TRIRELVLA: TRIADIC RELATIONAL STRUCTURE FOR GENERALIZABLE EMBODIED MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action (VLA) models perform well on training-seen robotic tasks but struggle to generalize to unseen scenes and objects. A key limitation is their implicit visual representations, which entangle object appearance, background, and scene layout, making policies sensitive to visual variations. Prior work improves transferability through structured intermediate representations, but these mainly capture scene semantics rather than action-relevant relations. We argue that manipulation actions are fundamentally governed by the object–hand–task relational structure, which defines interactions among task requirements, robot states, and object properties. Based on this observation, we propose **TriRelVLA**, a triadic relational VLA framework for generalizable embodied manipulation. Our approach consists of three components: 1) We construct explicit **object–hand–task triadic representations** from multimodal inputs as relational primitives. 2) We build a **task-grounded relational graph**. Task-guided cross-attention forms nodes, and a relation-aware graph transformer models interactions among them. 3) We perform **relation-conditioned action generation**. The relational structure is compressed into a bottleneck space and projected into the LLM for action prediction. This triadic relational bottleneck reduces reliance on appearance statistics and enables transfer across scenes, objects, and task compositions. We further introduce a real-world robotic dataset for fine-tuning. Experiments show strong performance on fine-tuned tasks and clear gains in cross-scene, cross-object, and cross-task generalization. Our code will be released upon acceptance.

1 INTRODUCTION

Vision-language-action (VLA) models have achieved strong performance on embodied tasks such as manipulation (Brohan et al., 2022; Kim et al., 2024) and navigation (Liu et al., 2024b). As shown in Fig. 1(a), existing VLAs (Kim et al., 2024; Black et al., 2024; Chi et al., 2025; Intelligence et al., 2025) encode visual observations and language instructions with LLMs to predict actions. They learn strong cross-modal action priors from large-scale demonstrations. However, their implicit visual representations entangle object appearance, background, and scene layout, making policies sensitive to visual variations and limiting generalization to unseen scenes and objects (Fig. 1(d)).

To address this issue, prior methods (Liu et al., 2024a; Huang et al., 2026; Li et al., 2025b; 2024c) improve transferability between perception and control by constructing intermediate representations that objectify visual content as summarized in Fig. 1(b). However, these representations mainly capture scene semantics instead of action-relevant relations. This leaves action prediction partially tied to visual appearance statistics. In contrast, robotic manipulation requires reasoning over task requirements, robot states, object attributes, and their interactions. Explicit modeling of the relations among these cues is thus critical for cross-task generalization in VLAs.

Motivated by this gap, we introduce the object–hand–task relational structure for action decisions in manipulation as illustrated in Fig. 1(c). Specifically, we observe that action-relevant visual information centers on the target object and the robot hand, and language specifies task goals and constraints. We therefore propose a triadic representation of object, hand, and task as a bridge between visual perception and action decision, and model their interactions to drive action generation and filter redundant scene factors. This representation reduces dependence on appearance statistics and supports more compact and transferable action prediction.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

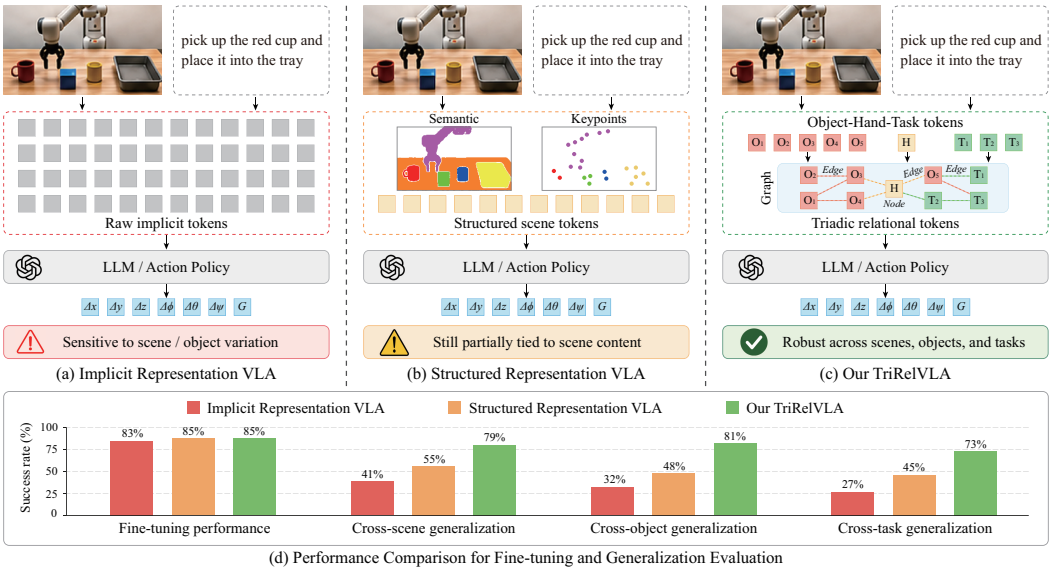

Figure 1: Illustration of VLA representation paradigms for robotic manipulation. (a) Implicit VLAs use dense latent tokens but are sensitive to scene variations. (b) Structured VLAs objectify visual content yet remain partially tied to scene appearance. (c) TriRelVLA models the object–hand–task relational structure for compact and transferable action prediction. (d) Results show competitive fine-tuning performance and stronger cross-scene, cross-object, and cross-task generalization.

Building on the need for explicit relational modeling, we introduce **TriRelVLA**, a VLA framework with a triadic relational structure for generalizable embodied manipulation. As illustrated in Fig. 2, the framework contains three components. First, we construct an object–hand–task representation. Multimodal inputs are encoded into visual and language latent spaces. Object and hand representations are derived from visual features, and task representations from language features. These form the basic relational primitives. Second, we build a task-grounded relational graph. Task-guided cross-attention organizes object and hand representations into nodes together with task representations. Pairwise edges define the relational structure, which is updated with a relation-aware graph transformer. Third, we perform relation-conditioned action generation. The updated relational structure is compressed into a bottleneck space and projected into the LLM for action prediction.

