# RoboStream: Weaving Spatio-Temporal Reasoning with Memory in Vision-Language Models for Robotics

Yuzhi Huang[1,2*‡♠] Jie Wu[1,2*‡] Weijue Bu[3*] Ziyi Xiong[2,4‡] Gaoyang Jiang[5] Ye Li[1] Kangye Ji[1,2‡] Shuzhao Xie[1] Yue Huang[6] Chenglei Wu[2] Jingyan Jiang[4†] Zhi Wang[1†]

[1] Shenzhen International Graduate School, Tsinghua University, China
[2] YuanxingGuangnian Robotics, China
[3] China University of Mining and Technology, China
[4] Shenzhen Technology University, China
[5] Huazhong University of Science and Technology, China
[6] Xiamen University, China

Website: https://robostream123.github.io/

**Abstract.** Enabling reliable long-horizon robotic manipulation is a crucial step toward open-world embodied intelligence. However, VLM-based planners treat each step as an isolated observation-to-action mapping, forcing them to reinfer scene geometry from raw pixels at every decision step while remaining unaware of how prior actions have reshaped the environment. Despite strong short-horizon performance, these systems lack the spatio-temporal reasoning required for persistent geometric anchoring and memory of action-triggered state transitions. Without persistent state tracking, perceptual errors accumulate across the execution horizon, temporarily occluded objects are catastrophically forgotten, and compounding failures lead to precondition violations that cascade through subsequent steps. In contrast, humans maintain a persistent mental model that continuously tracks spatial relations and action consequences across interactions rather than reconstructing them at each instant. Inspired by this human capacity for causal spatio-temporal reasoning with persistent memory, we propose **RoboStream**, a training-free framework that achieves geometric anchoring through Spatio-Temporal Fusion Tokens (STF-Tokens), which bind visual evidence to 3D geometric attributes for persistent object grounding, and maintains causal continuity via a Causal Spatio-Temporal Graph (CSTG) that records action-triggered state transitions across steps. This design enables the planner to trace causal chains and preserve object permanence under occlusion without additional training or fine-tuning. RoboStream achieves a 90.5% success rate on long-horizon RLBench tasks and a 44.4% success rate on challenging real-world block-building tasks, where both SoFar and VoxPoser score 11.1%, demonstrating that spatio-temporal reasoning and causal memory are critical missing components for reliable long-horizon manipulation.

**Keywords:** Robot Manipulation · Vision-Language Models · Long-Horizon Planning · Spatio-Temporal Reasoning · Causal Memory

---

[*] Equal contribution. [†] Corresponding author. ♠ Project lead.
[‡] Work done during an internship at Yuanxing Robotics.

## 1 Introduction

Vision-language models (VLMs) have emerged as powerful planners for robotic manipulation, leveraging internet-scale semantic knowledge and visual perception to translate high-level instructions into executable action sequences [14,32,34,81]. VLM-based systems have demonstrated strong performance on short-horizon manipulation tasks, including precise grasping, pose-aware placement, and decomposition of language instructions into subgoal sequences [4,26,47,56].

However, this success does not readily extend to long-horizon tasks, which demand sustained spatial and causal awareness across multi-step action sequences that irreversibly alter the environment [55,66,76]. Current VLM-based planners lack the persistent state tracking required to bridge observations across steps, forcing them to reinfer world state from partial observations at each decision step [19,28,79]. Without accumulated context, perceptual errors and action effects compound across steps, ultimately resulting in cascading failures, as illustrated in Fig. 1. In contrast, humans maintain a coherent mental model [35] that continuously integrates spatial relations and causal consequences as actions unfold, tracking not only where objects are but also how each action has cumulatively reshaped the world.

Replicating this human capacity for spatio-temporal reasoning requires persistent tracking of two interdependent quantities, namely the evolving geometric state of each object across steps and the causal record of how actions modify those states. Tracking alone is insufficient, as geometric states must be anchored in 3D space for reliable spatial reasoning while causal records must be structured to support inference over displaced or occluded objects. Current VLM-based planners fail at both, as illustrated in Fig. 1, exposing two tightly coupled representation deficits. ① *Ungrounded Spatial Perception*: spatial relations in manipulation are latent state components continually rewritten by contact and motion, yet current systems infer them implicitly via image-centric perception without geometric anchoring or cross-step identity binding [8,22,72]. This forces geometry reconstruction from raw pixels at every step, where small inaccuracies accumulate into spatial hallucinations and precondition violations. ② *Untracked Causal History*: although actions irreversibly alter the environment, existing methods retain no record linking actions to induced state transitions. Without causal tracking, planners remain unaware of displaced objects and lose track of occluded ones entirely, leading to precondition violations when subsequent actions depend on altered states [64,77].

To address these two deficits, we propose **RoboStream**, a training-free framework that weaves spatio-temporal reasoning and persistent memory directly into the VLM planning loop through *Spatio-Temporal Grounding* and *Causal Memory Tracking*. For grounding, Spatio-Temporal Fusion Tokens (STF-Tokens) distill visual appearance and 3D geometry (centroid and Gaussian shape) of each object into identity-persistent primitives, shifting VLM reasoning from pixel-level reconstruction to structured object-instance references that suppress cascading spatial errors. For tracking, a Causal Spatio-Temporal Graph (CSTG) records action-triggered state transitions alongside persistent object identities,

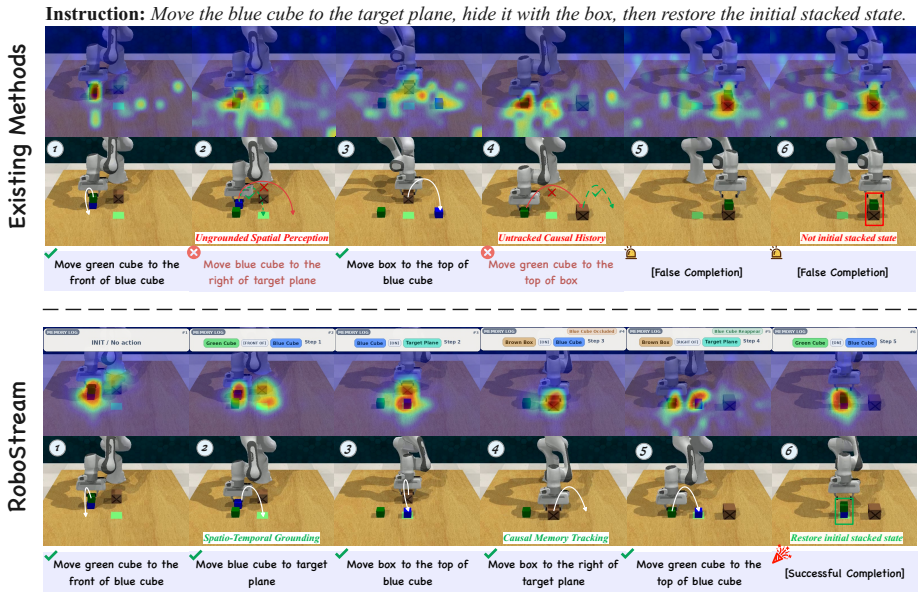


**Fig. 1: Qualitative comparison between existing methods [47] and RoboStream.** This long-horizon task involves object occlusion and state restoration. **Top (Existing Methods):** Diffuse attention heatmaps and absent memory logs reveal two cascading failures: *Ungrounded Spatial Perception* misplaces the blue cube at Step 2, while *Untracked Causal History* leaves the planner unaware of the occluded cube at Step 4, resulting in an incorrect final state. **Bottom (RoboStream):** Concentrated attention heatmaps and persistent memory logs show that *Spatio-Temporal Grounding* enables accurate object placement at Step 2, while *Causal Memory Tracking* maintains awareness of the occluded cube at Step 4, enabling successful task completion.

preserving object permanence under occlusion without direct visual evidence. We evaluate RoboStream on five benchmarks spanning simulated and real-world settings: SIMPLER [37], long-horizon RLBench tasks [31], 6-DoF SpatialBench [47], Open6DOR V2 [12], and long-horizon real-world Franka experiments including block building, disassembly, and hide-and-restore under occlusion. Across all settings, RoboStream outperforms existing VLM-based planners, demonstrating that spatio-temporal reasoning and persistent memory are essential for robust long-horizon manipulation. Our contributions are as follows:

1) We introduce **RoboStream**, a training-free framework weaving spatio-temporal reasoning with persistent memory to overcome spatial hallucination and causal blindness in VLM-based planners. By anchoring geometric evidence across steps and tracking causal state transitions, RoboStream builds persistent world representations enabling robust long-horizon reasoning without retraining.

