SemanticVLA: Semantic-Aligned Sparsification and Enhancement for Efficient Robotic Manipulation

Wei Li[1], Renshan Zhang[1], Rui Shao[1, 2*], Zhijian Fang[1], Kaiwen Zhou[3], Zhuotao Tian[1] and Liqiang Nie[1]

[1]Harbin Institute of Technology, Shenzhen
[2]Shenzhen Loop Area Institute
[3]Huawei Noah's Ark Lab
liwei2024@stu.hit.edu.cn, zhangrenshan@stu.hit.edu.cn, shaorui@hit.edu.cn

Abstract

Vision-Language-Action (VLA) models have advanced in robotic manipulation, yet practical deployment remains hindered by two key limitations: **1) perceptual redundancy**, where irrelevant visual inputs are processed inefficiently, and **2) superficial instruction-vision alignment**, which hampers semantic grounding of actions. In this paper, we propose **SemanticVLA**, a novel VLA framework that performs Semantic-Aligned Sparsification and Enhancement for Efficient Robotic Manipulation. Specifically, **1)** To sparsify redundant perception while preserving semantic alignment, **Semantic-guided Dual Visual Pruner (SD-Pruner)** performs: Instruction-driven Pruner (ID-Pruner) extracts global action cues and local semantic anchors in SigLIP; Spatial-aggregation Pruner (SA-Pruner) compacts geometry-rich features into task-adaptive tokens in DINOv2. **2)** To exploit sparsified features and integrate semantics with spatial geometry, **Semantic-complementary Hierarchical Fuser (SH-Fuser)** fuses dense patches and sparse tokens across SigLIP and DINOv2 for coherent representation. **3)** To enhance the transformation from perception to action, **Semantic-conditioned Action Coupler (SA-Coupler)** replaces the conventional observation-to-DoF approach, yielding more efficient and interpretable behavior modeling for manipulation tasks. Extensive experiments on simulation and real-world tasks show that SemanticVLA sets a new SOTA in both performance and efficiency. SemanticVLA surpasses OpenVLA on LIBERO benchmark by **21.1%** in success rate, while reducing training cost and inference latency by **3.0**× and **2.7**×.

Code — https://github.com/JiuTian-VL/SemanticVLA

1 Introduction

In recent years, advances in deep learning (Shao et al. 2024; Shao, Wu, and Liu 2023; Wang et al. 2025b; Zhou et al. 2025; Xie et al. 2025) have driven rapid development of intelligent agents (Li et al. 2025c; Chen et al. 2025; Lyu et al. 2025; Sun et al. 2025; Zhou et al. 2024; Li et al. 2024b; Zhang et al. 2025c,b). In particular, (VLA) (Kim et al. 2024; Black et al. 2024; Intelligence et al. 2025; Li et al. 2025b; Shao et al. 2025; Zeng et al. 2024) models have advanced robotic manipulation by enabling end-to-end mapping from language to action using pre-trained Vision-Language Models (VLMs) (Liu et al. 2023b; Chen et al. 2024a; Li et al. 2025a; Zhang et al. 2025a, 2024a; Zhu et al. 2025). Despite their progress, existing VLA models still face critical challenges in deployment, particularly in dynamic and cluttered environments. A core bottleneck lies in the lack of semantic-aligned perception and structured action representation, resulting in redundant computation and weak task grounding.

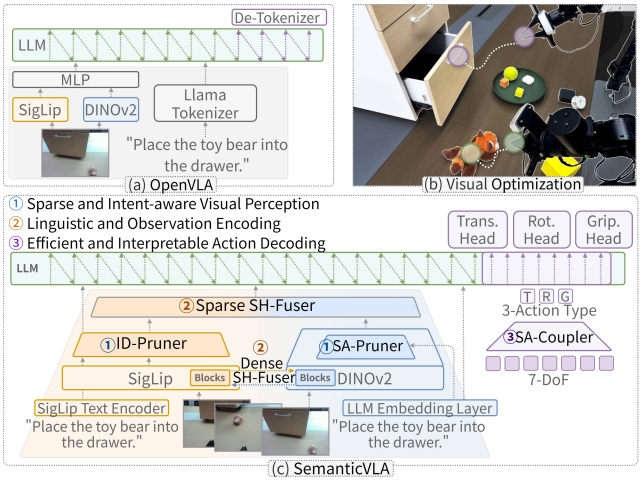

Figure 1: Comparison between OpenVLA and SemanticVLA. SemanticVLA dynamically performs instruction-guided visual sparsification, establishes tight perception-action correspondence, and accelerates parallel decoding via action type coupling, as illustrated in (b) and (c).

This bottleneck is mainly caused by two fundamental limitations: 1) **Redundancy in visual perception**. Prevailing VLA frameworks (Zhao et al. 2025; Bu et al. 2025) adopt generic instruction-agnostic visual encoders (Zhai et al. 2023; Caron et al. 2021; Oquab et al. 2023). They focus on processing all observed pixels uniformly without paying attention to the semantic relevance of the instructions. As a result, background clutter, task-irrelevant distractors, and environmental noise are encoded indiscriminately, leading to excessive computational cost and diluted attention to task-critical cues. 2) **Superficiality in instruction-vision seman-**

*Corresponding authors

tic alignment. Most VLA models (Zhang et al. 2025d; Kim et al. 2024; Black et al. 2024) rely solely on generic cross-modal alignment with large language models. This superficial alignment struggles to capture complex semantic relations in robotics and thus failing to capture fine-grained visual compositionality. Consequently, this significantly limits the VLA's ability to identify global action cues, local semantic anchors, and the structured instruction-spatial dependencies in robotic manipulation tasks.

