Novelty Assessment Report

Paper: RoboInter: A Holistic Intermediate Representation Suite Towards Robotic Manipulation
PDF URL: https://openreview.net/pdf?id=PGUC3mmMoi
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Large language and vision-language models have inspired end-to-end vision-language-action (VLA) systems in robotics, yet existing robot datasets remain costly, embodiment-specific, and insufficient, limiting robustness and generalization. Recent approaches address this by adopting a plan-then-execute paradigm, where high-level plans are generated before translating into low-level actions, but their success depends on fine-grained intermediate supervision that current datasets lack. To fill this gap, we present the RoboInter Manipulation Suite, a unified resource for data, benchmarking, and modeling of intermediate representations. It includes RoboInter-Tool, a lightweight GUI for semi-automatic per-frame annotation of embodied videos, and RoboInter-Data, a human-verified dataset with over 200k episodes across 571 diverse scenes, offering dense per-frame alignment across more than nine intermediate categories and surpassing prior work in both scale and quality. Building on this foundation, RoboInter-VQA introduces 8 spatial and 20 temporal embodied QA categories to benchmark and enhance the embodied capabilities of current large vision-language models, while RoboInter-VLA provides a flexible plan-then-execute framework with modular and end-to-end variants that link planning to execution. Together, these contributions establish RoboInter Manipulation Suite as a foundation for advancing generalizable and robust robotic learning through fine-grained intermediate supervision.

Core Task Landscape

This paper addresses: **intermediate representation learning for robotic manipulation**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Visual and Multi-Modal Representation Learning**
- **Language-Grounded and Vision-Language Representations**
- **Object-Centric and Structured Representations**
- **World Models and Predictive Dynamics**
- **Information-Theoretic and Latent Compression**
- **Policy Learning with Intermediate Representations**
- **Domain Adaptation and Transfer Learning**
- **Active Perception and Embodied Interaction**
- **Unified and Holistic Representation Frameworks**
- **Adversarial Robustness and Security**

Complete Taxonomy Tree

- intermediate representation learning for robotic manipulation Survey Taxonomy
- Visual and Multi-Modal Representation Learning
  - Self-Supervised Visual Pretraining (4 papers)
  - [1] Language-driven representation learning for robotics (Siddharth Karamcheti, 2023) View paper
  - [18] Multi-View Masked World Models for Visual Robotic Manipulation (Seo, 2023) View paper
  - [33] Manipulate by seeing: Creating manipulation controllers from pre-trained representations (Jianren Wang, 2023) View paper
  - [44] Robots Pre-train Robots: Manipulation-Centric Robotic Representation from Large-Scale Robot Datasets (Jiang, 2024) View paper
  - Multi-Modal Fusion and Pretraining (5 papers)
  - [2] UpViTaL: Unpaired Visual-Tactile Self-Supervised Representation Learning for Dexterous Robotic Manipulation (Guwen Han, 2025) View paper
  - [6] EmbodiedMAE: A Unified 3D Multi-Modal Representation for Robot Manipulation (Ni Fei, 2025) View paper
  - [15] Masked Visual-Tactile Pre-training for Robot Manipulation (Qingtao Liu, 2024) View paper
  - [36] MOSAIC: Learning Unified Multi-Sensory Object Property Representations for Robot Learning via Interactive Perception (Gyan Tatiya, 2023) View paper
  - [46] Contact-aware and multi-modal robotic manipulation (Jialiang, 2025) View paper
  - Spatially-Grounded and Depth-Aware Representations (3 papers)
  - [16] Bridging Perception and Action: Spatially-Grounded Mid-Level Representations for Robot Generalization (Yang, 2025) View paper
  - [17] Self-supervised sim-to-real adaptation for visual robotic manipulation (Rae Jeong, 2020) View paper
  - [24] Learning Sim-to-Real Dense Object Descriptors for Robotic Manipulation (Cao, 2023) View paper
- Language-Grounded and Vision-Language Representations
  - Vision-Language Model Adaptation for Manipulation (4 papers)