2) We design two synergistic mechanisms for Spatio-Temporal Grounding and Causal Memory Tracking. *Spatio-Temporal Fusion Tokens (STF-Tokens)* bind visual evidence to 3D geometry, converting pixel-level inference into deterministic

spatial references. A *Causal Spatio-Temporal Graph (CSTG)* maintains object identities and records action-triggered state transitions, enabling causal inference of occluded states and preventing error accumulation over long horizons.

3) RoboStream achieves state-of-the-art results across five benchmarks spanning long-horizon manipulation (RLBench, real-world Franka), zero-shot generalization (SIMPLER), and spatial reasoning (6-DoF SpatialBench, Open6DOR V2), validating that weaving spatio-temporal reasoning with persistent memory is essential for robust manipulation across diverse settings.

## 2 Related Work

### 2.1 Spatial Understanding with VLMs

Spatial understanding underpins robotic manipulation by recovering geometric relationships [29, 65]. Existing approaches address VLM limitations through two main strategies. Tool-augmented methods [12, 48, 59, 80] delegate VLMs to orchestrate external visual foundation models or geometric engines. SpaceTools [9] combines dual-interaction reinforcement learning with geometric reasoning, while TIGER [20] synthesizes code-based geometric routines. These externalize computation but add overhead and require module coupling. 3D-data-driven fine-tuning [10, 11, 51, 58, 62] strengthens spatial representations through pseudo-3D or real-world 3D data. RoboRefer [78] integrates depth encoders, SpatialBot [5] applies progressive spatial curricula, and SpatialVLM [8] leverages large-scale 3D spatial QA datasets. Fine-tuning incurs substantial data and training costs. RoboStream contrasts by anchoring visual evidence directly to explicit 3D geometry via STF-Tokens, requiring no fine-tuning and maintaining persistent spatial representations across sequential steps without external modules or retraining.

### 2.2 VLMs for Robotic Manipulation

Robotic manipulation bridges semantic reasoning with execution along three complementary axes. Logic-grounded task planning [24, 27, 57, 74] decomposes instructions into structured sequences. SayCan [4] scores action feasibility via affordance models, while Code as Policies [43] synthesizes executable programs for long-horizon scheduling. These methods prioritize logical structure but treat execution steps independently. Geometry-aware constraint generation [17, 23, 25, 52, 71, 72] grounds planning in spatial constraints. VoxPoser [26] synthesizes 3D value maps for zero-shot planning, CoPa [22] defines constraints over geometric keypoints, and SoFar [47] maps language to 6-DoF poses. However, these approaches reconstruct geometric constraints at each step without persistent tracking. Closed-loop reasoning [14, 15, 44, 79] enhances adaptability through perceptual feedback. Inner Monologue [28] implements self-correcting loops, and VILA [21] reasons over image sequences for physical intuition. Yet these methods remain reactive, recovering from failures without reasoning about action-induced state changes or preconditions. RoboStream unifies geometry and causality through a CSTG that records action-state transitions, enabling VLMs to maintain causal consistency while reasoning about occluded states and external disturbances.

### 2.3 Long-Horizon Task Planning

Long-horizon manipulation poses a distinct challenge, requiring systems to translate high-level instructions into coherent action sequences while maintaining consistency across extended temporal spans. Dynamic visual conditions and partial observability amplify the challenge of maintaining causal consistency across sequential decisions [19]. Hierarchical decomposition [2, 42, 54] organizes multi-step execution by encoding intermediate representations. HiRobot [55] segments instructions into atomic commands, DexVLA [64] couples VLM planning with diffusion-based policies, and RoboMatrix [45] implements three-tier hierarchical control. LoHoVLA [70] aligns semantic intent via natural-language sub-goals. These strategies organize execution without tracking how past actions reshape the environment. Chain-of-thought methods [46, 63, 73, 75] strengthen planning via explicit future-state reasoning. CoT-VLA [76] reasons over latent visual subgoals, ManualVLA [18] generates executable visual guidance, and GR-3 [7] learns flow-matching trajectories. While improving short-horizon decisions, these methods lack persistent memory to track action consequences. RoboStream addresses this gap through a CSTG recording actions, environmental consequences, and external disturbances, enabling VLMs to trace state evolution and reason causally about how past actions constrain future possibilities under partial observability.

## 3 Method

### 3.1 Overview

VLMs exhibit strong generalization in high-level manipulation planning [4, 34, 81], yet existing planners treat each step independently, lacking the persistent spatio-temporal memory required for long-horizon tasks. Without structured world-state tracking, perception errors compound over time and recovery from unexpected disturbances becomes substantially more difficult. We propose RoboStream, which constructs geometry-grounded STF-Tokens by binding visual evidence to 3D geometric attributes from RGB-D observations (Sec. 3.3), organizes them into a Causal Spatio-Temporal Graph (CSTG) that maintains persistent object identities and logs action-triggered state transitions in a sliding-window memory (Sec. 3.4), and feeds this structured memory into the VLM for chain-of-thought verification to generate 6-DoF action directives (Sec. 3.5), as illustrated in Fig. 2.

Formally, given RGB-D observation $(I_t, D_t)$, goal specification $I_{\text{goal}}$, and memory $\mathcal{M}_{<t}$, RoboStream constructs STF-Tokens encoding per-object 3D geometric properties and visual appearance:

$$\mathcal{T}_t^{\text{STF}} = \Phi_{\text{enc}}(I_t, D_t \mid \mathcal{M}_{<t}), \tag{1}$$

which are incorporated as nodes into the CSTG $\mathcal{G}_t^{\text{CST}}$. By recursively merging incoming STF-Tokens with historical graph state, the CSTG maintains persistent object identities and inter-object spatial relations across time without requiring model retraining:

$$\mathcal{G}_t^{\text{CST}} = \Psi_{\text{graph}}(\mathcal{T}_t^{\text{STF}},\ \mathcal{G}_{t-1}^{\text{CST}}). \tag{2}$$

