \tag{1}$$

where $\mathcal{N}$, $\mathcal{G}$, and σ denote layer normalization, graph convolution, and ReLU activation, and $\mathbf{A}$ is the adjacency matrix of G_t. The final node features are fed into the VLM for downstream reasoning and action prediction.

CoT Reasoning via Understanding, Feedback and Imagination

To improve reasoning under ambiguous instructions and novel visuals, and enable suggestion-based feedback in the VLA framework, we propose a structural CoT mechanism (Fig. 2) that integrates scene understanding, instruction interpretation, feedback generation, and future imagination.

Specifically, we first use Qwen2.5-VL-7B (Bai et al. 2025) to perform initial scene understanding from the head image by identifying objects in the environment, prompt example:

> The desk contains the necessary ingredients and a plate. Please observe the image and describe the items on the desk with very simple sentences.

Next, guided by task requirements, we perform feasibility analysis and decompose ambiguous instructions. If the scene does not satisfy the instruction, we query Qwen2.5-VL-7B with a targeted prompt to generate feedback and suggestions, prompt example:

> I want to eat spicy river seafood. Please recommend the ingredients that are still missing from the table and needed with very simple sentences.

Beyond understanding the current scene, CoT module incorporates future prediction through a generic frame sampling strategy. At each time step t, We set a fixed time interval $\Delta t = 30$ frames and perform object detection using Qwen2.5-VL-7B on the future frame at $t' = \lfloor \frac{t}{\Delta t} \rfloor \times \Delta t + \Delta t$. Additionally, the robot's state at frame $t'' = t + \Delta t$ is converted into a textual description and appended to the CoT.

Overall Framework of GraphCoT-VLA

We adopt π_0 as the baseline. At each time step t, the robot receives multi-view images from the head, left and right wrist cameras: $I_{\text{head},t}$, $I_{\text{left},t}$, and $I_{\text{right},t}$. Additionally, proprioception q_t, an ambiguous instruction $\mathcal{L}$, and a Pose-Object Graph $\mathcal{G}_t$ are fed into model. Our model learns the conditional distribution of future actions as:

$$P(a_{t+1:t+\Delta t} \mid I_{\text{head},t}, I_{\text{left},t}, I_{\text{right},t}, \mathcal{G}_t, q_t, \mathcal{L}). \quad (2)$$

To model this distribution, we first encode the visual inputs $\{I_{\text{head},t}, I_{\text{left},t}, I_{\text{right},t}\}$ by vision transformer (Dosovitskiy et al. 2020). The Pose-Object Graph $\mathcal{G}_t$, representing the robot's configuration, is encoded using a graph encoder that treats each node as a graph token. Then the embeddings are concatenated with the tokenized language instruction $\mathcal{L}$ to construct the input sequence for the model.

The input sequence is fed into a VLM built upon PaliGemma (Beyer et al. 2024). The output tokens from the VLM are divided into two parts: the first is used to autoregressively generate a CoT explanation, while the second is passed to an action expert module based on flow matching. This module also takes as input the robot state q_t and action noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The action expert then generates the predicted sequence of future actions $\hat{A}_t = \hat{a}_{t+1:t+\Delta t}$.

Co-Training with Dropout

CoT Optimization. To supervise the generation of CoT reasoning, we apply a standard cross-entropy loss (Mao, Mohri, and Zhong 2023):

$$\mathcal{L}_{\text{CoT}} = -\sum_{i=1}^{T_{\text{CoT}}} \log P(y_i \mid y_{<i}, o_t^{\text{CoT}}), \quad (3)$$

where y_i denotes the i-th token in the ground-truth CoT sequence. T_{CoT} denotes the number of tokens in the output CoT sequence, and $o_t^{CoT} = \{I_{\text{head},t}, I_{\text{left},t}, I_{\text{right},t}, \mathcal{G}_t, q_t, \mathcal{L}\}$.