To address these challenges, we propose **SemanticVLA**, a novel framework designed to perform Semantic-Aligned Sparsification and Enhancement for efficient and interpretable robotic manipulation (Fig. 1). SemanticVLA hinges on three-level complementary semantics: **1) instruction-level linguistic intent semantics** conveyed by task prompts; **2) vision-level spatial semantics** describing objects and their layout; and **3) control-level action semantics** governing translation, rotation, and gripper state.

Specifically, SemanticVLA unifies sparsification and enhancement aligned to these semantics through three integrated modules: **1) Semantic-guided Dual Visual Pruner (SD-Pruner).** Exploiting encoder specialization—SigLIP (Zhai et al. 2023) for instruction grounding and DINOv2 (Oquab et al. 2023) for spatial geometry, SD-Pruner independently prunes each encoder to retain the most task-relevant evidence under occlusion and noise. **i) Instruction-driven Pruner (ID-Pruner) for SigLIP.** We compute instruction–image cross-modality similarity to derive token importance scores, enabling two complementary paths: Vision-to-Language Mapping preserves global action cues from complete instruction inputs, resolving the "know the goal but not the steps" issue. Language-to-Vision Filtering enhances local semantic anchors from complete visual inputs, mitigating the "can't do what you can't see" problem. Together, they form a robust and sparsified perception pipeline that retains essential visual-language-action alignment under occlusion and noise. **ii) Spatial-aggregation Pruner (SA-Pruner) for DINOv2.** Inspired by the register design in previous work (Zhang et al. 2025e), SA-Pruner aggregates DINOv2 features into compact, geometry-rich tokens, further modulated via FiLM (Perez et al. 2018) to reflect instruction relevance, thereby complementing SigLIP's semantics. **2) Semantic-complementary Hierarchical Fuser (SH-Fuser)**. To enhance coherence between spatial vision and task intent, SH-Fuser performs a two-stream fusion that spans the entire visual encoding stage. Dense-Fuser exchanges patch-level information between corresponding blocks of SigLIP and DINOv2, while Sparse-Fuser merges salient tokens produced by ID-Pruner and SA-Pruner. This hierarchical design propagates complementary cues early and late in the stack, producing a unified representation that is both semantically grounded and geometrically faithful. **3) Semantic-conditioned Action Coupler (SA-Coupler).** To enable efficient and interpretable observation-to-action mappings, SA-Coupler adopts a structured formulation that explicitly maps perception representations to semantic action types. Joint control is modulated based on them accordingly, which further enhances both the efficiency and interpretability of action decoding.

We evaluate SemanticVLA on extensive simulation and real-world manipulation tasks, showing improved task success with reduced perceptual redundancy, enhanced instruction reasoning, and lower computational cost over SOTA baselines. Our contributions are summarized as follows:

- We propose **SD-Pruner**, which jointly prunes SigLIP and DINOv2 encoders via instruction-aware token filtering and geometry-aware aggregation via **ID-Pruner** & **SA-Pruner**, significantly pruning redundant perception.
- We propose **SH-Fuser**, a two-stream fusion module that integrates dense patch features and sparse semantic tokens across SigLip and DINOv2, enhancing instruction semantics and spatial structure alignment.
- We design **SA-Coupler** to enable a more efficient mapping from sparsified perception to semantic action types.
- Extensive experiments on standard VLA benchmarks and real-world robot deployments demonstrate that SemanticVLA achieves SOTA performance and efficiency.

2 Related Work

Robot Manipulation with Lightweight Models. These models (Chi et al. 2023; Hao et al. 2025; Lv et al. 2025) typically excels in deterministic and real-time control. However, these approaches heavily rely on pre-defined objects and environments, lacking the semantic generalization capabilities necessary for open-world settings.

Two Branches of VLM-based VLA Models. VLA can be broadly categorized into two types. Monolithic architectures (Kim et al. 2024; Qu et al. 2025), which maintain causal and semantically consistent multi-step reasoning, making them well-suited for open-ended environments. Nevertheless, their reliance on autoregressive action decoding introduces substantial efficiency bottlenecks. Hierarchical expert models (Black et al. 2024; Li et al. 2024a; Intelligence et al. 2025; Shukor et al. 2025) which leverage diffusion or flow-matching mechanisms for high-frequency action prediction. While effective, these models suffer from a disconnect between the VLM and action experts, thereby underutilizing the reasoning capacity of the VLM.

Efficient VLA Modeling Approaches. Efficient VLA models generally fall into three categories: 1) Algorithmic strategies like FAST (Pertsch et al. 2025), which merges action bins using discrete cosine transform (DCT), and PD-VLA (Song et al. 2025) or OpenVLA-OFT (Kim et al. 2025), which employ parallel decoding, aim to accelerate inference. 2) Architectural innovations include RoboMamba's Mamba backbone (Liu et al. 2024), Deer-VLA's multi-exit design (Yue et al. 2024), and MoLe-VLA's dynamic layer-skipping mechanism (Zhang et al. 2025d), all of which reduce redundant computation. 3) Compression-based approaches such as NORA (Hung et al. 2025) and BitVLA (Wang et al. 2025a) focus on downsizing models while retaining task performance. While these strategies enhance internal efficiency, they often neglect the alignment between visual inputs and instruction semantics—an essential component for embodied creativity, which is inherently visual.