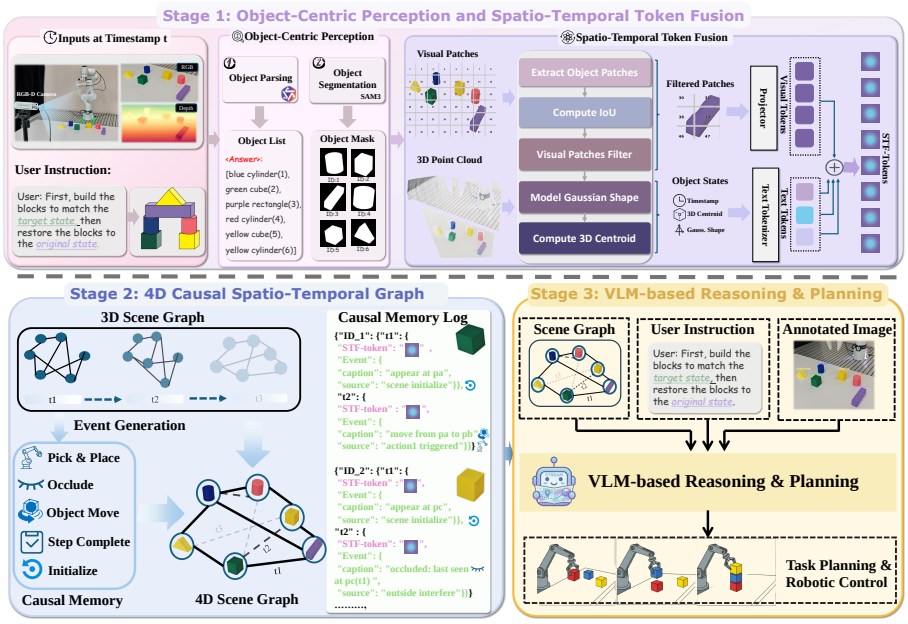


**Fig. 2: Overview of the RoboStream Framework.** The system processes multi-modal RGB-D inputs through three hierarchical stages: (1) Object-Centric Perception and Spatio-Temporal Token Fusion, where raw visual data is distilled into STF-Tokens by grounding visual evidence within 3D geometric primitives (centroids and Gaussian shapes); (2) 4D Causal Spatio-Temporal Graph Construction, which maintains persistent memory encoding object identities and state transitions across time steps; and (3) VLM-based Reasoning and Planning, where the VLM leverages both the CSTG and visual inputs to perform high-level task planning and precise robotic manipulation.

Action generation is then decoupled to balance VLM semantic generalization with robot execution precision. Given a context-augmented prompt $\mathcal{P}_t$ assembled from $\mathcal{G}_t^{\text{CST}}$ and the current observation, the VLM produces a semantic action directive $\hat{a}_t$ conditioned on the goal $I_{\text{goal}}$ and the accumulated causal history:

$$\hat{a}_t = \Pi_\theta(I_{\text{goal}},\ \mathcal{P}_t \mid \mathcal{G}_t^{\text{CST}}), \tag{3}$$

which is deterministically instantiated into a 6-DoF pose by resolving object identities against the geometry of STF-Tokens stored in $\mathcal{G}_t^{\text{CST}}$:

$$a_t = \Phi_{\text{inst}}(\hat{a}_t,\ \mathcal{G}_t^{\text{CST}}). \tag{4}$$

This decoupled design preserves VLM semantic reasoning while delegating precise coordinate extraction to deterministic geometry resolution, enabling training-free deployment across various scenarios. We elaborate on each stage in Sec. 3.3–3.5.

### 3.2 Object-Centric Perception

To construct scene representations prioritizing task-relevant entities, we employ a two-stage perception pipeline. Given the RGB-D observation $(I_t, D_t)$, goal specification $I_{\text{goal}}$, and accumulated memory $\mathcal{M}_{<t}$, we prompt a VLM to identify task-relevant objects and output open-vocabulary descriptors $\mathcal{D} = \{d_1, \ldots, d_N\}$ [47,78]. The goal specification $I_{\text{goal}}$ accepts either a reference goal image or a natural language instruction, both applicable in real-world and simulation settings. Conditioning on $\mathcal{M}_{<t}$ allows the descriptor set to maintain consistent object identities across steps, implicitly filtering task-irrelevant entities and reducing computational overhead in downstream modules. The resulting descriptors condition an open-vocabulary segmentation model to produce pixel-accurate masks $\{M_i\}_{i=1}^{N} = \text{SAM3}(I_t, \mathcal{D})$ [6], where $M_i \in \{0,1\}^{H \times W}$. Combined with the depth map $D_t$ [69], each mask defines a spatially coherent region for 3D point cloud extraction, providing the inputs to $\Phi_{\text{enc}}$ for constructing $\mathcal{T}_t^{\text{STF}}$ as in Sec. 3.3.

### 3.3 Spatio-Temporal Token Fusion

Given instance masks $\{M_i\}_{i=1}^{N}$ and the point cloud from the perception stage, we construct a per-object Spatio-Temporal Fusion Token $\tau_i^t$ by integrating visual evidence with metric geometry into a unified, object-centric representation; the full set $\mathcal{T}_t^{\text{STF}} = \{\tau_i^t\}_{i=1}^{N}$ constitutes the encoder output $\Phi_{\text{enc}}$ defined in Sec. 3.1. For each object $o_i$, the mask $M_i$ is applied to the RGB observation $I_t$ to obtain a masked image $\hat{I}_i^t$, which is passed through the VLM visual encoder [13, 50] to produce a $16 \times 16$ grid of patch tokens. Patches with an IoU exceeding $I_{th}$ relative to $M_i$ are selected, projected into the language space, and aggregated to form the visual evidence $\mathbf{v}_i^t$. Unlike methods that pass multiple masked images directly to the VLM, this design avoids redundant background processing by compressing each object's filtered visual tokens, spatial position, geometry, and temporal trajectory into a compact token sequence, shifting VLM reasoning from raw pixels to structured object instances.This compact-token design is orthogonal to recent efficient-inference methods that reduce VLA or vision-transformer computation through adaptive reasoning, model scheduling, token pruning, speculative decoding, or structural pruning [38–41].

In parallel, we define a shape representation vector $\mathbf{s}_i^t$ constructed from Gaussian distribution parameters and geometric extrema along each spatial axis: $\mathbf{s}_i^t = \{(\mu_a, \sigma_a, a_{\min}, a_{\max}) \mid a \in \{x, y, z\}\}$. Unlike traditional bounding boxes, this captures geometric extent more completely, with the joint mean-standard-deviation encoding suppressing outlier noise while correcting Gaussian approximation bias. The centroid $\mathbf{c}_i^t = \text{median}(\mathbf{P}_i) \in \mathbb{R}^3$ is extracted from the object-specific point cloud $\mathbf{P}_i$. Both $\mathbf{c}_i^t$ and $\mathbf{s}_i^t$ are serialized via the text tokenizer into the same language embedding space as $\mathbf{v}_i^t$, placing visual appearance and 3D geometry within a single context window. This shared space allows the VLM to jointly attend over heterogeneous visual and spatial signals without architectural modification, encoding the complete per-object state as:

$$\tau_i^t = \left\langle \mathbf{v}_i^t,\ \mathbf{c}_i^t,\ \mathbf{s}_i^t,\ t \right\rangle, \tag{5}$$

where $t$ denotes the current timestamp. By grounding visual tokens in explicit 3D geometry and temporal context, STF-Tokens enable deterministic spatial queries and suppress cascading errors in long-horizon manipulation tasks.

### 3.4 4D Causal Spatio-Temporal Graph

While STF-Tokens stabilize per-object spatial perception, long-horizon manipulation additionally requires reasoning about inter-object spatial relations and action causality. We address this by constructing the Causal Spatio-Temporal Graph (CSTG) $\mathcal{G}_t^{\mathrm{CST}}$ [1, 61], which realizes the graph update operator $\Psi_{\mathrm{graph}}$ in Sec. 3.1 by integrating incoming STF-Tokens $\mathcal{T}_t^{\mathrm{STF}}$ with the historical state $\mathcal{G}_{t-1}^{\mathrm{CST}}$ through a spatial scene graph and a causal memory log. At each timestamp $t$, we construct a spatial scene graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ where each node $v_i \in \mathcal{V}_t$ corresponds to an object in $\mathcal{T}_t^{\mathrm{STF}}$ and maintains a sliding-window buffer $\{\tau_i^{t-K+1}, \ldots, \tau_i^t\}$ of length $K$. Edges $e_{ij} \in \mathcal{E}_t$ encode pairwise spatial relations via Euclidean distance $d_{ij} = \|\mathbf{c}_i - \mathbf{c}_j\|_2$ and directional offset $\Delta\mathbf{c}_{ij} = \mathbf{c}_j - \mathbf{c}_i$.