Action Optimization. A conditional flow matching loss (Lipman et al. 2022) is adopted to learn a time-dependent denoising vector field. Given a future action sequence $A_t = [a_{t+1}, \ldots, a_{t+\Delta t}]$, we sample a time coefficient $\tau \sim \text{Beta}(\alpha, \beta)$ and noise $\epsilon \sim \mathcal{N}(0, I)$, and compute a perturbed action A_t^τ as a weighted combination of the original action A_t and noise ϵ, given by $A_t^\tau = \tau A_t + (1-\tau)\epsilon$. The network is trained to predict the denoising vector $\mathbf{u}(A_t^\tau \mid A_t) = \epsilon - A_t$, and loss defined as:

$$\mathcal{L}_{\text{action}} = \mathbb{E}_{A_t,\tau,\epsilon}\left[\left\|\mathbf{v}_\theta(A_t^\tau, o_t^{\text{action}}) - (\epsilon - A_t)\right\|^2\right], \quad (4)$$

where $\mathbf{v}_\theta$ is the learned vector field conditioned on the full observation. o_t^{action} refers to the portion of the VLM output tokens following the CoT reasoning.

To support real-time inference, we adopt a joint training strategy with CoT supervision dropout. Each sample is randomly dropped with probability p, enabling the model to learn from both reasoning-guided and direct action prediction modes. The training loss is defined as:

$$\mathcal{L}_{\text{total}} = (1-d)\cdot(\lambda_{\text{CoT}}\cdot\mathcal{L}_{\text{CoT}} + \lambda_{\text{action}}\cdot\mathcal{L}_{\text{action}}) + d\cdot\mathcal{L}_{\text{action}}, \quad (5)$$

where $d \sim \text{Bernoulli}(p)$ indicates whether CoT supervision is dropped for the sample ($d = 1$ means dropout), and $\lambda_{\text{CoT}}, \lambda_{\text{action}}$ are balancing weights, set to 1 in our experiments. During inference, the model adopts a hybrid reasoning strategy: CoT is generated only for the first frame to provide feedback and guidance, while subsequent frames skip reasoning and predict actions for robot control.

Experiments

Experimental Setup

Datasets. We collect data on a bimanual robot (Huang et al. 2025) with 7 degrees of freedom per arm(Fig. 2(b)), including RGB-D images from the head, neck, and hand-mounted cameras. As data streams run at different rates, with cameras at 30 Hz and proportional-derivative (PD) control at 150 Hz, we align all modalities using the head camera's timestamp, actions are defined in joint angle space.

Task Descriptions. The tasks, Food Preparation and Outfit Selection, are designed to evaluate the robot's ability to

Food Preparation				
Subtask	**Egg**	**Tomato**	**Fish**	**Pepper**
Subtask 1	✓	✓	✓	✓
Subtask 2	✓	✓		✓
Subtask 3	✓	✓		
Outfit Selection				
Subtask	**Sweater**	**T-shirt**	**Shorts**	
Subtask 1	✓	✓	✓	
Subtask 2	✓	✓		
Subtask 3	✓			

Table 1: Available objects per subtask.

Method	Food Preparation				Outfit Selection				No Task Confusion
	fish & pepper	pepper	nothing	Avg.	T-shirt & shorts	T-shirt	nothing	Avg.	
ACT	45	65	65	58.33	50	45	0	31.67	✗
Diffusion Policy	35	65	60	53.33	45	30	5	26.67	✗
Octo fine-tuned	60	60	**80**	66.67	55	50	45	50.00	✗
π_0 fine-tuned	55	65	50	56.67	50	55	50	51.67	✗
Ours	**75**	**80**	75	**76.67**	**65**	**70**	**75**	**70.00**	✓

Table 2: Comparison of our method and baselines. Success rates (%) for two tasks, the last column indicates task confusion.