Our TriRelVLA is a unified design that directly addresses the limitations of prior representations, which improves data efficiency and compositional generalization across diverse scenes, objects, and task combinations. We also introduce a real-world robotic dataset to fine-tune the model for generalizable manipulation. Our main contributions are summarized as follows:

- We propose a novel triadic relational VLA framework for generalizable embodied manipulation. Our model improves the transferability of VLAs to unseen robotic scenes with a triadic relational structure as an intermediate representation between visual perception and action control.
- We design an explicit triadic representation that reveals the core cues governing action decision, and motivates an object–hand–task structure to organize visual representations and provide a more compact basis for action prediction.
- We construct a task-grounded relational graph by observing a strong interaction between triadic representations and action generation, which motivates a task-guided graph to optimize the relational structure and provide transferable support for action prediction.
- We build a real-world robotic dataset to fine-tune our model, and extensive experiments show that the framework achieves state-of-the-art performance across diverse generalizable robotic tasks.

2 Related Work

Vision-Language-Action Models. Vision-language models (Liu et al., 2023b; Alayrac et al., 2022; Liu et al., 2023c; Li et al., 2023; Zhou & Lee, 2025) have advanced scene reasoning and inspired

VLA models for embodied manipulation. Existing VLAs (Kim et al., 2024; Black et al., 2024; Team et al., 2024; Zhou et al., 2025a; Ma et al., 2026; Wen et al., 2025a; Fan et al., 2025; Zhou et al., 2025b) typically align multi-view visual latents with language embeddings, and feed them into an LLM to learn action generation from large-scale demonstrations. However, their implicit visual representations entangle object appearance, background texture, and scene layout, making policies sensitive to unseen scene and object variations. In this work, we learn discriminative and structured scene representations for generalizable embodied manipulation.

Generalizable Structured Representation. Improving the generalization of VLA models (Zheng et al., 2024; Kim et al., 2025; Wen et al., 2025b; Zhao et al., 2025; Huang et al., 2025) to unseen scenes and objects requires extracting structured cues that remain stable across environmental variations. Existing methods (Liu et al., 2024a; Huang et al., 2026; Li et al., 2025b; 2024c) construct intermediate representations, such as objects, regions, semantic maps, or keypoints, to objectify visual observations and improve transferability between perception and control. However, these representations mainly describe scene semantics rather than the action-relevant relations that govern manipulation decisions. As a result, action prediction remains partially tied to visual appearance statistics, limiting generalization across scenes, objects, and tasks. This motivates us to build an action-centric relational representation that explicitly structures the core cues required for compositional manipulation.

Relational Reasoning for Action Generation. Mining task-relevant cues and modeling their relations are crucial for compositional generalization in visual reasoning and embodied decision-making. Existing relational modeling methods (Vaswani et al., 2017; Xie et al., 2025; Huang et al., 2024b; Wang et al., 2025a) mainly fall into two categories: implicit attention-based reasoning and explicit structure-based reasoning. The former (Ma et al., 2022; Kim et al., 2024; Zhou et al., 2025a) models interactions through Transformer self-attention but lacks explicit structural inductive bias. The latter (Huang et al., 2024b;a; 2026) builds relations over intermediate entities (*e.g.*, objects, keypoints, and scene graphs), yet mainly targets scene understanding rather than action generation. In contrast, TriRelVLA explicitly models object–hand–task relations, yielding a compact action-centric representation for compositional generalization across unseen scenes, objects, and tasks.

3 Our TriRelVLA

Overview. Fig. 2 shows the architecture of our TriRelVLA. Given multi-view images and instruction texts, our TriRelVLA achieves generalizable embodied manipulation through the three stages:

1) **Object–Hand–Task Triadic Representation (Sec. 3.1).** This stage constructs a triadic representation from visual and linguistic inputs. We first encode the image and text into a visual latent $\mathbf{F}_v$ and a linguistic latent $\mathbf{F}_l$. We then derive object, hand, and task tokens with dedicated operators:

$$\mathbf{Z}_o = \mathrm{Ground}(\mathbf{F}_v), \quad \mathbf{Z}_h = \mathrm{Anchor}(\mathbf{F}_v, \mathbf{F}_p), \quad \mathbf{Z}_t = \mathrm{Decomp}(\mathbf{F}_l), \tag{1}$$

where $\mathbf{Z}_o$, $\mathbf{Z}_h$, and $\mathbf{Z}_t$ denote the object, hand, and task tokens that form the triadic structural primitive. Here, $\mathbf{F}_p$ is the robot proprioceptive embedding used to condition hand token extraction.

2) **Task-Grounded Relational Graph (Sec. 3.2).** This stage models the triadic relational structure. We transform the triadic tokens into nodes $\mathcal{V}$ through task-guided cross-attention, and define pairwise edges $\mathcal{E}$ to form a relational structure, which is updated by a graph transformer:

$$\mathcal{V} = \mathrm{CAtt}([\mathbf{Z}_o \parallel \mathbf{Z}_h \parallel \mathbf{Z}_t]), \quad \mathcal{E} = \{\mathbf{e}_{ij} \mid \mathbf{v}_i, \mathbf{v}_j \in \mathcal{V}, i \neq j\}, \quad \hat{\mathcal{V}} = \mathrm{GraphTrans}(\mathcal{V}, \mathcal{E}), \tag{2}$$

where $\hat{\mathcal{V}}$ denotes the relation-enhanced triadic nodes.

3) **Relation-Conditioned Action Generation (Sec. 3.3).** This stage designs conditional action generation. We compress the updated triadic nodes into a bottleneck space: $\mathbf{R} = \mathrm{Bott}(\hat{\mathcal{V}})$, and then project them into the language embedding space: $\mathbf{X}_r = \mathrm{Proj}(\mathbf{R})$, where $\mathbf{X}_r$ denotes the relational tokens aligned with the linguistic tokens $\mathbf{X}_l$. The LLM then learns the mapping from the aligned multimodal tokens to actions: $\hat{\mathbf{a}} = \mathrm{LLM}([\mathbf{X}_l \parallel \mathbf{X}_r])$.

Remarks. The output $\hat{\mathbf{a}}$ denotes the final action parameters determined by the triadic relational structure. Our framework uses triadic relations as an intermediate bridge between perception and control, improving data efficiency and compositional generalization across scenes, objects, and tasks.