Building upon this structure, semantic events are detected by comparing STF-Tokens across the sliding window, covering both object-level events (e.g., planned displacement, unintended collision, occlusion) and task-level events (e.g., action execution, subtask completion). Each event is recorded with its timestamp, location, and causal source in the memory log $\mathcal{H}_t^K$, which is integrated with the spatial graph to form the complete CSTG:

$$\mathcal{G}_t^{\mathrm{CST}} = \left(\mathcal{G}_t,\ \mathcal{H}_t^K\right). \tag{6}$$

This structured representation enables deterministic verification of action preconditions at each planning step, triggering re-planning upon detected failure to prevent error accumulation over long horizons.

### 3.5 VLM-based Causal Reasoning and Planning

At each planning step, RoboStream assembles the context-augmented prompt $\mathcal{P}_t$ from the CSTG $\mathcal{G}_t^{\mathrm{CST}}$ encoding current object states and spatial relations, the annotated observation with object identity labels overlaid [68], and the goal specification $I_{\mathrm{goal}}$. This fusion of structured causal memory with grounded visual evidence enables $\Pi_\theta$ to reason jointly over historical context and current scene geometry. The VLM then performs chain-of-thought spatial verification [63], reviewing historical states, checking action preconditions against object geometry, and resolving causal conflicts from occlusion or dependencies, before generating the semantic action directive $\hat{a}_t$ as defined in Eq. 3. Finally, $\hat{a}_t$ is instantiated into the 6-DoF action $a_t$ via $\Phi_{\mathrm{inst}}$ by resolving object identities against STF-Token geometry stored in $\mathcal{G}_t^{\mathrm{CST}}$ (Sec. 3.1).

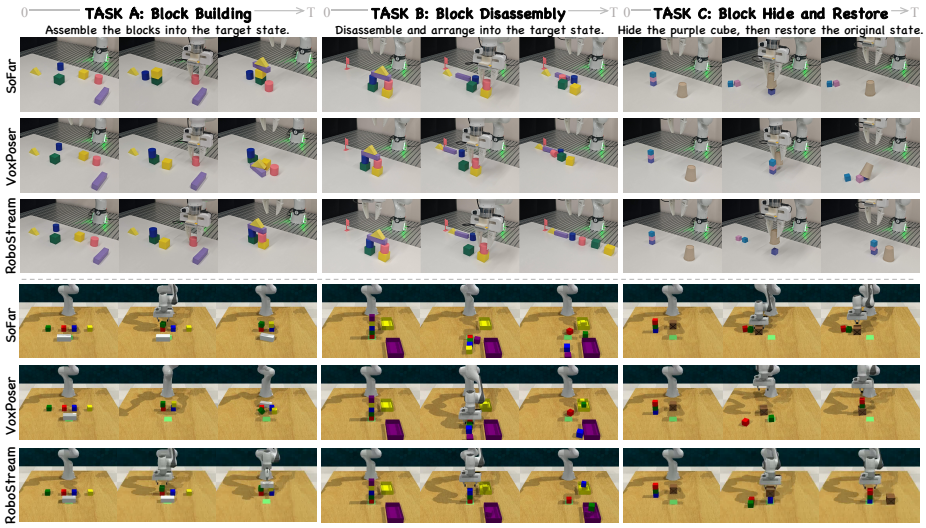


**Fig. 3: Qualitative results in real-world and simulated manipulation tasks.** We evaluate on three challenging tasks: (a) Block Building, demanding precise bottom-to-top assembly toward a target configuration; (b) Block Disassembly, requiring structured top-to-bottom deconstruction and rearrangement; and (c) Block Hide and Restore, requiring causal memory maintenance under full occlusion. RoboStream consistently outperforms SoFar [47] and VoxPoser [26], particularly in spatio-temporal perception and causal memory over extended horizons.

# 4 Experiments

## 4.1 Experimental Setup

We evaluate RoboStream across five benchmarks spanning physical and simulated environments, covering real-world manipulation on a Franka Research 3 arm (Sec. 4.2), long-horizon tasks on RLBench [31] (Sec. 4.4), zero-shot generalization on SIMPLER [37] (Sec. 4.3), and 6-DoF spatial reasoning on SpatialBench [47] and Open6DOR V2 [12] (Sec. 4.5). Short-horizon evaluation validates that the geometry grounding of STF-Tokens suppresses spatial hallucination in individual actions; long-horizon evaluation tests whether the full framework sustains causal state tracking and correct action ordering. We instantiate RoboStream with three VLM backbones, Qwen3-VL-8B, Qwen3-VL-32B, and Qwen3-VL-235B [67], hereafter referred to as RoboStream-8B, RoboStream-32B, and RoboStream-235B respectively. All variants operate without environment-specific adaptation or fine-tuning across any benchmark. We compare against state-of-the-art baselines including SoFar [47], VoxPoser [26], and RT-2-X [81], measuring execution success rate for manipulation tasks and accuracy for spatial reasoning.

To accommodate task diversity, we adopt two goal-specification strategies. For real-world experiments and RLBench long-horizon tasks, where target configurations are difficult to fully capture in text, we employ image-guided goal

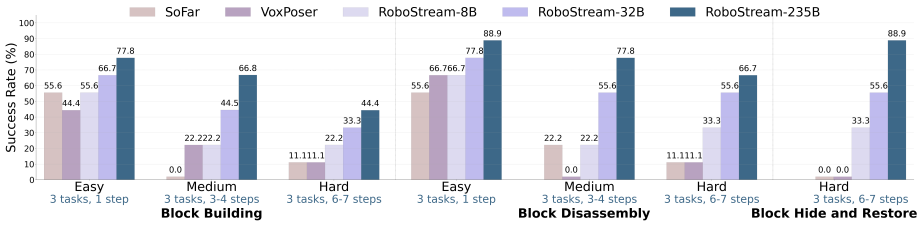


**Fig. 4: Performance comparison on zero-shot real-world manipulation tasks.** We design 21 diverse tasks covering block building (Task A), block disassembly (Task B), and block hide and restore (Task C).

specification with images depicting the desired final state, such as blocks arranged in a target tower configuration. For SIMPLER short-horizon tasks, and tasks involving occlusion or state restoration, we employ text-guided natural language instructions, such as "hide the blue cube, then restore the original state."

### 4.2 Real-World Robotic Manipulation

**Tasks and Evaluation.** We construct 21 real-world tasks on a Franka Research 3 (FR3) arm equipped with a parallel gripper and a front-view Intel RealSense D435i RGB-D camera, utilizing 17 colored objects, as shown in Fig. 3. The tasks are categorized into three types evaluating different cognitive and physical reasoning dimensions. Task A (Block Building) assesses spatio-temporal grounding and precondition adherence by stacking objects bottom-to-top across three difficulty levels (Easy, Medium, Hard), requiring the robot to verify base stability before proceeding. Task B (Block Disassembly) evaluates sequential state tracking and support relation maintenance during top-to-bottom removal, requiring accurate causal memory of stacking order and support dependencies. Task C (Block Hide and Restore) stress-tests causal memory and object permanence under full occlusion, where the robot must conceal a block beneath a cup, execute distractor steps, and restore the original configuration by retrieving object identity and last known pose from causal history. Tasks A and B are specified by goal images, while Task C is guided by natural language instructions. Each task is executed three times.

**Results.** As shown in Fig. 4, all RoboStream variants consistently outperform baselines across all difficulty levels, with success rates scaling proportionally with backbone capacity. On Hard-level block building and disassembly tasks, RoboStream-235B achieves success rates of 44.4% and 66.7%, respectively. In the hard block-building setting, SoFar [47] and VoxPoser [26] struggle to exceed 11.1% without persistent causal tracking and physical state evolution awareness. In Task C (Block Hide and Restore), requiring accurate tracking under total occlusion, RoboStream-235B achieves 88.9% while even our lightweight 8B variant maintains 33.3%, whereas SoFar and VoxPoser completely fail at 0%. This contrast highlights that structured spatio-temporal memory is far more decisive for hidden-state

**Table 1. Quantitative success rate comparison on the SIMPLER benchmark**. We evaluate: (a) the Google Robot setup under Variant Aggregation (VA) and Visual Matching (VM) protocols, and (b) the WidowX + Bridge setup. For the WidowX tasks, we use the following abbreviations: Spoon (Put Spoon on Towel), Carrot (Put Carrot on Plate), Stack (Stack Green Block on Yellow Block), and Egg (Put Eggplant in Basket). "Zero" in the Data column indicates zero-shot evaluation.