    - [3] PEEK: Guiding and Minimal Image Representations for Zero-Shot Generalization of Robot Manipulation Policies (Zhang, 2025) View paper
    - [20] LaVA-Man: Learning Visual Action Representations for Robot Manipulation (Zhu, 2025) View paper
    - [34] RoboGround: Robotic Manipulation with Grounded Vision-Language Priors (Huang Hai-feng, 2025) View paper
    - [39] KALIE: Fine-Tuning Vision-Language Models for Open-World Manipulation Without Robot Data (Grace Tang, 2024) View paper
    - Hierarchical and Symbolic Planning with VLMs (3 papers)
    - [13] Rethinking Intermediate Representation for VLM-based Robot Manipulation (Weiliang Tang, 2025) View paper
    - [22] HAMSTER: Hierarchical Action Models For Open-World Robot Manipulation (Li Yi, 2025) View paper
    - [25] Learning to Recover from Plan Execution Errors during Robot Manipulation: A Neuro-symbolic Approach (Namasivayam K, 2024) View paper
- Object-Centric and Structured Representations
    - Object-Centric Representation Learning (4 papers)
    - [8] Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints (Mingjie Pan, 2025) View paper
    - [9] Is an object-centric representation beneficial for robotic manipulation ? (Alexandre Chapin, 2025) View paper
    - [23] Viola: Imitation learning for vision-based manipulation with object proposal priors (Zhu Yi-feng, 2023) View paper
    - [43] Disentangled Object-Centric Image Representation for Robotic Manipulation (Emukpere, 2025) View paper
    - Keypoint and Affordance Representations (3 papers)
    - [29] The Treachery of Images: Bayesian Scene Keypoints for Deep Policy Learning in Robotic Manipulation (Jan Ole von Hartz, 2023) View paper
    - [32] Self-Supervised Learning of Multi-Object Keypoints for Robotic Manipulation (von Hartz, 2022) View paper
    - [49] RT-Affordance: Affordances are Versatile Intermediate Representations for Robot Manipulation (Soroush Nasiriany, 2024) View paper
    - Graph-Based and Relational Representations (3 papers)
    - [31] Learning graph-structured representations for robotic manipulation (Rezazadeh, 2024) View paper
    - [38] GNN Topology Representation Learning for Deformable Multi-Linear Objects Dual-Arm Robotic Manipulation (Alessio Caporali, 2025) View paper
    - [50] Dynamics Learning With Object-Centric Interaction Networks for Robot Manipulation (Jiayu Wang, 2021) View paper
- World Models and Predictive Dynamics
    - Visual World Models for Manipulation (2 papers)
    - [5] A step toward world models: A survey on robotic manipulation (Zhang, 2025) View paper
    - [27] DreamerRL: empowering representation learning via predictive world models for robot manipulation tasks (Correia, 2025) View paper
    - Flow-Based and 3D Scene Dynamics (3 papers)
    - [7] FlowDreamer: A RGB-D World Model with Flow-based Motion Representations for Robot Manipulation (Guo Jun, 2025) View paper
    - [11] MSGField: A Unified Scene Representation Integrating Motion, Semantics, and Geometry for Robotic Manipulation (Sheng Yu, 2024) View paper
    - [41] Learning 3D Dynamic Scene Representations for Robot Manipulation (Zhenjia Xu, 2022) View paper
    - Deformable Object Dynamics (1 papers)
    - [45] Action-conditional implicit visual dynamics for deformable object manipulation (Bokui Shen, 2024) View paper
- Information-Theoretic and Latent Compression (2 papers)
    - [4] Rethinking latent representations in behavior cloning: An information bottleneck approach for robot manipulation (Shiwen Bai, 2025) View paper
    - [26] Moto: Latent Motion Token as the Bridging Language for Learning Robot Manipulation from Videos (Chen Yi, 2024) View paper
- Policy Learning with Intermediate Representations
    - Imitation Learning with Structured Representations (3 papers)
    - [10] Human-oriented Representation Learning for Robotic Manipulation (Mingxiao Huo, 2023) View paper
    - [35] Manipulator-independent representations for visual imitation (Zhou, 2021) View paper
    - [48] Ag2Manip: Learning Novel Manipulation Skills with Agent-Agnostic Visual and Action Representations (Puhao Li, 2024) View paper
    - Goal-Conditioned Reinforcement Learning (1 papers)
    - [40] MURM: Utilization of Multi-Views for Goal-Conditioned Reinforcement Learning in Robotic Manipulation (Seongwon Jang, 2023) View paper
    - Generative Policy Models (3 papers)
    - [12] DM1: MeanFlow with Dispersive Regularization for 1-Step Robotic Manipulation (Zou Guowei, 2025) View paper
    - [21] Generative artificial intelligence in robotic manipulation: A survey (Zhang Kun, 2025) View paper
    - [37] Zero-Shot Robotic Manipulation with Pretrained Image-Editing Diffusion Models (Black, 2023) View paper
- Domain Adaptation and Transfer Learning (1 papers)
    - [14] Learning for Sequential Manipulation (Driess, 2024) View paper
- Active Perception and Embodied Interaction (2 papers)
    - [28] Learning manipulation by predicting interaction (Jia Zeng, 2024) View paper
    - [42] Vision in Action: Learning Active Perception from Human Demonstrations (Xiong Hao-yu, 2025) View paper
- Unified and Holistic Representation Frameworks ★ (3 papers)
    - [0] RoboInter: A Holistic Intermediate Representation Suite Towards Robotic Manipulation (Anon et al., 2026) View paper
    - [19] XR-1: Towards Versatile Vision-Language-Action Models via Learning Unified Vision-Motion Representations (Fan Shichao, 2025) View paper
    - [30] Splat-MOVER: Multi-Stage, Open-Vocabulary Robotic Manipulation via Editable Gaussian Splatting (Shorinwa, 2024) View paper
- Adversarial Robustness and Security (1 papers)
    - [47] Rethinking the Intermediate Features in Adversarial Attacks: Misleading Robotic Models via Adversarial Distillation (Zhao Ke, 2024) View paper