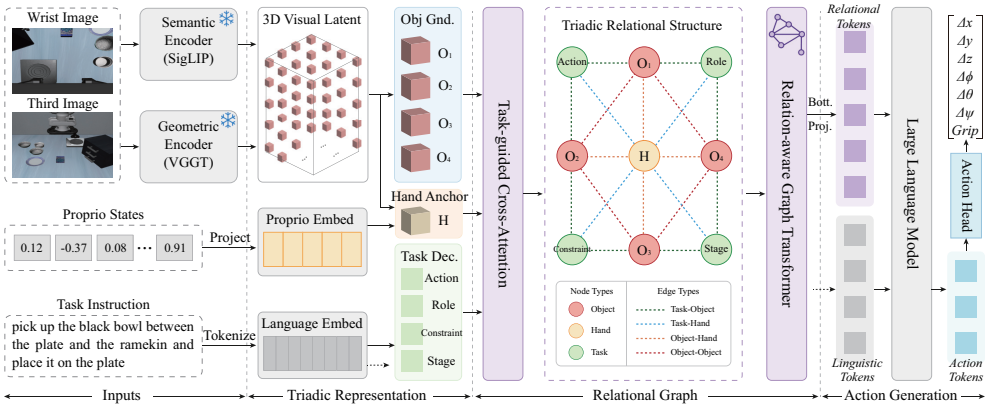

Figure 2: Our **TriRelVLA** consists of three stages: **1) Triadic representation.** Multi-view images, proprioceptive states, and task instructions are encoded to construct object, hand, and task tokens. **2) Relational graph.** The triadic tokens are transformed into task-grounded nodes and updated by a relation-aware graph transformer. **3) Action generation.** The relation-enhanced nodes are compressed into a bottleneck space and fed into the LLM with linguistic tokens for action prediction.

3.1 Object–Hand–Task Triadic Representation

Existing VLA models typically learn an implicit mapping from images and instructions to actions. However, this strategy introduces distractions that hinder generalization due to the conditioning of action generation on high-dimensional visual features entangled with irrelevant scene factors. In this section, we extract action-relevant cues to improve robustness against redundant information.

Multimodal Embedding. In VLA feature space, visual and action representations often lie in different coordinates: images are in 2D pixel space, while actions are executed in 3D world or robot coordinates. To reduce this mismatch, we construct a unified 3D visual latent. Specifically, we use SigLIP (Zhai et al., 2023) to extract multi-view semantic features $\mathbf{F}_{sem}$, and VGGT (Wang et al., 2025b) to obtain 3D geometric features $\mathbf{F}_{geo}$. These features are jointly fused into a unified 3D visual latent $\mathbf{F}_v$ for both wrist-view and third-view observations to facilitate more consistent visual-to-action feature mapping. For language and proprioception, we tokenize the instruction text into the linguistic latent $\mathbf{F}_l$, and optionally project the robot proprioceptive state into a learnable embedding space: $\mathbf{F}_p = \mathrm{Proj}_p(\mathbf{p})$, where $\mathbf{p}$ denotes the raw proprioceptive parameters.

Triadic Representation. The visual latent contains background content and distractors that may obscure action-relevant cues. We posit that action prediction should depend on task requirements, robot states, object properties, and their interactions instead of holistic scene recognition. To identify these cues, we extract visual and language embeddings from a pretrained VLA model such as OpenVLA (Kim et al., 2024) and visualize the attention distributions for action prediction. As shown in Fig. 3, visual attention mainly focuses on target objects and the robot hand, and language attention highlights task goals and action constraints. We are motivated by this observation to use object, hand, and task representations as primitives for action decision. For ***object representation***, we ground candidate object tokens from the 3D visual latent using learnable object queries $\mathbf{Q}_o$. These queries attend to visual features $\mathbf{F}_v$ to aggregate object-centric evidence. Specifically, we compute projected queries, keys, and values as:

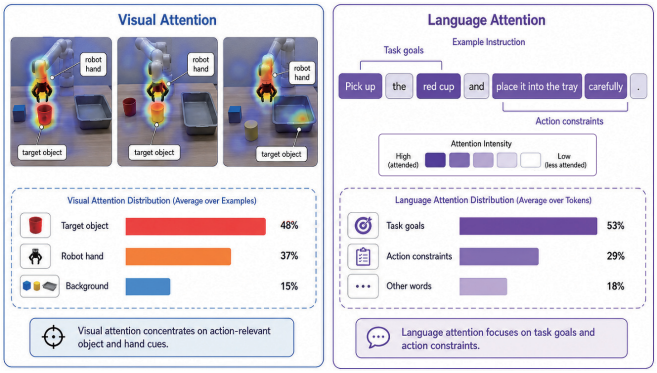

Figure 3: Visualization of action-relevant attention patterns.

$$\tilde{\mathbf{Q}}_o = \mathbf{Q}_o \mathbf{W}_q^o, \quad \tilde{\mathbf{K}}_o = \mathbf{F}_v \mathbf{W}_k^o, \quad \tilde{\mathbf{V}}_o = \mathbf{F}_v \mathbf{W}_v^o, \tag{3}$$

followed by attention: $\mathbf{Z}_o = \mathrm{Softmax}\left(\tilde{\mathbf{Q}}_o \tilde{\mathbf{K}}_o^\top / \sqrt{d}\right) \tilde{\mathbf{V}}_o$, where $\mathbf{W}_q^o$, $\mathbf{W}_k^o$, and $\mathbf{W}_v^o$ are learnable projection matrices, and $\mathbf{Z}_o$ denotes the resulting candidate object tokens. For ***hand representation***,

we introduce a proprioception-anchored hand query to capture hand cues. The robot proprioceptive embedding $\mathbf{F}_p$ serves as the query, which attends to the 3D visual latent $\mathbf{F}_v$ to aggregate hand-centric evidence. Specifically, we compute:

$$\tilde{\mathbf{Q}}_h = \mathbf{F}_p \mathbf{W}_q^h, \quad \tilde{\mathbf{K}}_h = \mathbf{F}_v \mathbf{W}_k^h, \quad \tilde{\mathbf{V}}_h = \mathbf{F}_v \mathbf{W}_v^h, \tag{4}$$