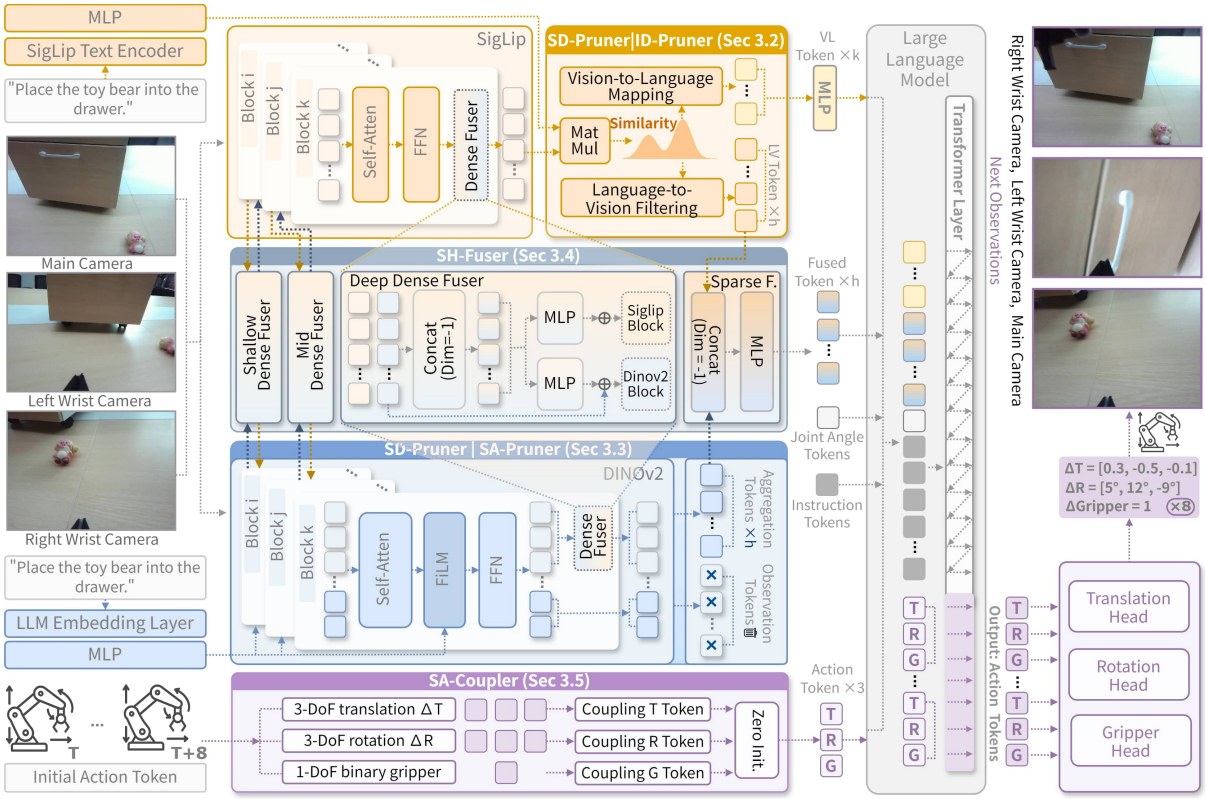

Figure 2: Overview of the SemanticVLA Framework. Observations are processed through two parallel pathways: instruction-aware encoding via SigLIP-based Instruction-driven Pruner and spatial-aware encoding via DINOv2-based Spatial-aggregation Pruner, tightly fused through a shared Semantic-complementary Hierarchical Fuser. Action inputs are initialized via Semantic-conditioned Action Coupler to optimize the sparsified perception to action type transition in large language model.

3 SemanticVLA

3.1 Proposed Framework

We define the input context as $\mathbf{X} = \{\mathcal{V}, \mathbf{q}, \ell\}$, where $\mathcal{V}$ denotes the real-time visual observation, $\mathbf{q}$ represents the robot's current proprioceptive state (e.g., joint angles and end-effector pose), and ℓ is the provided natural language instruction. The model predicts a chunk of K future actions $\mathbf{A} = [\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_{K-1}] \in \mathbb{R}^{(K\times D)\times d}$, where D signifies the dimensionality of each atomic action vector (e.g., $D = 7$ for 3-DoF translation, 3-DoF rotation, and gripper control).

As shown in Fig. 2, SemanticVLA processes $\mathcal{V}$ through two parallel pathways: 1) a SigLIP-based visual encoder whose outputs are sparsified via **ID-Pruner** according to instruction guidance; and 2) a DINOv2-based spatial encoder that captures dense geometric features via **SA-Pruner**. These two streams are hierarchically integrated through the **SH-Fuser** to produce a task-relevant representation $\mathbf{Z}$. And then, $\mathbf{Z}$ is concatenated with ℓ, $\mathbf{q}$, and K learnable action placeholders, constructing the input for parallel decoding:

$$\tilde{\mathbf{X}} = [\mathbf{Z}, \mathbf{q}, \ell, \mathbf{0}_0, \mathbf{0}_1, \ldots, \mathbf{0}_{K-1}] \tag{1}$$

where, following **SA-Coupler**, each placeholder $\mathbf{0}_i = \{\mathbf{t}_i^0, \mathbf{r}_i^0, \mathbf{g}_i^0\} \in \mathbb{R}^{3\times d}$ explicitly separates translation, rotation, and gripper tokens while jointly encoding the full 7-DoF atomic action vector. Lastly, a bidirectional decoding process $f_{\parallel}(\cdot)$ performs a single forward pass on $\tilde{\mathbf{X}}$ to concurrently generate all K future actions: $\mathbf{A} = f_{\parallel}(\tilde{\mathbf{X}})$. This structured pipeline enables SemanticVLA to achieve efficient and instruction-aware manipulation by tightly integrating semantic sparsification, hierarchical fusion, and compositional action modeling within a unified architecture.