Method	Food Preparation				Outfit Selection			
	fish & pepper	pepper	nothing	Avg.	T-shirt & shorts	T-shirt	nothing	Avg.
w/o CoT&Graph (π_0)	55	65	50	56.67	50	55	50	51.67
w/o PoseGraph	70	80	75	75.00	55	60	55	56.67
w/o CoT	70	80	55	68.33	55	60	60	58.33
Ours	**75**	**80**	**75**	**76.67**	**65**	**70**	**75**	**70.00**

Table 3: Ablation study. Success rates (%) of our model and variants on subtasks from two tasks.

perform under ambiguous instructions. The evaluation focuses on: (1) understanding ambiguous commands; (2) bimanual collaboration to test policy robustness and coordination; and (3) scenario diversity across subtasks to assess whether the model can distinguish tasks relying solely on visual and contextual information. Table 1 summarizes the varying object availability across subtasks for both Food Preparation and Outfit Selection, representing different situations the robot encounters. To encourage generalization, we vary object placement during data collection, items are randomly translated 10 cm and rotated up to 30°, clothing arrangements vary per subtask, with predefined permutations or uniform sampling across 10 hanger segments, we collect 100 demonstrations per subtask, totaling 600 for training.

Baselines. We compare our method with four state-of-the-art baselines: **ACT**(Zhao et al. 2023), a transformer-based trained from scratch to map images to actions; **Diffusion Policy**(Chi et al. 2023), an imitation learning method generating actions via diffusion, trained from scratch per task; **Octo**(Team et al. 2024), a pretrained transformer on 800k Open X-Embodiment (O'Neill et al. 2024) episodes, fine-tuned with released weights; $\boldsymbol{\pi_0}$(Black et al. 2024), a general-purpose VLA model using both image and language inputs, also fine-tuned from released weights.

Training and Evaluation Protocol. We follow ACT's (Zhao et al. 2023) task failure rate evaluation protocol, testing each task 20 times. A trial is successful only if the entire task is completed correctly. The details of the experimental setup and the criteria for identifying failure cases are provided in the appendix.

Comparative Results with SOTA Models

Table 2 compares our method with state-of-the-art approaches. In Food Preparation, Octo had the highest baseline accuracy, which we improved by 10%. In Outfit Selection, π_0 led among baselines, and we outperformed it by 18.33%. These results demonstrate that our approach offers higher accuracy and robustness, particularly in scenarios with task ambiguity. During real-world robot inference, as shown in Fig. 4, our method exhibits smoother body control. Notably, all baseline methods showed task confusion when handling ambiguous instructions. For example, as illustrated in Fig. 4(6), when the task required the robot to grasp a T-shirt, other methods incorrectly targeted a sweater. In contrast, our method correctly inferred the task intent and successfully executed the grasp, demonstrating superior task understanding and execution consistency.

Ablation Studies

3D Spatial Awareness through Pose-Object Graph. As shown in Table 3, the introduction of the Pose-Object Graph increases the success rate by up to 18.33%. During the experiments, we observed that with graph reasoning (illustrated in Fig. 4(a)), the robot's movements when grasping the target object become more decisive, coherent and natural. Furthermore, in the second task, which requires high end-effector precision to successfully grasp the hanger, the Pose-Object Graph leads to more accurate execution and greater operational stability. This demonstrates that the Pose-Object Graph improves the model's accuracy and the fluency of action generation.

Enhancing Comprehension and Reasoning with CoT. As shown in Fig. 5(b), the CoT module enables the model to form a global understanding of the task based on varying environmental observations under ambiguous instructions. For instance, in the Food Selection task, when only chili peppers satisfy the given conditions, the model successfully provides reasonable suggestions. Additionally, it demonstrates preliminary sub-task planning abilities and the capacity to