followed by attention $\mathbf{Z}_h = \mathrm{Softmax}\left(\tilde{\mathbf{Q}}_h \tilde{\mathbf{K}}_h^\top / \sqrt{d}\right) \tilde{\mathbf{V}}_h$, where $\mathbf{W}_q^h$, $\mathbf{W}_k^h$, and $\mathbf{W}_v^h$ are learnable projection matrices, and $\mathbf{Z}_h$ denotes the resulting hand token. For ***task representation***, we decompose the linguistic latent into four task-aware tokens: action, role, constraint, and stage. Four learnable category queries attend over $\mathbf{F}_l$ to aggregate distinct task cues. Specifically, we compute $\tilde{\mathbf{F}}_l = \mathbf{F}_l \mathbf{W}_t$, followed by attention $\boldsymbol{\alpha}^m = \mathrm{Softmax}\left(\mathbf{q}_t^m \tilde{\mathbf{F}}_l^\top / \sqrt{d_t}\right)$ and $\mathbf{z}_t^m = \boldsymbol{\alpha}^m \tilde{\mathbf{F}}_l$, where $m \in \{\text{act}, \text{role}, \text{con}, \text{stage}\}$. Here, $\mathbf{q}_t^m$ denotes the learnable query for the m-th task category, and $\mathbf{z}_t^m$ denotes the corresponding task token. The final task token is defined: $\mathbf{Z}_t = [\mathbf{z}_t^{\text{act}} \parallel \mathbf{z}_t^{\text{role}} \parallel \mathbf{z}_t^{\text{con}} \parallel \mathbf{z}_t^{\text{stage}}]$ as structured task primitives.

3.2 Task-Grounded Relational Graph

Despite structured visual representations, existing VLAs generalize poorly across scenes and objects due to underexplored relations between triadic representations and action decisions. We thus explicitly model triadic relations to support generalizable action generation.

Task-Guided Triadic Nodes & Edges. The relational structure among triadic representations is crucial for cross-scene and cross-task transfer. Scene graphs provide an explicit and interpretable way to model how key cues jointly govern action generation. As shown in Fig. 4, action generation depends mainly on target-object affordances, hand states, and task constraints to form an object–hand–task interaction pattern. We therefore refine the scene graph into an object–hand–task relational structure to focus on action-relevant interactions while suppressing redundant scene factors. For ***node construction***, we use task-guided cross-attention to inject task context into object and hand tokens to form graph nodes. Specifically, we first compute the projected queries, keys, and values:

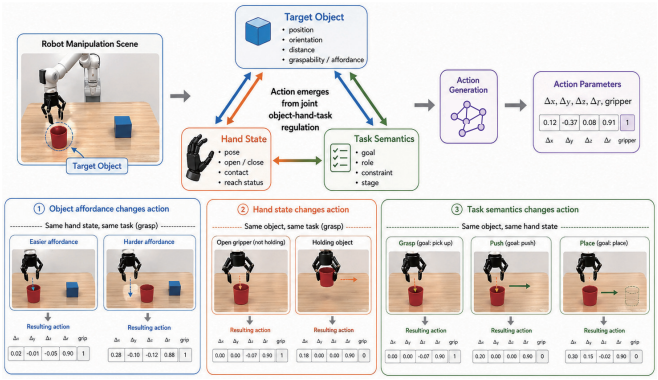

Figure 4: Illustration of triadic interaction for action decision.

$$\mathbf{Q}_o^n = \mathbf{Z}_o \mathbf{W}_{q,o}^n, \quad \mathbf{Q}_h^n = \mathbf{Z}_h \mathbf{W}_{q,h}^n, \quad \mathbf{K}_t^n = \mathbf{Z}_t \mathbf{W}_{k,t}^n, \quad \mathbf{V}_t^n = \mathbf{Z}_t \mathbf{W}_{v,t}^n, \tag{5}$$

followed by task-guided attention:

$$\tilde{\mathbf{Z}}_o = \mathbf{Z}_o + \mathrm{Softmax}\left(\mathbf{Q}_o^n (\mathbf{K}_t^n)^\top / \sqrt{d}\right) \mathbf{V}_t^n, \quad \tilde{\mathbf{Z}}_h = \mathbf{Z}_h + \mathrm{Softmax}\left(\mathbf{Q}_h^n (\mathbf{K}_t^n)^\top / \sqrt{d}\right) \mathbf{V}_t^n, \tag{6}$$

where $\mathbf{W}_{q,o}^n$, $\mathbf{W}_{q,h}^n$, $\mathbf{W}_{k,t}^n$, and $\mathbf{W}_{v,t}^n$ are learnable projection matrices. The resulting node set is defined as $\mathcal{V} = [\tilde{\mathbf{Z}}_o \parallel \tilde{\mathbf{Z}}_h \parallel \mathbf{Z}_t]$. For ***edge construction***, we model four relation types: task–object ($\mathbf{e}_{to}$), task–hand ($\mathbf{e}_{th}$), object–hand ($\mathbf{e}_{oh}$), and object–object ($\mathbf{e}_{oo}$). Each edge is computed via a relation-specific encoder applied to concatenated features. Concretely, we define $\mathbf{e}_{oh} = \phi_{oh}([\tilde{\mathbf{Z}}_h; \tilde{\mathbf{Z}}_o; \Delta\mathbf{p}_{oh}; \Delta\mathbf{R}_{oh}; d_{oh}; s_{oh}])$ and $\mathbf{e}_{oo} = \phi_{oo}([\tilde{\mathbf{Z}}_{o,i}; \tilde{\mathbf{Z}}_{o,j}; \Delta\mathbf{p}_{i,j}; \Delta\mathbf{R}_{i,j}; d_{i,j}])$ to capture geometric and interaction cues, and $\mathbf{e}_{to} = \phi_{to}([\mathbf{Z}_t; \tilde{\mathbf{Z}}_o])$ and $\mathbf{e}_{th} = \phi_{th}([\mathbf{Z}_t; \tilde{\mathbf{Z}}_h])$ to encode task-conditioned dependencies. Here, ϕ_{oh}, ϕ_{oo}, ϕ_{to}, and ϕ_{th} denote relation-specific edge encoders, and $\Delta\mathbf{p}$, $\Delta\mathbf{R}$, d, and s represent relative position, relative pose, distance, and contact or reachability score. The resulting edge set is defined as: $\mathcal{E} = \{\mathbf{e}_{to}, \mathbf{e}_{th}, \mathbf{e}_{oh}, \mathbf{e}_{oo}\}$.