3.2 ID-Pruner for SigLIP

ID-Pruner executes a dual pruning mechanism between visual tokens $\{\mathbf{v}_i^{Sig}\}_{i=1}^{N} \in \mathbb{R}^{N\times d_v^{Sig}}$, extracted via SigLIP encoder, and instruction embeddings $\{\mathbf{l}_j^{Sig}\}_{j=1}^{M} \in \mathbb{R}^{M\times d_l^{Sig}}$ generated by the SigLip text encoder. This process operates through two pathways: Vision-to-Language Mapping for global action cues, and Language-to-Vision Filtering for local semantic anchors, as shown in Fig. 3 (left).

Step 1: Cosine Similarity Matrix Construction. Each instruction token $\mathbf{l}_j^{Sig}$ is projected into the visual token space using a transformation matrix $\mathbf{W}_l \in \mathbb{R}^{d_v^{Sig}\times d_l^{Sig}}$, followed by cosine similarity computation with each visual token:

$$\mathbf{S}_{ij} = \mathrm{sim}\left(\mathbf{W}_l \cdot \mathbf{l}_j^{Sig},\ \mathbf{v}_i^{Sig}\right) = \frac{(\mathbf{W}_l \cdot \mathbf{l}_j^{Sig})^\top \mathbf{v}_i^{Sig}}{\|\mathbf{W}_l \cdot \mathbf{l}_j^{Sig}\| \cdot \|\mathbf{v}_i^{Sig}\|} \tag{2}$$

where $\mathbf{S} \in \mathbb{R}^{N\times M}$ denotes the visual-instruction similarity matrix, and captures the fine-grained relevance between

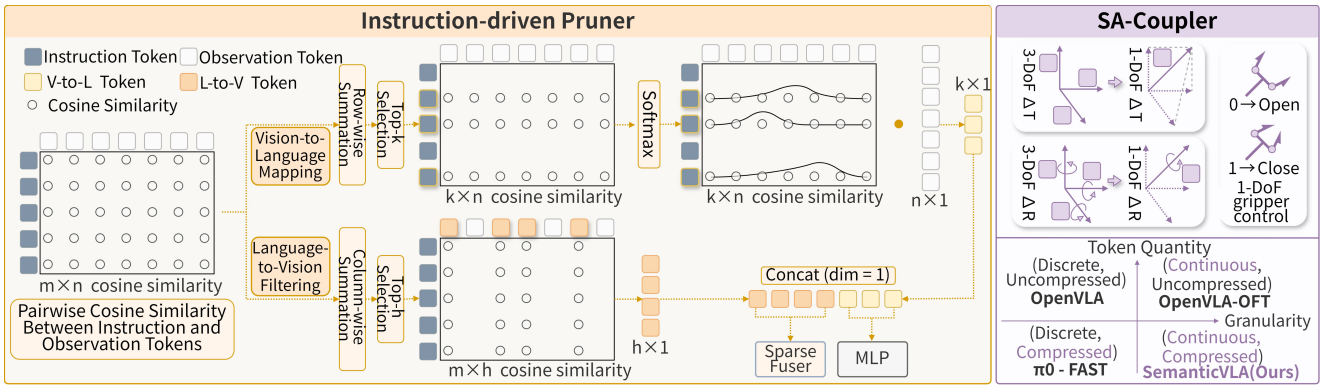

Figure 3: Illustration of Instruction-driven Pruner for SigLIP (left) and Semantic-conditioned Action Coupler (right).

every visual patch and every word in the instruction.

Step 2: Vision-to-Language Mapping. This pathway identifies key instruction tokens (e.g., target nouns, action verbs) and pinpoint their visual correspondents in the scene. We compute a saliency score s_j^{VL} for each instruction token $\mathbf{l}_j^{Sig}$ by aggregating its similarities across all visual tokens.

$$s_j^{\mathrm{VL}} = \sum_{i=1}^{N} \mathbf{S}_{ij}, \quad \mathcal{I}_{\mathrm{top}-k} = \arg\max_k \{s_j^{\mathrm{VL}}\} \tag{3}$$

where $\mathcal{I}_{\mathrm{top}-k} = \{i_1, \ldots, i_k\}$ represents the indices of the top-k most salient instruction tokens, typically capturing the task's key elements. Subsequently, for each selected instruction token $\mathbf{l}_p^{Sig}$ (where $p \in \mathcal{I}_{\mathrm{top}-k}$), it is aggregated using softmax-normalized weights to form a global cue vector:

$$\alpha_{j_p} = \frac{\exp\left(s_{j_p}^{\mathrm{VL}}\right)}{\sum_{q=1}^{k} \exp\left(s_{j_p}^{\mathrm{VL}}\right)}, \quad \mathbf{v}_{\mathrm{p}}^{Sig} = \sum_{p=1}^{k} \alpha_{j_p} \cdot \mathbf{v}_{j_p}^{Sig} \tag{4}$$

Finally, we obtain$\mathcal{V}^{\mathrm{VL}} = \{\mathbf{v}_p^{Sig} | p \in \mathcal{I}_{\mathrm{top}-k}\} \in \mathbb{R}^{k \times d_v}$ as the instruction-aware global action cue feature.