Narrative

Core task: intermediate representation learning for robotic manipulation. The field organizes itself around several complementary perspectives on how robots should encode sensory input and task structure. Visual and Multi-Modal Representation Learning explores perceptual encodings from camera and tactile streams, often leveraging self-supervised pretraining (e.g., EmbodiedMAE[6], Multi View Masked[18]). Language-Grounded and Vision-Language Representations integrate linguistic instructions with visual observations to ground commands in perception (Language Driven Representation[1], RoboGround[34]). Object-Centric and Structured Representations decompose scenes into entities or keypoints (Dense Object Descriptors[24], Multi Object Keypoints[32]), while World Models and Predictive Dynamics learn forward models that anticipate future states (World Models Survey[5], DreamerRL[27]). Information-Theoretic and Latent Compression methods distill high-dimensional inputs into compact bottlenecks (Information Bottleneck[4], PEEK[3]), and Policy Learning with Intermediate Representations directly couples learned features to action selection. Domain Adaptation and Transfer Learning address sim-to-real gaps (Sim to Real Adaptation[17]), Active Perception and Embodied Interaction emphasize feedback loops between sensing and acting, and Adversarial Robustness and Security consider safety under perturbations (Adversarial Distillation[47]).

A particularly active line of work focuses on Unified and Holistic Representation Frameworks that synthesize multiple modalities, task structures, and learning objectives into a single architecture. RoboInter[0] exemplifies this direction by proposing an integrated approach that combines visual, language, and action representations within a coherent framework, aiming to capture the full spectrum of manipulation-relevant information. This contrasts with more specialized efforts: PEEK[3] emphasizes information-theoretic compression to isolate task-relevant features, while Splat-MOVER[30] leverages 3D Gaussian splatting for spatially grounded scene understanding. The tension between holistic integration and modular specialization remains a central open question—whether a single unified encoder can match or exceed the performance of carefully tailored representations for distinct subtasks. RoboInter[0] sits squarely in the holistic camp, seeking to demonstrate that end-to-end learning over diverse data can yield representations that generalize broadly across manipulation scenarios.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. XR-1: Towards Versatile Vision-Language-Action Models via Learning Unified Vision-Motion Representations

Authors: Fan Shichao, Wu Kun, Che, Zhengping, Wang Xin-hua, et al. (18 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Recent progress in large-scale robotic datasets and vision-language models (VLMs) has advanced research on vision-language-action (VLA) models. However, existing VLA models still face two fundamental challenges: (i) producing precise low-level actions from high-dimensional observations, (ii) bridging domain gaps across heterogeneous data sources, including diverse robot embodiments and human demonstrations. Existing methods often encode latent variables from either visual dynamics or robotic act...

Relationship Analysis

Both papers belong to the Unified and Holistic Representation Frameworks category, providing comprehensive approaches to intermediate representation learning for robotic manipulation. They overlap in addressing the challenge of bridging high-level planning and low-level execution through intermediate representations, with both offering large-scale datasets and unified frameworks. However, RoboInter focuses on dense per-frame annotation of diverse intermediate types (9+ categories including affordances, traces, bounding boxes) with human verification and a plan-then-execute VLA architecture, while XR-1 introduces Unified Vision-Motion Codes (UVMC) as a discrete latent representation learned via dual-branch VQ-VAE to align visual dynamics and robotic motion across heterogeneous embodiments.