Relation-Aware Graph Transformer. To learn the relational structure among triadic nodes and edges, we design a relation-aware graph transformer that injects edge features directly into attention for interactions to be modeled and propagated across nodes. For each node i, we first compute projected queries, keys, values, and edge embeddings:

$$\mathbf{q}_i = \mathbf{W}_q^g \mathbf{v}_i, \quad \mathbf{k}_j = \mathbf{W}_k^g \mathbf{v}_j, \quad \mathbf{v}_j^g = \mathbf{W}_v^g \mathbf{v}_j, \quad \mathbf{r}_{ij} = \mathbf{W}_r^g \mathbf{e}_{ij}, \tag{7}$$

followed by relation-aware attention and aggregation:

$$\alpha_{ij} = \mathrm{Softmax}_j\left(\mathbf{q}_i(\mathbf{k}_j + \mathbf{r}_{ij})^\top/\sqrt{d}\right), \quad \hat{\mathbf{v}}_i = \mathbf{v}_i + \sum\nolimits_{j\in\mathcal{N}(i)} \alpha_{ij}(\mathbf{v}_j^g + \mathbf{r}_{ij}). \tag{8}$$

$\mathbf{v}_i$ denotes the feature of node i, $\mathbf{e}_{ij}$ the edge feature, and $\hat{\mathbf{v}}_i$ the relation-enhanced node feature. By incorporating $\mathbf{r}_{ij}$ into both attention weights and value updates, the model captures interactions that depend on node similarity as well as geometric and semantic relations. The final node set $\hat{\mathcal{V}} = \{\hat{\mathbf{v}}_i\}_{i=1}^{|\mathcal{V}|}$ provides structured and transferable representations for action prediction.

3.3 Relation-Conditioned Action Generation

Although triadic relational representations capture object–hand–task interactions, directly feeding them into the LLM may cause weak cross-modal alignment. We therefore compress and align them into relational conditioning embeddings to guide task- and scene-aware action generation.

Relational Bottleneck. Although the relational graph captures object–hand–task dependencies, it may still include redundant components such as task-irrelevant object nodes. We introduce a relational bottleneck to compress relation-enhanced nodes into compact action-relevant tokens to improve alignment efficiency and generalization. Specifically, we compute an importance weight $\alpha_{k,i}$ for each node $\hat{\mathbf{v}}_i$ and aggregate them into K relational bottleneck tokens:

$$\alpha_{k,i} = \mathrm{Softmax}_i\left(\mathbf{w}_k^\top \hat{\mathbf{v}}_i\right), \quad \mathbf{r}_k = \sum\nolimits_{i=1}^{|\hat{\mathcal{V}}|} \alpha_{k,i}\hat{\mathbf{v}}_i, \quad \mathbf{R} = [\mathbf{r}_1 \parallel \mathbf{r}_2 \parallel \cdots \parallel \mathbf{r}_K], \tag{9}$$

where $\mathbf{R}$ denotes the relational bottleneck representation, which helps further filter out irrelevant object interference and improves the data efficiency of the model.

Conditional Action Generation. After obtaining the relational bottleneck, we project it into the language embedding space with an MLP: $\mathbf{X}_r = \mathrm{MLP}(\mathbf{R})$, where $\mathbf{X}_r$ denotes relational tokens. We then concatenate $\mathbf{X}_r$ with linguistic tokens $\mathbf{X}_l$ and feed them into the LLM for action prediction: $\hat{\mathbf{a}} = \mathcal{H}(\mathrm{LLM}([\mathbf{X}_l \parallel \mathbf{X}_r]))$, where $\mathcal{H}$ is the action head. Generalization and stability are improved by the reliance of action prediction on the compact triadic relations instead of raw visual appearance.

3.4 Optimization

To stabilize triadic representation learning and prevent token drift, we add object grounding and hand anchoring constraints besides action supervision. First, we employ L1 loss to supervise the prediction of continuous action parameters: $\mathcal{L}_{act} = ||\hat{a} - a||_1$, where $\hat{a}$ and a denote the predicted and ground-truth actions, respectively. Since object and hand representations are latent tokens, we supervise their attention maps instead of token values by aligning them with the corresponding region masks: $\mathcal{L}_{obj} = \mathrm{BCE}(\mathbf{A}_o, \mathbf{M}_o)$, $\mathcal{L}_{hand} = \mathrm{BCE}(\mathbf{A}_h, \mathbf{M}_h)$, where $\mathbf{A}_o$, $\mathbf{A}_h$ denote the attention maps for the object and hand, respectively. $\mathbf{M}_o$ is the object region mask and $\mathbf{M}_h$ is the hand region mask. Note that the losses $\mathcal{L}_{obj}$ and $\mathcal{L}_{hand}$ are also applicable to the attention constraints of the object and hand nodes in the relational graph. The overall optimization objective is formulated as: $\mathcal{L}_{total} = \mathcal{L}_{act} + \lambda_o\mathcal{L}_{obj} + \lambda_h\mathcal{L}_{hand}$, where λ_o, λ_h are the weighting coefficients for the auxiliary constraints. The auxiliary losses stabilize object-hand alignment, and action loss drives relational representations toward final action decisions.

4 Dataset and Training Pipeline

4.1 Robotic Dataset

To improve the basic action prediction ability, real-world transferability, and triadic relational modeling capability of our model, we adopt multiple complementary datasets to train our model.

OXE. Open X-Embodiment (O'Neill et al., 2024) is a large-scale robotic manipulation dataset for pre-training, covering diverse robots, environments, tasks, and basic skills such as grasping, placing, moving, opening/closing, pushing, and pulling.

DROID. DROID (Khazatsky et al., 2024) is a large-scale real-world manipulation dataset collected in diverse household and tabletop scenes, emphasizing visual variation and manipulation complexity for improving real-scene transferability. We further semi-automatically annotate DROID with hand and object masks.

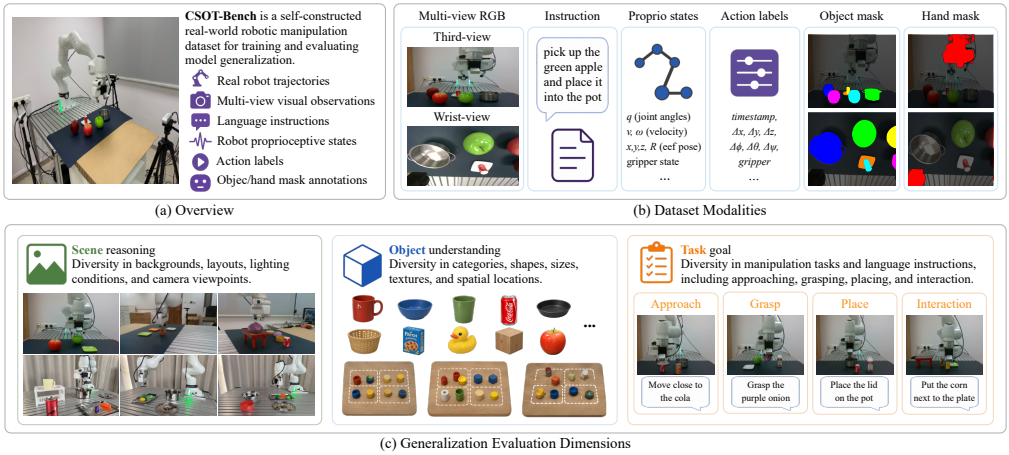

Figure 5: Illustration of CSOT-Bench. (a) Real-world robotic manipulation dataset. (b) Multi-modal annotations, including multi-view RGB images, instructions, proprioception, actions, and object/hand masks. (c) Generalization evaluation dimensions across scenes, objects, and tasks.