**(a) Google Robot Setup**

| Set | Policy | Pick Coke | Move Near |
|---|---|---|---|
| VA | RT-1-X [3] | 49.0 | 32.3 |
| | RT-2-X [81] | 82.3 | 79.2 |
| | Octo-B [60] | 0.6 | 3.1 |
| | OpenVLA [34] | 54.5 | 47.7 |
| | SoFar [47] | 90.7 | 74.0 |
| | **RoboStream-8B** | **94.2** | **79.8** |
| VM | RT-1-X [3] | 56.7 | 31.7 |
| | RT-2-X [81] | 78.7 | 77.9 |
| | Octo-B [60] | 17.0 | 4.2 |
| | OpenVLA [34] | 16.3 | 46.2 |
| | SoFar [47] | 92.3 | 91.7 |
| | **RoboStream-8B** | **95.7** | **95.8** |

**(b) WidowX + Bridge Setup**

| Policy | Data | Spoon | Carrot | Stack | Egg | Avg |
|---|---|---|---|---|---|---|
| RT-1-X [3] | OXE | 0.0 | 4.2 | 0.0 | 0.0 | 1.1 |
| Octo-B [60] | OXE | 12.5 | 8.3 | 0.0 | 43.1 | 16.0 |
| Octo-S [60] | OXE | 47.2 | 9.7 | 4.2 | 56.9 | 30.0 |
| OpenVLA [34] | OXE | 0.0 | 0.0 | 0.0 | 4.1 | 1.0 |
| RoboVLM [36] | OXE | 20.8 | 25.0 | 8.3 | 0.0 | 13.5 |
| RoboVLM [36] | Bridge | 29.2 | 25.0 | 12.5 | 58.3 | 31.3 |
| SpatialVLA [49] | OXE | 20.8 | 20.8 | 25.0 | 70.8 | 34.4 |
| SpatialVLA [49] | Bridge | 16.7 | 25.0 | 29.2 | **100.0** | 42.7 |
| SoFar [47] | Zero | 58.3 | 66.7 | 70.8 | 37.5 | 58.3 |
| **RoboStream-8B** | **Zero** | **62.5** | **71.9** | **77.1** | 87.5 | **74.8** |

tasks than scaling reactive systems. The monotonic gains across model scales confirm that STF-Tokens and CSTG provide complementary capabilities growing increasingly effective with model capacity, yielding robustness that purely reactive systems cannot replicate.

### 4.3 Simulation Object Manipulation Evaluation on SIMPLER

We evaluate zero-shot cross-embodiment generalization on the SIMPLER benchmark [37] across Google Robot and WidowX + Bridge configurations, representing substantial domain shifts from training environments. These short-horizon tasks isolate STF-Token contributions by testing whether spatial grounding transfers across embodiments and viewpoints, independent of long-horizon reasoning.

**Google Robot Configuration.** As shown in Tab. 1, RoboStream-8B achieves superior success rates on both "Pick Coke Can" and the spatially demanding "Move Near" tasks under VM evaluation, surpassing the prior state-of-the-art SoFar [47]. Notably, RoboStream-8B also outperforms OXE-finetuned RT-2-X [81] by a substantial margin despite operating fully zero-shot. Under the more challenging VA evaluation, RoboStream-8B maintains competitive performance, again exceeding SoFar. Consistent gains across both protocols and tasks with varying geometry confirm stable spatial grounding independent of embodiment.

**WidowX + Bridge Configuration.** RoboStream achieves a higher average success rate across all subtasks compared to zero-shot SoFar and Bridge-finetuned SpatialVLA [49] without requiring any fine-tuning or domain adaptation. Notably, this advantage holds despite SpatialVLA being explicitly trained on in-domain Bridge data, highlighting the superior transferability of our geometry-grounded representation. These results confirm that spatio-temporal reasoning grounded in

**Table 2. Simulation evaluation of long-horizon manipulation on RLBench.** All RoboStream variants are evaluated in a zero-shot manner without in-domain fine-tuning.

| Tasks | SoFar [47] | VoxPoser [26] | RoboStream-8B | RoboStream-32B | RoboStream-235B |
|---|---|---|---|---|---|
| Bridge Between Towers | 52.0 | 8.0 | 76.0 | **88.0** | **88.0** |
| Cover Top Block with Box | 4.0 | 4.0 | 40.0 | 84.0 | **92.0** |
| Cover Bottom Block with Box | 0.0 | 0.0 | 16.0 | 68.0 | **96.0** |
| Place Blocks in Two Containers | **100.0** | 96.0 | **100.0** | **100.0** | **100.0** |
| Place Blocks in Two Cont. (Hard) | 4.0 | 4.0 | 24.0 | 76.0 | **88.0** |
| Stack Five Colors | 4.0 | 4.0 | 60.0 | 72.0 | **80.0** |
| Stack Three Colors | 60.0 | **96.0** | **96.0** | **96.0** | **96.0** |
| Unstack then Stack Four Colors | 0.0 | 0.0 | 52.0 | 76.0 | **84.0** |
| **Average** | 28.0 | 26.5 | 58.0 | 82.5 | **90.5** |

explicit geometric primitives transfers robustly across embodiments and extends naturally from short tasks to subsequent long-horizon benchmarks.

### 4.4 Simulation Long-Horizon Task Evaluation on RLBench

**Tasks and Evaluation.** To assess long-horizon manipulation capabilities, we introduce 8 complex tasks within RLBench [31] (Tab. 2), constructed by extending and reconfiguring native assets. Each method is evaluated over 25 episodes per task using different random seeds. Following Sec. 4.1, tasks are categorized by guidance modality. Five goal-image-guided tasks require visually complex target configurations, including Bridge Between Towers, Place Blocks in Two Containers, Place Blocks in Two Cont. (Hard), and Stack Three/Five Colors. Three text-guided tasks stress-test action ordering, restoration, and object permanence under occlusion, including Cover Top/Bottom Block with Box and Unstack then Stack Four Colors.

**Results.** As shown in Tab. 2, RoboStream consistently outperforms both SoFar [47] and VoxPoser [26] across these long-horizon scenarios. Notably, our framework demonstrates a strong scaling effect, with performance increasing alongside model capacity to reach a 90.5% average success rate on RoboStream-235B without additional training, as larger capacity provides the reasoning depth necessary to maintain logical consistency over extended action sequences. However, scaling the VLM alone is not a panacea for long-horizon stability. Even our smallest variant, RoboStream-8B, substantially exceeds much larger baselines, demonstrating that spatio-temporal reasoning and structured memory are more critical for physical consistency than simply increasing parameter scale. We further compare against memory-augmented approaches SAM2Act [16] and MemoryVLA [53] on four RLBench long-horizon cases; RoboStream outperforms them across all model scales.