---

2. Splat-MOVER: Multi-Stage, Open-Vocabulary Robotic Manipulation via Editable Gaussian Splatting

Authors: Shorinwa, Ola, Tucker, Johnathan, Swann, et al. (11 authors total) | **Year/Venue**: 2024 • Conference on Robot Learning | **URL**: View paper

Abstract

We present Splat-MOVER, a modular robotics stack for open-vocabulary robotic manipulation, which leverages the editability of Gaussian Splatting (GSplat) scene representations to enable multi-stage manipulation tasks. Splat-MOVER consists of: (i) ASK-Splat, a GSplat representation that distills semantic and grasp affordance features into the 3D scene. ASK-Splat enables geometric, semantic, and affordance understanding of 3D scenes, which is critical in many robotics tasks; (ii) SEE-Splat, a real...

Relationship Analysis

Both papers belong to the Unified and Holistic Representation Frameworks category, providing comprehensive systems that integrate multiple intermediate representations for robotic manipulation. While RoboInter focuses on creating a large-scale annotation suite with diverse intermediate representations (traces, affordances, bounding boxes) and a plan-then-execute VLA framework trained on human-verified data, Splat-MOVER takes a modular approach using Gaussian Splatting for real-time 3D scene understanding, editing, and grasp generation. The key difference is that RoboInter emphasizes data-driven learning with extensive annotations and benchmarking, whereas Splat-MOVER emphasizes geometric scene representation and real-time editability for multi-stage manipulation without requiring large-scale training data.

Contributions Analysis

Overall novelty summary. The paper introduces RoboInter, a unified resource combining data annotation tools, a large-scale dataset with dense per-frame intermediate annotations, embodied VQA benchmarks, and a plan-then-execute VLA framework. It resides in the 'Unified and Holistic Representation Frameworks' leaf, which contains only three papers total. This is one of the sparsest leaves in the taxonomy, suggesting that comprehensive frameworks integrating multiple intermediate representation types remain relatively underexplored compared to specialized single-modality or single-task approaches.

The taxonomy reveals that most research concentrates in specialized directions: Visual and Multi-Modal Representation Learning (nine papers across three subcategories), Language-Grounded and Vision-Language Representations (seven papers), and Object-Centric and Structured Representations (ten papers across three subcategories). RoboInter's holistic approach contrasts with these focused efforts—it aims to bridge visual pretraining, language grounding, object-centric reasoning, and policy learning within a single framework. The taxonomy's scope_note for this leaf explicitly highlights integration of 'multiple intermediate representation types' and 'unified benchmarks and tooling,' positioning RoboInter as a synthesis effort rather than a specialized method.

Among 25 candidates examined, the dataset contribution (RoboInter-Data) shows one refutable candidate from six examined, suggesting some overlap with prior large-scale annotation efforts. The VQA benchmark (RoboInter-VQA) and VLA framework (RoboInter-VLA) show no refutable candidates among nine and ten examined respectively, indicating these components appear more novel within the limited search scope. The statistics suggest the data contribution faces more direct prior work, while the benchmark and framework components occupy less crowded territory, though this assessment is constrained by the top-25 semantic search scope.

Given the sparse population of the 'Unified and Holistic Representation Frameworks' leaf and the limited search scope, the work appears to occupy a relatively open research direction. However, the single refutable candidate for the dataset contribution and the modest search scale (25 candidates total) mean this assessment captures only a snapshot of the immediate semantic neighborhood, not an exhaustive field survey. The framework's integration of annotation tools, benchmarks, and modeling represents a systems-level contribution that may be harder to directly compare against specialized prior work.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: RoboInter-Data: Large-scale human-verified dataset with dense per-frame intermediate annotations

Description: A large-scale manipulation dataset containing over 200,000 episodes across 571 scenes with human-verified, dense per-frame annotations of nine intermediate representation categories (subtasks, primitive skills, affordances, target points, gripper/object bounding boxes, traces, contact points, and placement affordance). This dataset surpasses prior work in both scale and annotation quality.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. StreamingCoT: A Dataset for Temporal Dynamics and Multimodal Chain-of-Thought Reasoning in Streaming VideoQA

URL: View paper

Brief Assessment

StreamingCoT[74] focuses on streaming video question answering with temporal dynamics and chain-of-thought reasoning for video understanding tasks, not robotic manipulation datasets with intermediate representations like subtasks, primitive skills, affordances, and gripper trajectories.