LIBERO. LIBERO (Liu et al., 2023a) is a public simulation benchmark covering instruction-driven manipulation tasks in spatial reasoning, object understanding, goal completion, and long-horizon planning. We use it for generalization evaluation with object/hand masks for auxiliary supervision.

CSOT-Bench. CSOT-Bench is our self-constructed real-world robotic manipulation dataset with three generalization evaluation suites, including scene reasoning, object understanding and task goal. As shown in Fig. 5, it includes robot trajectories, multi-view observations, instructions, proprioceptive states, action labels, and object/hand masks. It varies scene factors such as backgrounds, and viewpoints; object factors such as categories, shapes, and locations; and task factors such as approaching, grasping, placing, and goal-conditioned interactions. These variations allow CSOT-Bench to evaluate whether models learn structured object–hand–task relations rather than visual appearance cues.

4.2 Training Pipeline

Our TriRelVLA model first loads the weights of several pretrained models to initialize its visual reasoning capability. To fully enhance the learning capability of the triadic relational structure for generalizable embodied tasks, we divide the training process into three stages:

Stage 1: Multimodal Alignment. This stage learns basic vision-action alignment using the large-scale OXE dataset with only the action loss $\mathcal{L}_{act}$. It provides a general action prior for subsequent relational representation learning. At this stage, we only update the projector and the action head, while freezing the remaining modules.

Stage 2: Relational Structure Modeling. This stage learns triadic representations and relational structures on the realistic DROID dataset with the full loss $\mathcal{L}_{total}$. It shifts the model from vision-action mapping to relation-driven action generation, improving generalization under complex scenes and object variations. At this stage, we update the triadic representation, relational graph, bottleneck space, and projector, while freezing the semantic encoder, geometric encoder, action head, and LLM.

Stage 3: Robotic Task Fine-tuning. This stage fine-tunes the model on LIBERO and CSOT-Bench with the full loss $\mathcal{L}_{total}$ for task adaptation and compositional generalization. It adapts the model to specific tasks, objects, and scene variations while preserving prior vision-action alignment and relational modeling abilities. At this stage, we update the triadic representation, relational graph, bottleneck space, projector, and action head, and optimize the LLM components using LoRA, while keeping the semantic encoder and geometric encoder frozen.

5 Experiments

Implementation Details. TriRelVLA uses Qwen3-4B (Yang et al., 2025) as the LLM backbone, SigLIP (Zhai et al., 2023) as the semantic encoder, and DINOv2 from VGGT (Wang et al., 2025b) as the geometry encoder. The model is trained on 8×A800 GPUs with AdamW. The learning rates and batch sizes are set to 1×10^{-4}/64, 2×10^{-5}/32, and 1×10^{-5}/32 for stages 1, 2, and 3, respectively.

Table 1: Quantitative results of VLAs for fine-tuned robotic manipulation tasks.

Representation Category	Method	LIBERO (Liu et al., 2023a)					CSOT-Bench (Ours)			
		Spatial	Object	Goal	Long	Average	Scene	Object	Task	Average
Implicit	OpenVLA (Kim et al., 2024)	84.7	88.4	79.2	53.7	76.5	74.6	82.4	72.0	76.3
	Octo (Team et al., 2024)	78.9	85.7	84.6	51.1	75.1	66.3	78.0	71.5	72.0
	CogACT (Li et al., 2024b)	87.5	90.2	78.4	53.2	76.5	76.1	84.5	73.8	78.1
	DiffusionPolicy (Chi et al., 2025)	78.3	92.5	68.3	50.5	72.4	68.7	80.2	65.3	71.4
	SpatialVLA (Qu et al., 2025)	88.2	89.9	78.6	55.5	78.1	79.5	87.6	75.4	80.8
Structured	CoA-VLA (Li et al., 2024a)	85.3	93.1	85.8	55.0	79.8	–	–	–	–
	CogVLA (Li et al., 2025a)	**98.6**	98.8	96.6	**95.4**	97.4	81.2	88.1	78.4	82.5
	SemanticVLA (Li et al., 2026)	**98.6**	**99.6**	97.6	94.8	**97.7**	82.7	91.3	78.5	84.2
	TriRelVLA (Ours)	98.2	99.0	**97.8**	94.8	97.6	**83.1**	**91.5**	**80.3**	**84.9**

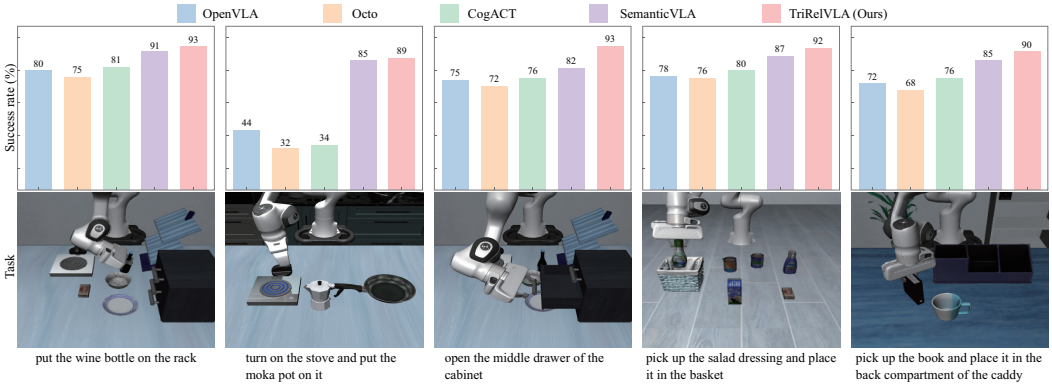

Figure 6: Quantitative comparison of VLAs for generalization evaluation on LIBERO.