### 4.5 6-DoF Spatial Reasoning Evaluation

To evaluate how STF-Tokens bridge high-level visual features with precise instance-level 3D grounding, we extend our evaluation to rotation and full 6-DoF

**Table 3. Spatial reasoning evaluation on 6-DoF SpatialBench [47].** Results are categorized by spatial VQA task types, covering absolute (*abs.*) and relative (*rel.*) reasoning for both position and orientation.

| Method | Position | | Orientation | | Total |
|---|---|---|---|---|---|
| | rel. | abs. | rel. | abs. | |
| GPT-4o [30] | 49.4 | 28.4 | 44.2 | 25.8 | 36.2 |
| SpaceMantis [8] | 33.6 | 29.2 | 27.2 | 25.0 | 28.9 |
| RoboPoint [72] | 43.8 | 30.8 | 33.8 | 25.8 | 33.5 |
| SoFar [47] | 59.6 | 33.8 | **54.6** | 31.3 | 43.9 |
| SoFar-8B [47] | 41.5 | 24.3 | 27.1 | 31.3 | 30.5 |
| SoFar-32B [47] | 60.4 | 41.9 | 45.8 | 33.3 | 45.3 |
| **RoboStream-8B** | 47.2 | 31.1 | 35.4 | 35.4 | 36.8 |
| **RoboStream-32B** | **66.0** | **44.6** | 48.0 | **37.5** | **48.9** |

**Table 4. 6-DoF object rearrangement evaluation on Open6DOR [12].** Results are categorized by position (*Pos.*), rotation (*Rot.*), and integrated 6-DoF success rates.

| Method | Pos. | Rot. | 6-DoF |
|---|---|---|---|
| Dream2Real [33] | 15.9 | 31.3 | 13.5 |
| VoxPoser [26] | 32.6 | - | - |
| Open6DOR-GPT [12] | 74.9 | 41.1 | 35.6 |
| SoFar [47] | 93.0 | 57.0 | 48.7 |
| SoFar-8B [47] | 92.4 | 42.9 | 45.1 |
| SoFar-32B [47] | 93.1 | 54.6 | 48.4 |
| **RoboStream-8B** | 93.1 | 47.6 | 47.3 |
| **RoboStream-32B** | **93.8** | **58.8** | **52.2** |

**Table 5. Ablation study on RoboStream-235B.** Success rates (%) across 8 long-horizon tasks, isolating contributions of CSTG and STF-Tokens.

| *STF-Tokens* | *CSTG* | Bridge Between Towers | Cover Top Block | Cover Bottom Block | Place in Containers | Place in Containers (Hard) | Stack Five Colors | Stack Three Colors | Unstack then Stack | **Average** |
|---|---|---|---|---|---|---|---|---|---|---|
| ✗ | ✗ | 0.0 | 0.0 | 0.0 | 84.0 | 12.0 | 0.0 | 0.0 | 0.0 | 12.0 |
| ✓ | ✗ | 0.0 | 0.0 | 0.0 | 92.0 | 24.0 | 0.0 | 0.0 | 0.0 | 14.5 |
| ✗ | ✓ | 72.0 | 84.0 | 88.0 | **100.0** | 72.0 | 60.0 | **96.0** | 64.0 | 79.5 |
| ✓ | ✓ | **88.0** | **92.0** | **96.0** | **100.0** | **88.0** | **80.0** | **96.0** | **84.0** | **90.5** |

pose estimation. Following SoFar [47], we incorporate PointSO for 6-DoF prediction and evaluate on Open6DOR V2 [12] and 6-DoF SpatialBench [47]. Since SoFar serves as our primary baseline, we include variants using Qwen3-VL-8B and Qwen3-VL-32B backbones for fair cross-scale comparison with RoboStream.

**Spatial Reasoning on 6-DoF SpatialBench.** As shown in Tab. 3, RoboStream-32B achieves superior overall performance, outperforming both SoFar and GPT-4o [30]. The most pronounced gains appear on absolute and relative position tasks, where pure image-text alignment is most susceptible to spatial ambiguity. These results confirm that the core advantage of STF-Tokens lies not in coordinate injection alone, but in serving as a geometric anchor providing deterministic spatial reference when visual features and language descriptions conflict.

**6-DoF Object Rearrangement on Open6DOR V2.** As shown in Tab. 4, RoboStream-32B consistently outperforms SoFar across position, rotation, and integrated 6-DoF metrics. Unlike baselines treating spatial coordinates as text annotations, STF-Tokens distill per-object visual evidence into structured instance representations encoding centroid and Gaussian shape. This shift from pixel-level to object-level reasoning underlies consistent gains across all tracks, most notably on 6-DoF where tight placement constraints demand precise geometric grounding and spatial alignment.

### 4.6 Ablation Study

We ablate RoboStream-235B across all 8 long-horizon RLBench [31] tasks to isolate the contribution of each architectural component (Tab. 5). Removing the CSTG module (*w/o CSTG*) causes the average success rate to collapse from 90.5% to 14.5%, confirming that persistent state tracking is the foundational driver for long-horizon tasks. However, while CSTG provides logical structure, STF-Tokens provide physical precision. Removing STF-Tokens (*w/o STF-Tokens*) results in an 11.0% absolute drop (from 90.5% to 79.5%), demonstrating that without explicit 3D geometric grounding, even a CSTG-equipped agent struggles with compounding spatial inaccuracies during contact-rich execution. The mutual dependence of both components becomes evident when removing them simultaneously (*w/o Both*), which yields only 12.0% success rate. Notably, comparing this to the *w/o CSTG* variant (14.5%) shows that STF-Tokens still provide marginal improvement in short-horizon spatial accuracy even when long-horizon logic fails. Their full potential, however, is unlocked only when coupled with the CSTG, where CSTG ensures correct action sequencing and STF-Tokens ensure each step is executed with deterministic spatial precision.

## 5 Limitations

As a decoupled planning-execution framework, RoboStream shares a limitation inherent to the broader VLM-for-robotics paradigm. Imperfections in any submodule, including unstable grasping or perceptual uncertainty, can propagate into execution failures even when high-level planning is correct. This reflects a fundamental tension between the semantic generalization of VLM-based planners and the physical precision demanded by low-level control, a challenge broadly applicable to decoupled systems beyond RoboStream. Future work includes exploring hybrid approaches that bridge VLM planner generalization with end-to-end execution fluency, either by adopting a Vision-Language-Action (VLA) model as the downstream controller for smoother and more compliant motion, or by developing a fully end-to-end VLA that internalizes spatio-temporal reasoning and causal memory within the action generation loop. Extending RoboStream to contact-rich, dexterous, and open-world manipulation also constitutes a promising direction.

## 6 Conclusion

RoboStream addresses two key deficits behind long-horizon VLM planning failures: ungrounded spatial perception and untracked causal history. STF-Tokens convert per-object appearance and 3D geometry into identity-persistent primitives, while the CSTG records action-induced state transitions and preserves hidden object states under occlusion. Across RLBench, SIMPLER, SpatialBench, Open6DOR V2, and real-world Franka experiments, RoboStream consistently improves long-horizon stability, showing that explicit spatio-temporal grounding and persistent memory are more decisive than model scaling alone.

**Acknowledgments.** This work is supported by the National Key R&D Program of China (2022ZD0161700), the National Natural Science Foundation of China (Grant Nos. 92467204 and 62472249), and the Shenzhen Science and Technology Program (Grant No. KJZD20240903102300001). This study was also supported by the Shenzhen Science and Technology Program (Grant No. JCYJ20250604145014018) and the Natural Science Foundation of Top Talent of SZTU (Grant No. GDRC202413). We would like to thank YuanxingGuangnian Robotics for sponsoring this research.