---

2. RoboBrain: A Unified Brain Model for Robotic Manipulation from Abstract to Concrete

URL: View paper

Prior Art Analysis

RoboBrain[70] demonstrates that similar prior work exists in creating large-scale manipulation datasets with human-verified, dense annotations of intermediate representations. The candidate paper introduces ShareRobot, a dataset that labels multi-dimensional information including task planning, object affordance, and end-effector trajectory, with diversity and accuracy refined by three human annotators. This directly challenges the novelty claim of RoboInter-Data being the first to provide such comprehensive human-verified annotations at scale.

Evidence

Evidence 1 - **Rationale**: Both papers claim to provide large-scale, human-verified datasets with dense intermediate annotations. ShareRobot explicitly mentions human refinement by three annotators and covers task planning, affordance, and trajectory - similar intermediate representations to RoboInter-Data's nine categories. - **Original**: robointer-data, a human-verified dataset with over 200k episodes across 571 diverse scenes, offering dense perframe alignment across more than nine intermediate categories and surpassing prior work in both scale and quality - **Candidate**: we introduce sharerobot, a high-quality heterogeneous dataset that labels multi-dimensional information such as task planning, object affordance, and end-effector trajectory. sharerobotâs diversity and accuracy have been meticulously refined by three human annotators

Evidence 2 - **Rationale**: The original paper explicitly mentions ShareRobot as prior work that it claims to surpass. However, the candidate (RoboBrain[70]/ShareRobot) demonstrates that such multi-dimensional, human-verified annotation work existed prior to the original paper's submission, challenging the novelty claim of being first to provide this type of dataset. - **Original**: robointer-dataprovides over 200k episodes across 571 distinct scenes, surpassing llarv a niu et al. (2024), ecot zawalski et al. (2024), and sharerobot ji et al. (2025) in both scale and diversity - **Candidate**: we introduce sharerobot, a high-quality heterogeneous dataset that labels multi-dimensional information such as task planning, object affordance, and end-effector trajectory

---

3. Human-in-the-loop Online Rejection Sampling for Robotic Manipulation

URL: View paper

Brief Assessment

Online Rejection Sampling[72] focuses on a post-training method for fine-tuning VLA models using rejection sampling and human-in-the-loop corrections, not on creating large-scale annotated datasets with dense per-frame intermediate representations.

---

4. EPFL-Smart-Kitchen-30: Densely annotated cooking dataset with 3D kinematics to challenge video and language models

URL: View paper

Brief Assessment

Smart Kitchen Dataset[75] focuses on cooking activities with action annotations and 3D kinematics, not robotic manipulation with intermediate representations like subtasks, affordances, and gripper traces.

---

5. DexCanvas: Bridging Human Demonstrations and Robot Learning for Dexterous Manipulation

URL: View paper

Brief Assessment

DexCanvas[73] focuses on dexterous hand manipulation with force/contact annotations from physics simulation, not dense per-frame intermediate representations (subtasks, skills, affordances, traces, bounding boxes) for general robotic manipulation as in the original paper.

---

6. Scalable Vision-Language-Action Model Pretraining for Robotic Manipulation with Real-Life Human Activity Videos

URL: View paper

Brief Assessment

Human Activity Pretraining[71] focuses on pretraining VLA models using unscripted human hand activity videos without annotations, not on creating a human-verified dataset with dense per-frame intermediate annotations for robot manipulation.

Contribution 2: RoboInter-VQA: Spatial and temporal embodied VQA benchmark and training data

Description: A curated embodied visual question answering dataset and benchmark comprising 8 spatial and 20 temporal QA categories designed to systematically evaluate and improve the embodied reasoning and grounding capabilities of vision-language models in manipulation scenarios.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Robovqa: Multimodal long-horizon reasoning for robotics

URL: View paper

Brief Assessment

Robovqa[60] focuses on long-horizon reasoning tasks in office environments with different question types (affordance, success, planning), while RoboInter-VQA emphasizes systematic spatial (8 categories) and temporal (20 categories) QA for manipulation scenarios with dense per-frame annotations.

2. 3DLLM-Mem: Long-Term Spatial-Temporal Memory for Embodied 3D Large Language Model

URL: View paper

Brief Assessment

3DLLM-Mem[67] focuses on 3D spatial-temporal memory for navigation and reasoning in multi-room environments, not manipulation-specific VQA benchmarks with spatial and temporal categories for robotic manipulation tasks.

3. Vlaser: Vision-Language-Action Model with Synergistic Embodied Reasoning

URL: View paper

Brief Assessment

Vlaser[65] focuses on general embodied reasoning across multiple benchmarks (grounding, planning, spatial intelligence) rather than specifically on spatial and temporal VQA categories for manipulation. The candidate does not demonstrate prior work that directly challenges the novelty of RoboInter-VQA's 8 spatial and 20 temporal QA categories designed for manipulation scenarios.