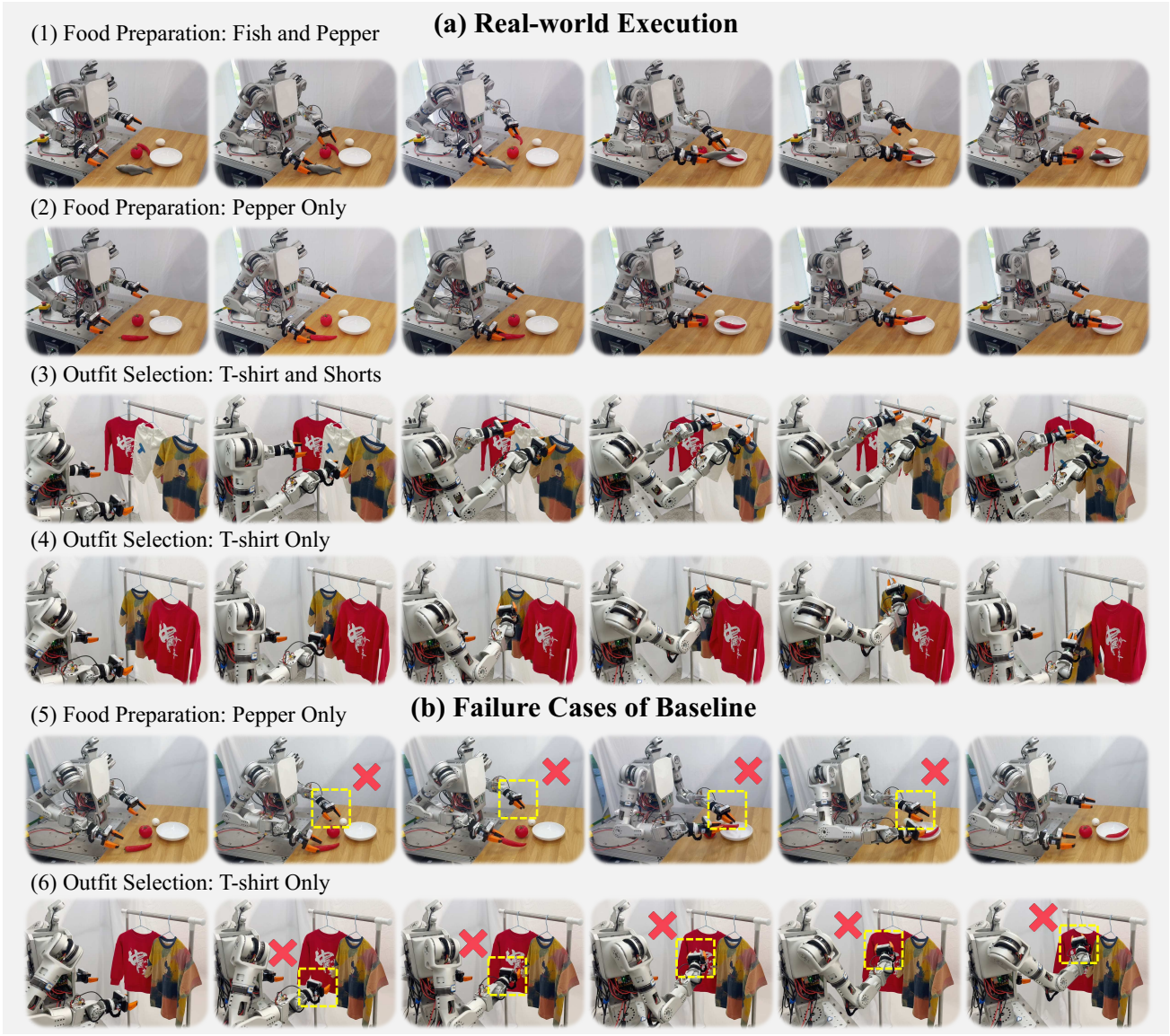


Figure 4: Real-world Executions. Each row shows a task in order, with baseline failures highlighted in yellow.

anticipate future steps. In addition, as shown in Table 3, the introduction of the CoT module leads to a significant improvement in the success rates. We also observed that the CoT module enhances the model's ability to interpret and respond to chaotic or ambiguous situations. The baseline models often exhibit issues such as "insufficient or continuous waving", which tends to make decisions based solely on current observations, lacking the capacity to reason about future task steps. With the CoT, the model demonstrates stronger task planning capabilities and a greater ability to predict future actions, resulting in more coherent behavior.