Step 3: Language-to-Vision Filtering. In contrast to the VL Mapping path, this pathway aims to identify and preserve visual regions that are most relevant to the local semantic anchors of overall instruction. We compute a comprehensive relevance score s_i^{LV} for each visual token $\mathbf{v}_i^{Sig}$ by aggregating its similarities across all instruction tokens. This evaluates the total "response strength" of each visual region to the instruction as a whole:

$$s_i^{\mathrm{LV}} = \sum_{j=1}^{M} \mathbf{S}_{ij}, \quad \mathcal{I}_{\mathrm{top}-h} = \arg\max_h \{s_i^{\mathrm{LV}}\} \tag{5}$$

where $\mathcal{I}_{\mathrm{top}-h} = \{i_1, \ldots, i_h\}$ represents the indices of the top-h visual tokens. We select the Top-h visual tokens with the highest scores, forming a sparse yet critical visual subset $\mathcal{V}^{\mathrm{LV}} = \{\mathbf{v}_q | q \in \mathcal{I}_{\mathrm{top}-h}\}$. This step effectively filters out background noise and irrelevant distractors, achieving an initial focus on key regions with local semantic anchors.

Step 4. The final output of ID-Pruner is the union of the two pruned sets of visual tokens: $\mathcal{V}^{\mathrm{VL}} \cup \mathcal{V}^{\mathrm{LV}} \in \mathbb{R}^{(k+h) \times d_v^{Sig}}$. This dual-path design balances global action cues (to avoid misinterpreting manipulation details) with local semantic anchors (to avoid missing key regions). It achieves efficient visual compression while maximally preserving the essential vision-language-action joint information.

3.3 SA-Pruner for DINOv2

Parallel to the SigLIP-based pruning branch, we employ the DINOv2-based SA-Pruner to extract dense spatial representations from the observation tokens $\mathcal{V}^{Din} \in \mathbb{R}^{N \times d_v^{Din}}$. To facilitate spatial aggregation, a set of zero-initialized aggregation tokens $\mathcal{V}^{\mathrm{Agg}} \in \mathbb{R}^{(N/8) \times d_v^{Din}}$ is appended to $\mathcal{V}^{Din}$. As a self-supervised model, DINOv2 excels at capturing fine-grained spatial structure and geometric details. This makes it an ideal source of dense spatial features to complement the sparse, object-centric features provided by ID-Pruner.

To align dense spatial features with task semantics, we introduce lightweight instruction modulation via a FiLM layer. A pooled instruction representation $\bar{\ell}^{Din}$ is passed to produce affine transformation parameters (scale γ and shift β):

$$(\gamma, \beta) = \mathrm{FiLM}(\bar{\ell}^{Din}) \in \mathbb{R}^{d_v^{Din} \times 2} \tag{6}$$

These parameters are subsequently applied to the concatenated $\mathcal{V}^{Din} \cup \mathcal{V}^{\mathrm{Agg}}$,yielding the modulated representation:

$$(\mathcal{V}^{Din} \cup \mathcal{V}^{\mathrm{Agg}})' = (\mathbf{1} + \gamma) \odot \mathrm{Attn}(\mathcal{V}^{Din} \cup \mathcal{V}^{\mathrm{Agg}}) + \beta \tag{7}$$

where $\odot$ denotes element-wise multiplication. This formulation allows spatial features to be dynamically adjusted based on task context, enabling aggregation onto $\mathcal{V}^{\mathrm{Agg}}$, while maintaining semantic relevance. The resulting representations of SA-Pruner are structurally aligned with the outputs of ID-Pruner, facilitating effective cross-modal fusion.

3.4 SH-Fuser cross SigLip & DINOv2

SH-Fuser hierarchically integrates sparse semantic features from ID-Pruner with dense geometric-rich features from SA-Pruner through dynamic, layer-wise modulation rather than simple late-stage concatenation, as shown in Fig. 2.

Dense-Fuser. This module is inserted between multiple Transformer blocks at different depths (for example, once in superficial, intermediate, and deep layers). This hierarchical integration makes sure that semantic cues (from SigLIP) are enhanced with corresponding spatial-geometric priors (from DINOv2) at each stage, thus enabling the synergistic enhancement of the two complementary visual streams. For the $b-th$ block, the fusion operation is defined as:

$$\mathcal{V}_b^{Fusion} = \mathrm{MLP}(\mathrm{Concat}(\mathcal{V}_b^{Sig}, \mathcal{V}_b^{Din})) \in \mathbb{R}^{N \times d_v} \tag{8}$$

Method	Spatial		Object		Goal		Long		Overall	
	SR ↑	RK ↓	SR ↑	RK ↓	SR ↑	RK ↓	SR ↑	RK ↓	SR ↑	RK ↓
Octo fine-tuned *[RSS'24]* (Team et al. 2024)	78.9	10	85.7	10	84.6	8	51.1	10	75.1	10
π_0 fine-tuned *[arXiv'24]* (Black et al. 2024)	96.8	4	98.8	2	95.8	3	85.2	6	94.2	6
OpenVLA *[CoRL'24]* (Kim et al. 2024)	84.7	9	88.4	9	79.2	9	53.7	9	76.5	9
OpenVLA-OFT *[arXiv'25]* (Kim et al. 2025)	97.6	2	98.4	3	**97.9**	**1**	94.5	2	97.1	2
STAR *[ICML'25]* (Hao et al. 2025)	95.5	5	98.3	5	95.0	5	88.5	5	94.3	5
CoT-VLA *[CVPR'25]* (Zhao et al. 2025)	87.5	8	91.6	7	87.6	7	69.0	7	83.9	7
PD-VLA† *[arXiv'25]* (Song et al. 2025)	95.5	5	96.7	6	94.9	6	91.7	3	94.7	4
SpatialVLA *[RSS'25]* (Qu et al. 2025)	88.2	7	89.9	9	78.6	10	55.5	8	78.1	8
SemanticVLA-Lite	97.0	3	98.4	3	95.4	4	92.4	3	95.8	3
SemanticVLA	**98.6**	**1**	**99.6**	**1**	97.6	2	**94.8**	**1**	**97.7**	**1**

Table 1: Simulation Results. Comparison of task success rates (SR) and ranks (RK) across four suites in the LIBERO.