4. Cosmos-reason1: From physical common sense to embodied reasoning

URL: View paper

Brief Assessment

Cosmos Reason[63] focuses on physical common sense reasoning and embodied reasoning across diverse agents (humans, robots, vehicles) with a different ontology structure, rather than specifically targeting spatial and temporal VQA benchmarks for manipulation tasks as in the original paper.

5. SpatialPIN: Enhancing Spatial Reasoning Capabilities of Vision-Language Models through Prompting and Interacting 3D Priors

URL: View paper

Brief Assessment

SpatialPIN[69] focuses on enhancing spatial reasoning through 3D foundation model priors for VQA tasks, not on creating embodied manipulation benchmarks with temporal reasoning categories or dense per-frame annotations for robotic manipulation scenarios.

6. Robotvqaâ□□a scene-graph-and deep-learning-based visual question answering system for robot manipulation

URL: View paper

Brief Assessment

RobotVQA Scene Graph[66] focuses on scene graph generation and object-level visual question answering for manipulation, not on spatial and temporal embodied VQA benchmarks with systematic QA categories as proposed in the original paper.

7. Embodied Intelligence for 3D Understanding: A Survey on 3D Scene Question Answering

URL: View paper

Brief Assessment

3D Scene QA[68] focuses on 3D scene understanding with point clouds and multi-view images, not robotic manipulation scenarios with spatial-temporal embodied VQA for manipulation tasks.

8. ReMEmbR: Building and Reasoning Over Long-Horizon Spatio-Temporal Memory for Robot Navigation

URL: View paper

Brief Assessment

ReMEmbR[64] focuses on long-horizon video question answering for robot navigation with the NavQA dataset, while the original paper targets manipulation tasks with spatial and temporal QA categories for VLMs. The domains and task types differ substantially.

9. Robo2VLM: Visual Question Answering from Large-Scale In-the-Wild Robot Manipulation Datasets

URL: View paper

Brief Assessment

Robo2VLM[62] focuses on VQA generation from robot trajectories using non-visual sensory modalities and manipulation phases, whereas the original paper emphasizes dense per-frame annotations across 8 spatial and 20 temporal QA categories with human verification. The approaches and annotation methodologies differ substantially.

Contribution 3: RoboInter-VLA: Flexible plan-then-execute framework with modular and end-to-end variants

Description: A flexible plan-then-execute framework that supports multiple architectural variants (implicitly-conditioned, explicitly-conditioned, and modular) for robotic manipulation. The framework uses a VLM-based Planner and an Executor, connected through Flexible Chain-of-Thought intermediate representations to bridge high-level planning and low-level action execution.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Recent trends in task and motion planning for robotics: A survey

URL: View paper

Brief Assessment

Task Motion Survey[52] provides a high-level survey of TAMP approaches that integrate abstract reasoning with geometric feasibility, but does not present a specific VLA framework with flexible chain-of-thought representations or architectural variants for robotic manipulation.

2. Human-Aware Reactive Task Planning of Sequential Robotic Manipulation Tasks

URL: View paper

Brief Assessment

Human Aware Planning[51] focuses on human-robot collaboration in manufacturing with mode switching between human-aware and fully automatic operations, not on general plan-then-execute frameworks for robotic manipulation with VLM-based planning and flexible intermediate representations.

---

3. Learning for Sequential Manipulation

URL: View paper

Brief Assessment

Sequential Manipulation[14] focuses on learning methods for sequential manipulation problems in robotics, including amortizing TAMP solutions into learned policies. However, it does not describe the specific flexible plan-then-execute framework with VLM-based Planner, Executor, and Flexible Chain-of-Thought intermediate representations that RoboInter-VLA proposes.

---

4. Practical task and motion planning for robotic food preparation

URL: View paper

Brief Assessment

Food Preparation Planning[53] focuses on task and motion planning specifically for food preparation scenarios, not on general robotic manipulation frameworks with VLM-based planners and flexible intermediate representations.

---

5. Task and motion planning for execution in the real

URL: View paper

Brief Assessment

Execution in Real[54] focuses on task and motion planning (TAMP) for robotic manipulation with partial grounding and execution-time behaviors, not on VLM-based planning architectures or intermediate representation learning that RoboInter-VLA proposes.

---

6. Payload-Aware Trajectory Optimisation for Non-Holonomic Mobile Multi-Robot Manipulation With Tip-Over Avoidance

URL: View paper

Brief Assessment

Payload Aware Trajectory[59] focuses on multi-robot mobile manipulation with trajectory optimization for non-holonomic constraints and tip-over avoidance, not on vision-language-action frameworks linking high-level planning to low-level execution through intermediate representations.