Qualitative Visualization

Pose-Object Graph Visualization. We visualize its structure in Fig. 6(a), where nodes represent robot joints or object centers in the world frame. Joints are connected via the kinematic chain, and the end-effector links to all objects to form a unified graph. This visualization clearly illustrates the spatial configuration between the robot and surrounding objects, demonstrating that the graph effectively encodes scene layout for downstream action generation.

Chain-of-thoughts visualization. Fig. 5 shows CoT's reasoning under varying task conditions. In ambiguous scenes, where both fish and chili peppers are valid(Fig. 5(a), CoT captures global context (purple box) and generates appropriate subtasks (orange box). When only one valid option remains (Fig. 5(b)), it adapts its plan. The comparison between Fig. 5(c) and (d) further demonstrates that CoT adapts its interpretation and reasoning based on observed changes. These examples highlight CoT's ability to interpret scene semantics and adjust decisions based on observation.

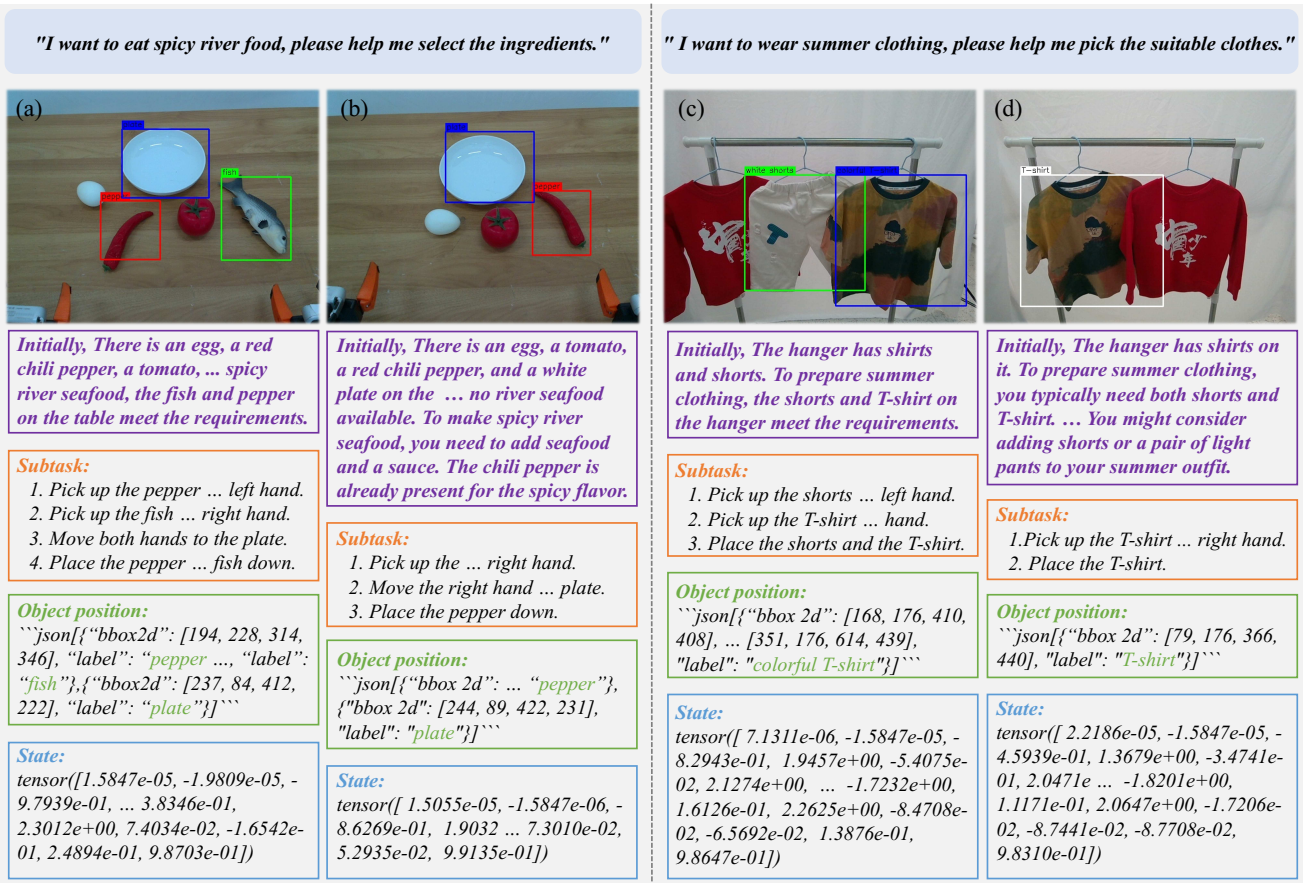
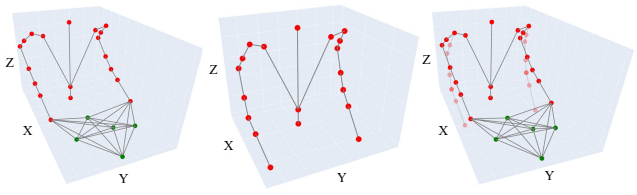