Sparse-Fuser. At the final stage, Sparse-Fuser merges the salient outputs from both pruning paths, where $\mathcal{V}^{LV}$ originates from the ID-Pruner and $\mathcal{V}^{Agg}$ from the SA-Pruner, forming a compact representation:

$$\mathbf{Z}^{Fusion} = \text{MLP}(\text{Concat}(\mathcal{V}^{LV}, \mathcal{V}^{Agg})) \in \mathbb{R}^{h \times d_l} \quad (9)$$

The Semantic-complementary Hierarchical Fuser reduces visual tokens by 8–16× while preserving discriminative representations. This design not only greatly improves computational efficiency (through token pruning) but also enhances task performance by making use of the complementary advantages of semantic grounding and geometric precision.

3.5 Semantic-conditioned Action Coupler

Building on Sections 3.1-3.3, we construct the visual input to the LLM as a composite token set $\mathbf{Z} = \mathbf{Z}^{\text{VL}} \cup \mathbf{Z}^{\text{Fusion}}$ ($\mathbf{Z}^{\text{VL}} = \text{MLP}(\mathcal{V}^{\text{VL}}) \in \mathbb{R}^{h \times d_l}$), which integrates both semantic and spatially-aligned visual information.

To optimize the mapping from vision to action, we depart from the conventional VLA formulation that discretizes 7-DoF actions into 7 independent binned tokens, each corresponding to a single DoF. Instead, as illustrated in Fig. 2 and Fig. 3 (right), we introduce a novel Semantic-conditioned Action Coupler that restructures the visual-to-action pipeline in a more structured and intuitive manner:

Token-Level Semantic Alignment. Each of the three fundamental motion primitives (3-DoF translation, 3-DoF rotation, and 1-DoF gripper control) is represented by a single token, enabling unified and semantically coherent modeling of atomic action types:

$$\mathbf{0}_i = \{\mathbf{t}_i^0, \mathbf{r}_i^0, \mathbf{g}_i^0\} \in \mathbb{R}^{3 \times d_l} \quad (10)$$

Head-Level Modularity for Action Prediction. The input sequence $[\mathbf{Z}, \mathbf{q}, \ell, \mathbf{0}_0, \mathbf{0}_1, \ldots, \mathbf{0}_{K-1}]$ is fed into the $f_{\|}(\cdot)$, yielding an updated action representation $\mathbf{h}_i = \{\mathbf{t}_i^h, \mathbf{r}_i^h, \mathbf{g}_i^h\} \in \mathbb{R}^{3 \times d_l}$. During token-to-value decoding, we design three prediction heads, each specialized for one action type, to directly regress continuous motion parameters:

$$\mathbf{d}_{i,u} = \mathbf{W}_u \mathbf{h}_i + \mathbf{b}_u, \quad \mathbf{W}_u \in \mathbb{R}^{D_u \times d_l}, \mathbf{b}_u \in \mathbb{R}^{D_u} \quad (11)$$

where $u \in \{\text{trans}, \text{rot}, \text{grip}\}$, and D_u denotes the dimensionality of degrees of freedom for action type u.

4 Experiments

4.1 Experiment Settings

All experiments are conducted on 8× A800 (80GB) GPUs.

Simulation & Real-World Setup. Simulation evaluations are performed on the LIBERO benchmark (Liu et al. 2023a), which includes four task suites: Spatial, Object, Goal, and Long, each with 500 human-teleoperated demonstrations. Real-world experiments are conducted on the AgileX Cobot Magic platform (Fu, Zhao, and Finn 2024), covering object placement, drawer manipulation, multi-step deformable tasks, and two additional scenarios, with 60, 60, 45, and 105 human-teleoperated demonstrations, respectively.

Baselines. We compare against **50+ SOTA baselines** on LIBERO, including OpenVLA, Octo and π_0. Efficiency comparisons are conducted under identical settings with OpenVLA and its accelerated variants, OpenVLA-OFT (the best-performing baseline) and PD-VLA. The top-3 performing models are further validated in real-world deployments.

4.2 Overall Performance & Efficiency

Simulation Experimental Analysis.

1) As shown in **Tab. 1**, SemanticVLA achieves the highest success rate (97.7%, rank 1), consistently outperforming recent SOTA methods through its Semantic-Aligned Sparsification and Enhancement design. SemanticVLA-Lite attains 95.8% (rank 3) with reduced complexity, highlighting the robustness and scalability.

2) As reported in **Tab. 2**, SemanticVLA achieves superior training (FLOPs, cost) and inference (latency, throughput) efficiency while utilizing only 1/16 or 1/8 of visual inputs and 3/7 of action representations. It significantly outperforms OpenVLA and other efficient baselines.

Real-world Experimental Analysis.

1) As shown in **Tab. 3**, SemanticVLA achieves a success rate of up to 77.8% on three challenging long-horizon tasks (Object Placement, Drawer Manipulation, and T-shirt Folding) outperforming OpenVLA-OFT by 22.2%.