---

7. A shared autonomy system for precise and efficient remote underwater manipulation

URL: View paper

Brief Assessment

Underwater Manipulation[58] focuses on shared autonomy for underwater robotic manipulation with VR interfaces and natural language commands, not on general plan-then-execute frameworks for robotic manipulation with VLM-based planners and flexible chain-of-thought representations.

---

8. Automatic Behavior Tree Expansion with LLMs for Robotic Manipulation

URL: View paper

Brief Assessment

Behavior Tree Expansion[56] focuses on dynamically expanding behavior trees using LLMs for robotic manipulation, which is a different architectural approach than RoboInter-VLA's VLM-based planner-executor framework with flexible chain-of-thought representations.

---

9. Reinforced Embodied Planning with Verifiable Reward for Real-World Robotic Manipulation

URL: View paper

Brief Assessment

Reinforced Embodied Planning[57] focuses on reinforcement learning with verifiable rewards for VLM-based planning, not on flexible architectural variants linking planning to execution through intermediate representations.

---

10. Robust planning for multi-stage forceful manipulation

URL: View paper

Brief Assessment

Forceful Manipulation Planning[55] focuses on task and motion planning for forceful manipulation tasks (opening bottles, twisting nuts, cutting vegetables) with torque and friction constraints, not on vision-language-action models or intermediate representations for general robotic manipulation.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] RoboInter: A Holistic Intermediate Representation Suite Towards Robotic Manipulation View paper
- [1] Language-driven representation learning for robotics View paper
- [2] UpViTaL: Unpaired Visual-Tactile Self-Supervised Representation Learning for Dexterous Robotic Manipulation View paper
- [3] PEEK: Guiding and Minimal Image Representations for Zero-Shot Generalization of Robot Manipulation Policies View paper
- [4] Rethinking latent representations in behavior cloning: An information bottleneck approach for robot manipulation View paper
- [5] A step toward world models: A survey on robotic manipulation View paper
- [6] EmbodiedMAE: A Unified 3D Multi-Modal Representation for Robot Manipulation View paper
- [7] FlowDreamer: A RGB-D World Model with Flow-based Motion Representations for Robot Manipulation View paper
- [8] Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints View paper
- [9] Is an object-centric representation beneficial for robotic manipulation ? View paper
- [10] Human-oriented Representation Learning for Robotic Manipulation View paper