Comparison Methods. We compare TriRelVLA with typical VLAs, including OpenVLA (Kim et al., 2024), Octo (Team et al., 2024), CogACT (Li et al., 2024b), DiffusionPolicy (Chi et al., 2025), SpatialVLA (Qu et al., 2025), CoA-VLA (Li et al., 2024a), CogVLA (Li et al., 2025a), and SemanticVLA (Li et al., 2026). The first five methods directly learn implicit vision-to-action mappings, while the latter three introduce intermediate representations to structure visual features.

Evaluation Metric. We evaluate all methods on LIBERO and CSOT-Bench under two settings: fine-tuning on all subtasks and zero-shot generalization across scenes, objects, and tasks. Task success rate is used as the primary metric. Ablation and discussion experiments in Sec. 5.2 are conducted on CSOT-Bench, with results reported for both fine-tuning and zero-shot generalization.

5.1 Comparison with State-of-the-Art Models

Comparison on Fine-Tuned Tasks. Table 1 compares fine-tuning performance on LIBERO and CSOT-Bench. TriRelVLA achieves performance comparable to existing VLAs, which indicates that the triadic relational structure does not compromise basic manipulation accuracy. Although dense visual tokens are replaced by a compact relational bottleneck, the model still preserves sufficient task-relevant information for accurate action prediction.

Comparison on Generalizable Robotic Tasks. Fig. 6 and Fig. 7 compare generalization across scenes, objects, and tasks on LIBERO and CSOT-Bench. Structured VLAs generally outperform implicit VLAs, while TriRelVLA achieves the best results, especially in cross-task generalization. The gains are modest in simulation but more pronounced in real-world settings, showing that object-hand-task relations reduce reliance on visual appearance and enable transferable action generation.

5.2 Ablation Study and Discussion

Effectiveness of Triadic Relational Architecture. Table 2 verifies the effectiveness of the basic architecture, triadic representations, and relational graph. The basic VLA maintains standard performance but shows limited generalization. Adding triadic representations improves generalization, while the relational graph brings further gains. This shows that object–hand–task modeling shifts the model from vision-to-action mapping to relation-driven action generation.

Table 2: Effect of triadic relational architecture.

Basic arch.	Triadic represent	Relational graph	Fine-tuned task	Generalized task
√	×	×	82.6	32.1
√	√	×	84.2	43.0
√	√	√	**84.9**	**79.5**

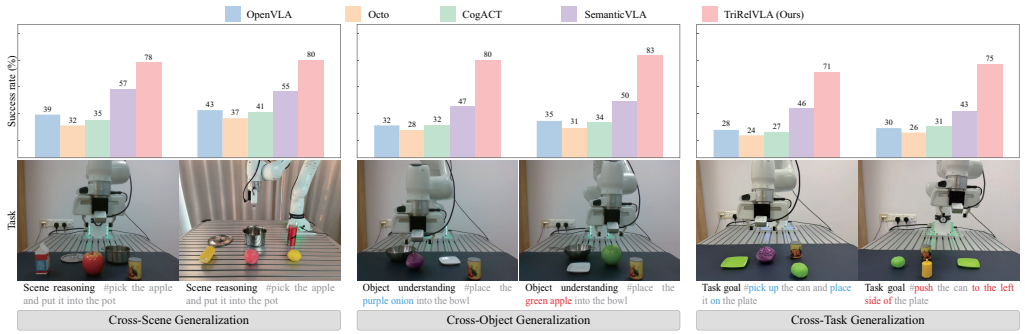

Figure 7: Generalization evaluation of VLAs on CSOT-Bench across scenes, objects, and tasks.

Impact of Structured Triadic Representations. Table 3 shows the impact of different triadic representation components. Using a single representation leads to limited generalization, which indicates that no single cue alone can support cross-scene, cross-object, and cross-task transfer. Jointly using object, hand, and task representations achieves the best results, especially on generalization tasks. This shows that structured triadic representations provide more complete and transferable cues for action prediction.

Table 3: Impact of structured representations.

Representation strategy	Fine-tuned task	Generalized task
w/ Only object	83.7	70.2
w/ Only hand	83.6	67.8
w/ Only task	83.2	68.4
w/ Object–hand–task	**84.9**	**79.5**

Effectiveness of Losses. Table 4 shows that adding object and hand auxiliary constraints to the action loss improves both fine-tuning and generalization by stabilizing object–hand alignment and enhancing relational modeling and action prediction.

Table 4: Effectiveness of loss combination.

Loss	Fine-tuned	Generalized
w/ $\mathcal{L}_{act}$	84.0	74.2
w/ $\mathcal{L}_{act} + \mathcal{L}_{obj} + \mathcal{L}_{hand}$	**84.9**	**79.5**

Effect of Proprioception on Hand Representation. Table 5 shows that proprio-anchored hand tokens outperform visual-only tokens, especially in generalization, by improving end-effector localization under occlusion and viewpoint changes and strengthening state–visual alignment.

Table 5: Effect of proprioception on hand.

Strategy	Fine-tuned	Generalized
w/o Proprio	83.2	71.6
w/ Proprio	**84.9**	**79.5**

Impact of Task Representations on Nodes. Table 6 analyzes the effect of task representations on object and hand node construction. Without task representations, the model performs well on fine-tuned tasks but generalizes poorly as nodes lack task association. Task representations improve both, especially generalization, by guiding nodes into task-driven representations for cross-object and cross-task transfer.

Table 6: Task representation on nodes.

Object	Hand	Fine-tuned	Generalized
w/o Task-Gnd	w/o Task-Gnd	84.1	77.2
w/ Task-Gnd	w/o Task-Gnd	84.3	77.8
w/ Task-Gnd	w/ Task-Gnd	**84.9**	**79.5**

Role of Relational Bottleneck Representation. Table 7 shows relational bottleneck reduces cost without sacrificing performance by compressing relation-enhanced nodes, preserving action-relevant cues, and filtering redundancy.

Table 7: Discussion on bottleneck space.