## References

1. Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J., Savarese, S.: 3d scene graph: A structure for unified semantics, 3d space, and camera. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5664–5673 (2019)
2. Belkhale, S., Ding, T., Xiao, T., Sermanet, P., Vuong, Q., Tompson, J., Chebotar, Y., Dwibedi, D., Sadigh, D.: Rt-h: Action hierarchies using language. arXiv preprint arXiv:2403.01823 (2024)
3. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022)
4. Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., Ibarz, J., Irpan, A., Jang, E., Julian, R., et al.: Do as i can, not as i say: Grounding language in robotic affordances. In: Conference on robot learning. pp. 287–318. PMLR (2023)
5. Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., Zhao, B.: Spatialbot: Precise spatial understanding with vision language models. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 9490–9498. IEEE (2025)
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
7. Cheang, C., Chen, S., Cui, Z., Hu, Y., Huang, L., Kong, T., Li, H., Li, Y., Liu, Y., Ma, X., et al.: Gr-3 technical report. arXiv preprint arXiv:2507.15493 (2025)
8. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024)
9. Chen, S., Uy, M.A., Song, C.H., Ladhak, F., Murali, A., Qu, Q., Birchfield, S., Blukis, V., Tremblay, J.: Spacetools: Tool-augmented spatial reasoning via double interactive rl. arXiv preprint arXiv:2512.04069 (2025)
10. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. Advances in Neural Information Processing Systems **37**, 135062–135093 (2024)
11. Daxberger, E., Wenzel, N., Griffiths, D., Gang, H., Lazarow, J., Kohavi, G., Kang, K., Eichner, M., Yang, Y., Dehghan, A., et al.: Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7395–7408 (2025)

12. Ding, Y., Geng, H., Xu, C., Fang, X., Zhang, J., Wei, S., Dai, Q., Zhang, Z., Wang, H.: Open6dor: Benchmarking open-instruction 6-dof object rearrangement and a vlm-based approach. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7359–7366. IEEE (2024)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. In: International Conference on Machine Learning. pp. 8469–8488. PMLR (2023)
15. Du, Y., Konyushkova, K., Denil, M., Raju, A., Landon, J., Hill, F., de Freitas, N., Cabi, S.: Vision-language models as success detectors. arXiv preprint arXiv:2303.07280 (2023)
16. Fang, H., Grotz, M., Pumacay, W., Wang, Y.R., Fox, D., Krishna, R., Duan, J.: Sam2act: Integrating visual foundation model with a memory architecture for robotic manipulation. arXiv preprint arXiv:2501.18564 (2025)
17. Fang, K., Liu, F., Abbeel, P., Levine, S.: Moka: Open-world robotic manipulation through mark-based visual prompting. Robotics: Science and Systems XX (2024)
18. Gu, C., Liu, J., Chen, H., Huang, R., Wuwu, Q., Liu, Z., Li, X., Li, Y., Zhang, R., Jia, P., et al.: Manualvla: A unified vla model for chain-of-thought manual generation and robotic manipulation. arXiv preprint arXiv:2512.02013 (2025)
19. Han, S., Qiu, B., Liao, Y., Huang, S., Gao, C., YAN, S., Liu, S.: Robocerebra: A large-scale benchmark for long-horizon robotic manipulation evaluation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
20. Han, Y., Chi, C., Zhou, E., Rong, S., An, J., Wang, P., Wang, Z., Sheng, L., Zhang, S.: Tiger: Tool-integrated geometric reasoning in vision-language models for robotics. arXiv preprint arXiv:2510.07181 (2025)
21. Hu, Y., Lin, F., Zhang, T., Yi, L., Gao, Y.: Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. In: First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024 (2024)
22. Huang, H., Lin, F., Hu, Y., Wang, S., Gao, Y.: Copa: General robotic manipulation through spatial constraints of parts with foundation models. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9488–9495. IEEE (2024)
23. Huang, S., Chang, H., Liu, Y., Zhu, Y., Dong, H., Boularias, A., Gao, P., Li, H.: A3vlm: Actionable articulation-aware vision language model. In: Conference on Robot Learning. pp. 1675–1690. PMLR (2025)
24. Huang, W., Abbeel, P., Pathak, D., Mordatch, I.: Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In: International conference on machine learning. pp. 9118–9147. PMLR (2022)
25. Huang, W., Wang, C., Li, Y., Zhang, R., Fei-Fei, L.: Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In: Conference on Robot Learning. pp. 4573–4602. PMLR (2025)
26. Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., Fei-Fei, L.: Voxposer: Composable 3d value maps for robotic manipulation with language models. In: Conference on Robot Learning. pp. 540–562. PMLR (2023)

27. Huang, W., Xia, F., Shah, D., Driess, D., Zeng, A., Lu, Y., Florence, P., Mordatch, I., Levine, S., Hausman, K., et al.: Grounded decoding: Guiding text generation with grounded models for embodied agents. Advances in Neural Information Processing Systems **36**, 59636–59661 (2023)
28. Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., et al.: Inner monologue: Embodied reasoning through planning with language models. In: Conference on Robot Learning. pp. 1769–1782. PMLR (2023)
29. Huang, Y., Wen, K., Gao, R., Liu, D., Lou, Y., Wu, J., Xu, J., Zhang, J., Yang, Z., Lin, Y., Li, C., Pan, P., Lu, J., Jiang, J., Ding, X., Huang, Y., Wang, Z.: Thinking in dynamics: How multimodal large language models perceive, track, and reason dynamics in physical 4d world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 33446–33456 (2026)
30. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
31. James, S., Ma, Z., Arrojo, D.R., Davison, A.J.: Rlbench: The robot learning benchmark & learning environment. IEEE Robotics and Automation Letters **5**(2), 3019–3026 (2020)
32. Jiang, Y., Gupta, A., Zhang, Z., Wang, G., Dou, Y., Chen, Y., Fei-Fei, L., Anandkumar, A., Zhu, Y., Fan, L.: Vima: General robot manipulation with multimodal prompts. In: International Conference on Machine Learning. pp. 14975–15022. PMLR (2023)
33. Kapelyukh, I., Ren, Y., Alzugaray, I., Johns, E.: Dream2Real: Zero-shot 3D object rearrangement with vision-language models. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 4796–4803. IEEE (2024)
34. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E.P., Sanketi, P.R., Vuong, Q., et al.: Openvla: An open-source vision-language-action model. In: Conference on Robot Learning. pp. 2679–2713. PMLR (2025)
35. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and brain sciences **40**, e253 (2017)
36. Li, X., Li, P., Qian, L., Liu, M., Wang, D., Liu, J., Kang, B., Ma, X., Wang, X., Guo, D., et al.: What matters in building vision–language–action models for generalist robots. Nature Machine Intelligence **8**, 158–172 (2026). https://doi.org/10.1038/s42256-025-01168-7
37. Li, X., Hsu, K., Gu, J., Mees, O., Pertsch, K., Walke, H.R., Fu, C., Lunawat, I., Sieh, I., Kirmani, S., et al.: Evaluating real-world robot manipulation policies in simulation. In: Conference on Robot Learning. pp. 3705–3728. PMLR (2025)
38. Li, Y., Feng, J., Meng, Y., Ji, K., Tang, C., Wen, X., Xia, S., Wang, Z., Zhu, W.: Ts-dp: Reinforcement speculative decoding for temporal adaptive diffusion policy acceleration. arXiv preprint arXiv:2512.15773 (2025)
39. Li, Y., Liu, H., Ji, K., Meng, Y., Fan, J., Wang, Y., Qin, S., Wu, C., Xia, S.T., Wang, Z.: Elegantvla: Learning when to think for efficient vision-language-action models. arXiv preprint arXiv:2605.29438 (2026)
40. Li, Y., Meng, Y., Sun, Z., Ji, K., Tang, C., Fan, J., Ma, X., Xia, S., Wang, Z., Zhu, W.: Sp-vla: A joint model scheduling and token pruning approach for vla model acceleration. arXiv preprint arXiv:2506.12723 (2025)
41. Li, Y., Tang, C., Meng, Y., Fan, J., Chai, Z., Ma, X., Wang, Z., Zhu, W.: Prance: Joint token-optimization and structural channel-pruning for adaptive vit inference. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)