Method	Z & H tokens ↓	FLOPs ↓	Training Cost ↓	Latency ↓	Throughput ↑	LIBERO SR ↑
OpenVLA† (Kim et al. 2024)	256 & 7	8.48 T	11.7 h	0.240 s	4.2 Hz	76.5%
OpenVLA-OFT† (Kim et al. 2025)	256 & 7	8.45 T	12.3 h	0.134 s	59.7 Hz	97.1%
PD-VLA† (Song et al. 2025)	256 & 7	8.48 T	11.7 h	0.143 s	55.9 Hz	94.7%
SemanticVLA-Lite	16 & 3	1.93 T	3.6 h	0.087 s	92.0 Hz	95.8%
SemanticVLA	32 & 3	2.37 T	3.9 h	0.089 s	89.9 Hz	97.7%

Table 2: Efficiency Results in Simulation. SemanticVLA-Lite and SemanticVLA achieve the highest efficiency and the best trade-off between efficiency and performance, respectively. "†" denotes our reproduced results. Z and H indicate the number of visual input tokens and initialized action tokens. Throughput refers to the number of actions predicted per second.

Method	Object Placement		Drawer Manipulation			T-shirt Folding			Overall
	Bear →Plate	+Raccoon→Bowl	Open	+Place	+Close	Step 1	+Step 2	+Step 3	SR
VQ-BeT (Lee et al. 2024)*	5/10	3/10	4/10	3/10	1/10	-	-	-	20.0%
QueST (Mete et al. 2024)*	6/10	4/10	3/10	1/10	0/10	-	-	-	20.0%
STAR* (Hao et al. 2025)	8/10	6/10	6/10	4/10	3/10	-	-	-	45.0%
PD-VLA†	7.3/10	6.7/10	6.0/10	5.3/10	4.7/10	6.7/10	5.3/10	4.0/10	51.1%
OpenVLA-OFT†	8.0/10	6.7/10	7.3/10	6.7/10	5.3/10	6.7/10	6.0/10	4.7/10	55.6%
SemanticVLA-Lite	8.0/10	7.3/10	8.0/10	6.7/10	5.3/10	8.7/10	8.0/10	6.7/10	62.2%
SemanticVLA	**9.3/10**	**9.3/10**	**8.7/10**	**7.3/10**	**6.0/10**	**9.3/10**	**8.7/10**	**8.0/10**	**77.8%**

Table 3: Real-World Results. Comparison of AgileX Cobot Magic task and subtask success rates. Decimal values arise from averaging over 15 trials. The overall SR reflects only the final subtask success rates. "†" denotes our reproduced results.

SigLip	DINOv2	Spatial	Object	Goal	Long	Overall
ID-Pruner	ID-Pruner	96.6	97.4	83.8	89.6	91.9
SA-Pruner	SA-Pruner	96.2	96.8	95.2	90.2	94.6
SA-Pruner	ID-Pruner	96.2	98.0	94.4	91.4	95.0
ID-Pruner	**SA-Pruner**	**98.2**	**99.0**	**97.2**	**93.8**	**97.1**

Table 4: Ablation on SD-Pruner.

Spf.Ratio	SR ↑	FLOPs ↓	Taining Cost ↓	Latency ↓
4 ×	97.7	3.28 T	4.5 h	0.093 s
8 ×	97.7	2.37 T	3.9 h	0.089 s
16 ×	95.8	1.93 T	3.6 h	0.087 s
32 ×	92.0	1.72 T	3.5 h	0.086 s
FastV†	88.8	2.71 T	-	0.091 s
SliME†	85.6	2.71 T	3.8 h	0.089 s

Table 5: Ablation on sparsification ratio (Spf.Ratio) & Comparison with FastV (Chen et al. 2024b) and SliME (Zhang et al. 2024b) under 8×. "†" denotes reproduced results.

2) Combined with **Tab. 2**, SemanticVLA demonstrates substantial improvements in both training and inference. Notably, in ALOHA setup with 25 actions per chunk, SA-Coupler reduces action tokens per inference from 350 to 150, substantially cutting inference overhead.

3) As visualized in **Fig. 4**, SemanticVLA consistently completes complex instructions across different execution stages, demonstrating strong instruction-following capability and generalization in real-world scenarios.

4.3 Ablation Studies

Ablation on SD-Pruner. As shown in **Tab. 4**, 1) SigLIP with ID-Pruner enables instruction-driven token pruning via language-supervised feature alignment, thus maximizing semantic density. 2) DINOv2 with SA-Pruner preserves global geometric structure through token aggregation, while injecting lightweight semantics through FiLM. Their targeted combination in SemanticVLA yields both semantic focusing and geometric perception, outperforming inverse or single configurations by 2.1%–5.2% in success rate.

Ablation on sparsification ratio. **Tab. 5** presents results under varying sparsification ratios $R \in \{4, 8, 16, 32\}$. The chosen setting $R = 8$ attains a 97.7% success rate, offering a favorable trade-off between performance and efficiency. $R = 16$ yields a marginal 1.9% drop and defines the SemanticVLA-Lite variant. In contrast, $R = 4$ retains redundancy, limiting speedup, while $R = 32$ discards critical semantic context, degrading performance. Furthermore, compared to other plug-and-play sparsification baselines at the same 8× compression level (e.g., FastV and SliME), SemanticVLA exhibits significantly better performance, underscoring that only instruction-aware pruning combined with structural preservation via HF-Fuser achieves Pareto-optimality in both performance and efficiency.

Ablation Study on HF-Fuser and SA-Coupler. As shown in **Tab. 6**, HF-Fuser and SA-Coupler provide complementary improvements across all tasks, with the largest gains in long-horizon tasks, highlighting the effectiveness of SemanticVLA's vision-action co-design. Specifically, HF-Fuser improves success rates by hierarchically integrating fine-grained observation tokens from both visual encoders.