Figure 5: Visualization CoT reasoning from model inference.

In addition to high-level reasoning, CoT accurately predicts low-level future states, including target object positions and the robot's configuration. As shown in Fig. 5, the boxed areas mark predicted object locations that closely match the ground truth. Fig. 6(c) further compares CoT's predicted body state with actual motion, showing strong alignment in end-effector position and overall kinematics. These results confirm CoT's capability in temporal reasoning and fine-grained state prediction.



(a) Pose-Object Graph (b) CoT-Predicted State (c) Combined Graph

Figure 6: Graph visualization. (a) Pose-Object Graph: Red and green nodes represent robot joints and object points, respectively. (b) CoT Prediction: Robot state predction by CoT. (c) Combined Graph: Overlay of (a) and (b), with transparent nodes indicating CoT predictions.

Model	Frequency
π_0	$\approx$10 Hz
Ours (co-training)	$\approx$10 Hz

Table 4: Comparison of inference frequencies.

Efficiency Study

The model's inference speed directly affects the smoothness of real-world robot execution. We measured the inference frequency of π_0 ten times, averaging about 10 Hz (Table 4). With CoT co-training on the same hardware, the first frame took slightly longer ($<$ 30 s), but subsequent inference also ran at 10 Hz, showing no added computational overhead.

Conclusion

In this work, we propose GraphCoT-VLA to address robotic manipulation tasks under ambiguous instructions. Specifically, we design a Pose-Object graph and a hierarchical CoT structure that enhance the model's understanding and reasoning capabilities. Currently, our approach relies on future imagination and lacks memory of past temporal information, future work would explore incorporating historical sequences to enable more coherent long-term reasoning.

Acknowledgements

The work of Helong Huang was partly supported by Noah's Ark Lab in Huawei. The work of Hong Zhang was partly supported by the National Natural Science Foundation of China (Grant No.7209121,12171451) and Anhui Center for Applied Mathematics.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Beyer, L.; Steiner, A.; Pinto, A. S.; Kolesnikov, A.; Wang, X.; Salz, D.; Neumann, M.; Alabdulmohsin, I.; Tschannen, M.; Bugliarello, E.; et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Black, K.; Brown, N.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Groom, L.; Hausman, K.; Ichter, B.; et al. 2024. pi_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Hsu, J.; et al. 2022. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*.

Cheang, C.-L.; Chen, G.; Jing, Y.; Kong, T.; Li, H.; Li, Y.; Liu, Y.; Wu, H.; Xu, J.; Yang, Y.; et al. 2024. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*.