Setting	Succ↑	GFLOPs↓	Mem (GB)↓
w/o Bottleneck	79.3	742.5	24.8
w/ Bottleneck	**79.5**	**618.3**	**19.6**

Limitation. Although TriRelVLA performs well in cross-scene, cross-object, and cross-task manipulation, it still struggles with long-horizon and unseen tasks. This is due to the lack of memory-based reasoning over triadic spatiotemporal relations and multi-stage subgoals. Our future work will introduce long-term memory (Lei et al., 2025), task decomposition (Ahn et al., 2022), and spatiotemporal relational graph modeling for complex compositional manipulation.

6 Conclusion

In this work, we propose **TriRelVLA**, a VLA model with a triadic relational structure for generalizable embodied manipulation. We extract object, hand, and task tokens from multimodal inputs as structured cues bridging perception and action. A task-guided relational graph models object–hand–task interactions to enable action decisions under structured constraints instead of entangled visual appearance. With a compact relational bottleneck, TriRelVLA aligns triadic relations with linguistic representations in the LLM for robust action prediction. We further construct datasets with object/hand mask annotations and introduce auxiliary supervision for relation learning. Experiments demonstrate the effectiveness of TriRelVLA, particularly in generalization to unseen scenes, objects, and tasks.

AI USE STATEMENT

No generative AI tools were used in the research, implementation, experimentation, analysis, or preparation of this manuscript. The authors take full responsibility for all content presented in this work.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Arthur Mensch, Katie Millican, David Moore, Michael Needham, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23732, 2022.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Yiguo Fan, Pengxiang Ding, Shuanghao Bai, Xinyang Tong, Yuyang Zhu, Hongchao Lu, Fengqi Dai, Wei Zhao, Yang Liu, Siteng Huang, et al. Long-vla: Unleashing long-horizon capability of vision language action model for robot manipulation. *arXiv preprint arXiv:2508.19958*, 2025.

Helong Huang, Min Cen, Kai Tan, Xingyue Quan, Guowei Huang, and Hong Zhang. Graphcot-vla: A 3d spatial-aware reasoning vision-language-action model for robotic manipulation with ambiguous instructions. *arXiv preprint arXiv:2508.07650*, 2025.

Helong Huang, Min Cen, Kai Tan, Xingyue Quan, Guowei Huang, and Hong Zhang. Graphcot-vla: A 3d spatial-aware reasoning vision-language-action model for robotic manipulation with ambiguous instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18324–18332, 2026.

Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024a.

Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, et al. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 2417–2425, 2024b.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

Mingcong Lei, Honghao Cai, Zezhou Cui, Liangchen Tan, Junkun Hong, Gehan Hu, Shuangyu Zhu, Yimou Wu, Shaohan Jiang, Ge Wang, et al. Robomemory: A brain-inspired multi-memory agentic framework for lifelong learning in physical embodied systems. In *NeurIPS 2025 Workshop on Space in Vision, Language, and Embodied AI*, 2025.

Jinming Li, Yichen Zhu, Zhibin Tang, Junjie Wen, Minjie Zhu, Xiaoyu Liu, Chengmeng Li, Ran Cheng, Yaxin Peng, Yan Peng, et al. Coa-vla: Improving vision-language-action models via visual-textual chain-of-affordance. *arXiv preprint arXiv:2412.20451*, 2024a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023.

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024b.

Wei Li, Renshan Zhang, Rui Shao, Jie He, and Liqiang Nie. Cogvla: Cognition-aligned vision-language-action model via instruction-driven routing & sparsification. *arXiv preprint arXiv:2508.21046*, 2025a.

Wei Li, Renshan Zhang, Rui Shao, Zhijian Fang, Kaiwen Zhou, Zhuotao Tian, and Liqiang Nie. Semanticvla: Semantic-aligned sparsification and enhancement for efficient robotic manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18397–18405, 2026.

Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18061–18070, 2024c.

Xiaoqi Li, Jingyun Xu, Mingxu Zhang, Jiaming Liu, Yan Shen, Iaroslav Ponomarenko, Jiahui Xu, Liang Heng, Siyuan Huang, Shanghang Zhang, et al. Object-centric prompt-driven vision-language-action model for robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27638–27648, 2025b.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023a.

Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-world robotic manipulation through mark-based visual prompting. *arXiv preprint arXiv:2403.03174*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Llava: Large language and vision assistant. *arXiv preprint arXiv:2304.08485*, 2023b.

Haotian Liu, Pengchuan Zhang, Jianwei Yang, and Lei Zhang. Improved visual instruction tuning for image-text alignment. *arXiv preprint*, abs/2310.00067, 2023c. URL https://arxiv.org/abs/2310.00067.

Rui Liu, Wenguan Wang, and Yi Yang. Volumetric environment representation for vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16317–16328, 2024b.

Chuanhao Ma, Hanyu Zhou, Shihan Peng, Yan Li, Tao Gu, and Luxin Yan. St-π: Structured spatiotemporal vla for robotic manipulation. *arXiv preprint arXiv:2604.17880*, 2026.

Xiaojian Ma, Weili Nie, Zhiding Yu, Huaizu Jiang, Chaowei Xiao, Yuke Zhu, Song-Chun Zhu, and Anima Anandkumar. Relvit: Concept-guided vision transformer for visual relational reasoning. *arXiv preprint arXiv:2204.11167*, 2022.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 30, 2017.

Haonan Wang, Hanyu Zhou, Haoyue Liu, and Luxin Yan. 4d-vggt: A general foundation model with spatiotemporal awareness for dynamic scene geometry estimation. *arXiv preprint arXiv:2511.18416*, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025b.

Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*, 2025a.

Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*, 2025b.

Chi Xie, Shuang Liang, Jie Li, Zhao Zhang, Feng Zhu, Rui Zhao, and Yichen Wei. Relationlmm: Large multimodal model as open and versatile visual relationship generalist. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, 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.

Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024.

Hanyu Zhou and Gim Hee Lee. Llava-4d: Embedding spatiotemporal prompt into lmms for 4d scene understanding. *arXiv preprint arXiv:2505.12253*, 2025.

Hanyu Zhou, Chuanhao Ma, and Gim Hee Lee. Vla-4d: Embedding 4d awareness into vision-language-action models for spatiotemporally coherent robotic manipulation. *arXiv preprint arXiv:2511.17199*, 2025a.

Zhongyi Zhou, Yichen Zhu, Minjie Zhu, Junjie Wen, Ning Liu, Zhiyuan Xu, Weibin Meng, Yaxin Peng, Chaomin Shen, Feifei Feng, et al. Chatvla: Unified multimodal understanding and robot control with vision-language-action model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5377–5395, 2025b.