- [11] MSGField: A Unified Scene Representation Integrating Motion, Semantics, and Geometry for Robotic Manipulation View paper
- [12] DM1: MeanFlow with Dispersive Regularization for 1-Step Robotic Manipulation View paper
- [13] Rethinking Intermediate Representation for VLM-based Robot Manipulation View paper
- [14] Learning for Sequential Manipulation View paper
- [15] Masked Visual-Tactile Pre-training for Robot Manipulation View paper
- [16] Bridging Perception and Action: Spatially-Grounded Mid-Level Representations for Robot Generalization View paper
- [17] Self-supervised sim-to-real adaptation for visual robotic manipulation View paper
- [18] Multi-View Masked World Models for Visual Robotic Manipulation View paper
- [19] XR-1: Towards Versatile Vision-Language-Action Models via Learning Unified Vision-Motion Representations View paper
- [20] LaVA-Man: Learning Visual Action Representations for Robot Manipulation View paper
- [21] Generative artificial intelligence in robotic manipulation: A survey View paper
- [22] HAMSTER: Hierarchical Action Models For Open-World Robot Manipulation View paper
- [23] Viola: Imitation learning for vision-based manipulation with object proposal priors View paper
- [24] Learning Sim-to-Real Dense Object Descriptors for Robotic Manipulation View paper
- [25] Learning to Recover from Plan Execution Errors during Robot Manipulation: A Neuro-symbolic Approach View paper
- [26] Moto: Latent Motion Token as the Bridging Language for Learning Robot Manipulation from Videos View paper
- [27] DreamerRL: empowering representation learning via predictive world models for robot manipulation tasks View paper
- [28] Learning manipulation by predicting interaction View paper
- [29] The Treachery of Images: Bayesian Scene Keypoints for Deep Policy Learning in Robotic Manipulation View paper
- [30] Splat-MOVER: Multi-Stage, Open-Vocabulary Robotic Manipulation via Editable Gaussian Splatting View paper
- [31] Learning graph-structured representations for robotic manipulation View paper
- [32] Self-Supervised Learning of Multi-Object Keypoints for Robotic Manipulation View paper
- [33] Manipulate by seeing: Creating manipulation controllers from pre-trained representations View paper
- [34] RoboGround: Robotic Manipulation with Grounded Vision-Language Priors View paper
- [35] Manipulator-independent representations for visual imitation View paper
- [36] MOSAIC: Learning Unified Multi-Sensory Object Property Representations for Robot Learning via Interactive Perception View paper
- [37] Zero-Shot Robotic Manipulation with Pretrained Image-Editing Diffusion Models View paper
- [38] GNN Topology Representation Learning for Deformable Multi-Linear Objects Dual-Arm Robotic Manipulation View paper
- [39] KALIE: Fine-Tuning Vision-Language Models for Open-World Manipulation Without Robot Data View paper
- [40] MURM: Utilization of Multi-Views for Goal-Conditioned Reinforcement Learning in Robotic Manipulation View paper
- [41] Learning 3D Dynamic Scene Representations for Robot Manipulation View paper
- [42] Vision in Action: Learning Active Perception from Human Demonstrations View paper
- [43] Disentangled Object-Centric Image Representation for Robotic Manipulation View paper
- [44] Robots Pre-train Robots: Manipulation-Centric Robotic Representation from Large-Scale Robot Datasets View paper
- [45] Action-conditional implicit visual dynamics for deformable object manipulation View paper
- [46] Contact-aware and multi-modal robotic manipulation View paper
- [47] Rethinking the Intermediate Features in Adversarial Attacks: Misleading Robotic Models via Adversarial Distillation View paper
- [48] Ag2Manip: Learning Novel Manipulation Skills with Agent-Agnostic Visual and Action Representations View paper
- [49] RT-Affordance: Affordances are Versatile Intermediate Representations for Robot Manipulation View paper
- [50] Dynamics Learning With Object-Centric Interaction Networks for Robot Manipulation View paper
- [51] Human-Aware Reactive Task Planning of Sequential Robotic Manipulation Tasks View paper
- [52] Recent trends in task and motion planning for robotics: A survey View paper
- [53] Practical task and motion planning for robotic food preparation View paper
- [54] Task and motion planning for execution in the real View paper
- [55] Robust planning for multi-stage forceful manipulation View paper
- [56] Automatic Behavior Tree Expansion with LLMs for Robotic Manipulation View paper
- [57] Reinforced Embodied Planning with Verifiable Reward for Real-World Robotic Manipulation View paper
- [58] A shared autonomy system for precise and efficient remote underwater manipulation View paper
- [59] Payload-Aware Trajectory Optimisation for Non-Holonomic Mobile Multi-Robot Manipulation With Tip-Over Avoidance View paper
- [60] Robovqa: Multimodal long-horizon reasoning for robotics View paper
- [61] Iqa: Visual question answering in interactive environments View paper
- [62] Robo2VLM: Visual Question Answering from Large-Scale In-the-Wild Robot Manipulation Datasets View paper
- [63] Cosmos-reason1: From physical common sense to embodied reasoning View paper
- [64] ReMEmbR: Building and Reasoning Over Long-Horizon Spatio-Temporal Memory for Robot Navigation View paper
- [65] Vlaser: Vision-Language-Action Model with Synergistic Embodied Reasoning View paper
- [66] RobotvqaâaÂ scene-graph-and deep-learning-based visual question answering system for robot manipulation View paper
- [67] 3DLLM-Mem: Long-Term Spatial-Temporal Memory for Embodied 3D Large Language Model View paper
- [68] Embodied Intelligence for 3D Understanding: A Survey on 3D Scene Question Answering View paper
- [69] SpatialPIN: Enhancing Spatial Reasoning Capabilities of Vision-Language Models through Prompting and Interacting 3D Priors View paper
- [70] RoboBrain: A Unified Brain Model for Robotic Manipulation from Abstract to Concrete View paper
- [71] Scalable Vision-Language-Action Model Pretraining for Robotic Manipulation with Real-Life Human Activity Videos View paper
- [72] Human-in-the-loop Online Rejection Sampling for Robotic Manipulation View paper
- [73] DexCanvas: Bridging Human Demonstrations and Robot Learning for Dexterous Manipulation View paper
- [74] StreamingCoT: A Dataset for Temporal Dynamics and Multimodal Chain-of-Thought Reasoning in Streaming VideoQA View paper
- [75] EPFL-Smart-Kitchen-30: Densely annotated cooking dataset with 3D kinematics to challenge video and language models View paper