42. Li, Y., Deng, Y., Zhang, J., Jang, J., Memmel, M., Garrett, C.R., Ramos, F., Fox, D., Li, A., Gupta, A., et al.: Hamster: Hierarchical action models for open-world robot manipulation. In: The Thirteenth International Conference on Learning Representations (2025)
43. Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., Zeng, A.: Code as policies: Language model programs for embodied control. In: 2023 IEEE International conference on robotics and automation (ICRA). pp. 9493–9500. IEEE (2023)
44. Liu, H., Chen, A., Zhu, Y., Swaminathan, A., Kolobov, A., Cheng, C.A.: Interactive robot learning from verbal correction. In: 2nd Workshop on Language and Robot Learning: Language as Grounding (2023)
45. Mao, W., Zhong, W., Jiang, Z., Fang, D., Zhang, Z., Lan, Z., Li, H., Jia, F., Wang, T., Fan, H., et al.: Robomatrix: A skill-centric hierarchical framework for scalable robot task planning and execution in open-world. arXiv preprint arXiv:2412.00171 (2024)
46. Mu, Y., Zhang, Q., Hu, M., Wang, W., Ding, M., Jin, J., Wang, B., Dai, J., Qiao, Y., Luo, P.: Embodiedgpt: Vision-language pre-training via embodied chain of thought. Advances in Neural Information Processing Systems **36**, 25081–25094 (2023)
47. Qi, Z., Zhang, W., Ding, Y., Dong, R., Yu, X., Li, J., Xu, L., Li, B., He, X., Fan, G., et al.: Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
48. Qin, Y., Kang, L., Song, X., Yin, Z., Liu, X., Liu, X., Zhang, R., Bai, L.: Robofactory: Exploring embodied agent collaboration with compositional constraints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10075–10085 (2025)
49. Qu, D., Song, H., Chen, Q., Yao, Y., Ye, X., Ding, Y., Wang, Z., Gu, J., Zhao, B., Wang, D., et al.: Spatialvla: Exploring spatial representations for visual-language-action model. arXiv preprint arXiv:2501.15830 (2025)
50. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
51. Ray, A., Duan, J., Brown II, E.L., Tan, R., Bashkirova, D., Hendrix, R., Ehsani, K., Kembhavi, A., Plummer, B.A., Krishna, R., et al.: Sat: Dynamic spatial aptitude training for multimodal language models. In: Second Conference on Language Modeling (2025)
52. Sharma, P., Sundaralingam, B., Blukis, V., Paxton, C., Hermans, T., Torralba, A., Andreas, J., Fox, D.: Correcting robot plans with natural language feedback. arXiv preprint arXiv:2204.05186 (2022)
53. Shi, H., Xie, B., Liu, Y., Sun, L., Liu, F., Wang, T., Zhou, E., Fan, H., Zhang, X., Huang, G.: Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. In: International Conference on Learning Representations (2026), https://openreview.net/forum?id=54U3XHf7qq, accessed 16 July 2026
54. Shi, L., Hu, Z., Zhao, T., Sharma, A., Pertsch, K., Luo, J., Levine, S., Finn, C.: Yell at your robot: Improving on-the-fly from language corrections. Robotics: Science and Systems XX (2024)
55. Shi, L.X., Ichter, B., Equi, M.R., Ke, L., Pertsch, K., Vuong, Q., Tanner, J., Walling, A., Wang, H., Fusai, N., et al.: Hi robot: Open-ended instruction following with

hierarchical vision-language-action models. In: International Conference on Machine Learning. pp. 54919–54933. PMLR (2025)
56. Shridhar, M., Manuelli, L., Fox, D.: Perceiver-actor: A multi-task transformer for robotic manipulation. In: Conference on Robot Learning. pp. 785–799. PMLR (2023)
57. Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., Garg, A.: Progprompt: Generating situated robot task plans using large language models. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 11523–11530. IEEE (2023)
58. Song, C.H., Blukis, V., Tremblay, J., Tyree, S., Su, Y., Birchfield, S.: Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15768–15780 (2025)
59. Tan, H., Hao, X., Chi, C., Lin, M., Lyu, Y., Cao, M., Liang, D., Chen, Z., Lyu, M., Peng, C., et al.: Roboos: A hierarchical embodied framework for cross-embodiment and multi-agent collaboration. arXiv preprint arXiv:2505.03673 (2025)
60. Team, O.M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al.: Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213 (2024)
61. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3961–3970 (2020)
62. Wang, X., Ma, W., Zhang, T., de Melo, C.M., Chen, J., Yuille, A.: Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24669–24679. IEEE (2025)
63. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
64. Wen, J., Zhu, Y., Li, J., Tang, Z., Shen, C., Feng, F.: Dexvla: Vision-language model with plug-in diffusion expert for general robot control. In: Conference on Robot Learning. pp. 3094–3114. PMLR (2025)
65. Wen, K., Huang, Y., Chen, R., Zheng, H., Lin, Y., Pan, P., Li, C., Cong, W., Zhang, J., Lu, J., Lin, C., Wang, D., Yan, Z., Xu, H., Theiss, J., Huang, Y., Ding, X., Ranjan, R., Fan, Z.: Dynamicverse: A physically-aware multimodal framework for 4d world modeling. In: Advances in Neural Information Processing Systems. vol. 38 (2025), https://proceedings.neurips.cc/paper_files/paper/2025/hash/9c20f16b05f5e5e70fa07e2a4364b80e-Abstract-Conference.html, accessed 16 July 2026
66. Wu, J., Antonova, R., Kan, A., Lepert, M., Zeng, A., Song, S., Bohg, J., Rusinkiewicz, S., Funkhouser, T.: Tidybot: Personalized robot assistance with large language models. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3546–3553. IEEE (2023)
67. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
68. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441 (2023)
69. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)

70. Yang, Y., Sun, J., Kou, S., Wang, Y., Deng, Z.: Lohovla: A unified vision-language-action model for long-horizon embodied tasks. arXiv preprint arXiv:2506.00411 (2025)
71. Yu, W., Gileadi, N., Fu, C., Kirmani, S., Lee, K.H., Arenas, M.G., Chiang, H.T.L., Erez, T., Hasenclever, L., Humplik, J., et al.: Language to rewards for robotic skill synthesis. In: Conference on Robot Learning. pp. 374–404. PMLR (2023)
72. Yuan, W., Duan, J., Blukis, V., Pumacay, W., Krishna, R., Murali, A., Mousavian, A., Fox, D.: Robopoint: A vision-language model for spatial affordance prediction in robotics. In: Conference on Robot Learning. pp. 4005–4020. PMLR (2025)
73. Zawalski, M., Chen, W., Pertsch, K., Mees, O., Finn, C., Levine, S.: Robotic control via embodied chain-of-thought reasoning. In: Conference on Robot Learning. pp. 3157–3181. PMLR (2025)
74. Zeng, A., Attarian, M., Choromanski, K.M., Wong, A., Welker, S., Tombari, F., Purohit, A., Ryoo, M.S., Sindhwani, V., Lee, J., et al.: Socratic models: Composing zero-shot multimodal reasoning with language. In: The Eleventh International Conference on Learning Representations (2023)
75. Zhang, K., Yin, Z.H., Ye, W., Gao, Y.: Learning manipulation skills through robot chain-of-thought with sparse failure guidance. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8012–8018. IEEE (2025)
76. Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1702–1713 (2025)
77. Zhen, H., Qiu, X., Chen, P., Yang, J., Yan, X., Du, Y., Hong, Y., Gan, C.: 3d-vla: A 3d vision-language-action generative world model. In: International Conference on Machine Learning. pp. 61229–61245. PMLR (2024)
78. Zhou, E., An, J., Chi, C., Han, Y., Rong, S., Zhang, C., Wang, P., Wang, Z., Huang, T., Sheng, L., et al.: Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
79. Zhou, E., Su, Q., Chi, C., Zhang, Z., Wang, Z., Huang, T., Sheng, L., Wang, H.: Code-as-monitor: Constraint-aware visual programming for reactive and proactive robotic failure detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6919–6929 (2025)
80. Zhu, K., Bai, F., Xiang, Y., Cai, Y., Chen, X., Li, R., Wang, X., Dong, H., Yang, Y., Fan, X., et al.: Dexflywheel: A scalable and self-improving data generation framework for dexterous manipulation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
81. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023)