Chen, W.; Belkhale, S.; Mirchandani, S.; Mees, O.; Driess, D.; Pertsch, K.; and Levine, S. 2025. Training Strategies for Efficient Embodied Reasoning. *arXiv preprint arXiv:2505.08243*.

Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; and Shan, Y. 2024. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16901–16911.

Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2023. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 02783649241273668.

Din, M. U.; Akram, W.; Saoud, L. S.; Rosell, J.; and Hussain, I. 2025. Vision Language Action Models in Robotic Manipulation: A Systematic Review. *arXiv preprint arXiv:2507.10672*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Huang, H.; Mower, C. E.; Huang, G.; Das, S.; Dierking, M.; Luo, G.; Tan, K.; Chen, X.; Yang, Y.; Chen, Y.; Zeng, Y.; Li, Y.; Zhang, Z.; Wu, S.; Zhang, Y.; Qiu, W.; Cao, T.; Qin, M.; Pakdamansavoji, S.; Liu, Y.; Zhuang, Y.; Tian, G.; Hao, J.; Wang, J.; Bou-Ammar, H.; and Quan, X. 2025. OpenPyRo-A1: An Open Python-based Low-Cost Bimanual Robot for Embodied AI. *IEEE Robotics and Automation Letters*.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Liu, S.; Wu, L.; Li, B.; Tan, H.; Chen, H.; Wang, Z.; Xu, K.; Su, H.; and Zhu, J. 2024. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*.

Mao, A.; Mohri, M.; and Zhong, Y. 2023. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, 23803–23828. pmlr.

O'Neill, A.; Rehman, A.; Maddukuri, A.; Gupta, A.; Padalkar, A.; Lee, A.; Pooley, A.; Gupta, A.; Mandlekar, A.; Jain, A.; et al. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 6892–6903. IEEE.

Qu, D.; Song, H.; Chen, Q.; Yao, Y.; Ye, X.; Ding, Y.; Wang, Z.; Gu, J.; Zhao, B.; Wang, D.; et al. 2025. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80.

Team, O. M.; Ghosh, D.; Walke, H.; Pertsch, K.; Black, K.; Mees, O.; Dasari, S.; Hejna, J.; Kreiman, T.; Xu, C.; et al. 2024. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, S. 2025. Roboflamingo-plus: Fusion of depth and rgb perception with vision-language models for enhanced robotic manipulation. *arXiv preprint arXiv:2503.19510*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wen, J.; Zhu, M.; Zhu, Y.; Tang, Z.; Li, J.; Zhou, Z.; Li, C.; Liu, X.; Peng, Y.; Shen, C.; et al. 2024. DiffusionVLA: Generalizable and Interpretable Robot Foundation Model via Self-Generated Reasoning. *arXiv preprint arXiv:2412.03293*.

Wen, J.; Zhu, Y.; Li, J.; Tang, Z.; Shen, C.; and Feng, F. 2025a. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*.

Wen, J.; Zhu, Y.; Li, J.; Zhu, M.; Tang, Z.; Wu, K.; Xu, Z.; Liu, N.; Cheng, R.; Shen, C.; et al. 2025b. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*.

Wu, H.; Jing, Y.; Cheang, C.; Chen, G.; Xu, J.; Li, X.; Liu, M.; Li, H.; and Kong, T. 2023. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*.

Zawalski, M.; Chen, W.; Pertsch, K.; Mees, O.; Finn, C.; and Levine, S. 2024. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*.

Zhang, Y.; and Yan, J. 2023. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*.

Zhao, Q.; Lu, Y.; Kim, M. J.; Fu, Z.; Zhang, Z.; Wu, Y.; Li, Z.; Ma, Q.; Han, S.; Finn, C.; et al. 2025. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1702–1713.

Zhao, T. Z.; Kumar, V.; Levine, S.; and Finn, C. 2023. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*.

Zitkovich, B.; Yu, T.; Xu, S.; Xu, P.; Xiao, T.; Xia, F.; Wu, J.; Wohlhart, P.; Welker, S.; Wahid, A.; et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2165–2183. PMLR.