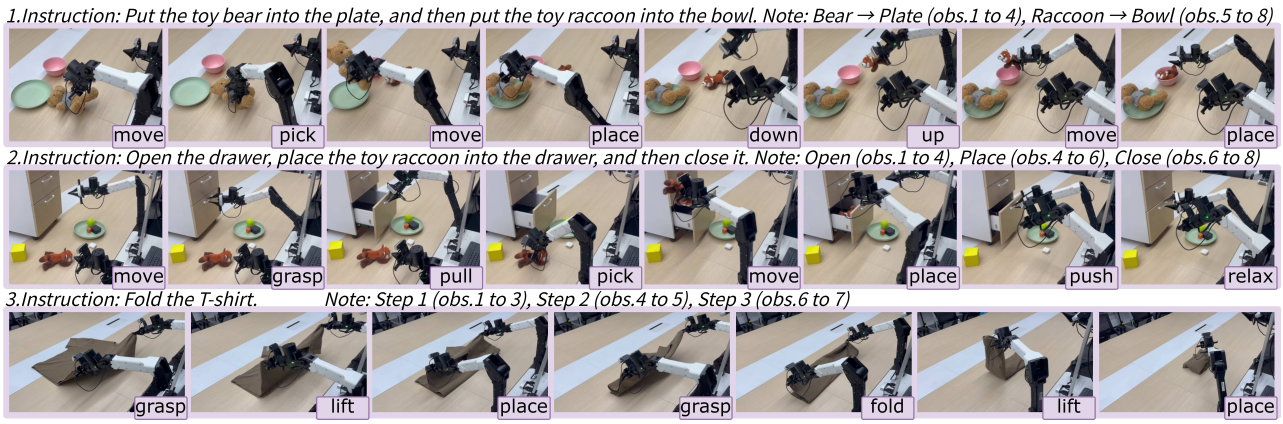

Figure 4: Visualization of SemanticVLA's manipulation process on three long-horizon real-world tasks.

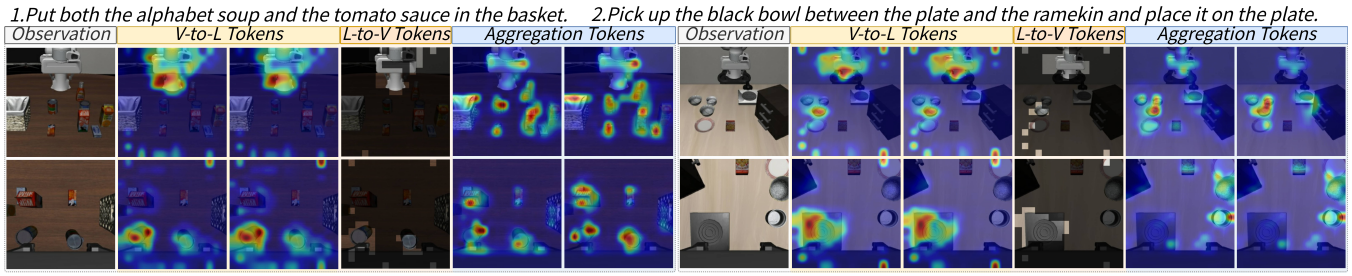

Figure 5: Visualization of 1) attention maps from V-to-L tokens to observation patches, capturing global action cues; 2) selected L-to-V token sets, where highlighted patches represent local semantic anchors; and 3) attention maps from aggregation tokens to patches, reflecting spatial features that complement ID-Pruner via HF-Fuser.

HF-F	SA-C	Spatial	Object	Goal	Long	Overall
×	×	95.2	96.0	94.4	88.6	93.6
✓	×	96.8	97.4	95.6	92.4	95.6
×	✓	95.6	96.4	95.2	89.2	94.1
✓	✓	**98.2**	**99.0**	**97.2**	**93.8**	**97.1**

Table 6: Ablation Study on HF-Fuser (HF-F) and SA-Coupler (SA-C). "×" for SA-Coupler denotes the use of uncompressed 7-DoF action tokens.

SA-Coupler eliminates redundant action tokens and reduces overfitting in the action space, particularly under sparse visual inputs. Together, these modules operate at different token granularities and enhance cross-modal alignment, yielding synergistic improvements beyond additive contributions.

4.4 Qualitative Analysis

Fig. 4 shows that the model consistently produces instruction-aligned action sequences with minimal deviation on three long-horizon real-world tasks. Each frame marks a key subgoal transition, demonstrating robustness in complex manipulation workflows. **Fig. 5** demonstrates that ID-Pruner in SigLIP emphasizes both global action cues and local semantic anchors, while SA-Pruner in DINOv2 focuses on geometric structure, highlighting their complementary strengths in semantic and spatial grounding. These visualizations indicate that the synergy among SD-Pruner, HF-Fuser, and SA-Coupler produces interpretable intermediate representations and reliable downstream control, validating the effectiveness of Semantic-Aligned Sparsification and Enhancement in real-world execution.

5 Conclusion

We present SemanticVLA, a novel framework for Semantic-Aligned Sparsification and Enhancement in robotic manipulation. By integrating Semantic-guided Dual Visual Pruner (SD-Pruner) for semantic-guided visual sparsification, Semantic-complementary Hierarchical Fuser (SH-Fuser) for cross-encoder semantic-structural fusion, and Semantic-conditioned Action Coupler (SA-Coupler) for modular action control, SemanticVLA achieves state-of-the-art task success with significantly reduced computational cost. Extensive evaluations on simulation and real-world tasks demonstrate its robustness, scalability, and efficiency.

Acknowledgments

This study is supported by National Natural Science Foundation of China (Grant No. 62306090), Natural Science Foundation of Beijing, China(Grant No. L254018), Jiangsu Science and Technology Major Program (Grant No. BG2024041), Natural Science Foundation of Guangdong

Province of China (Grant No. 2024A1515010147), Natural Science Foundation of Shenzhen City of China (Grant No. JCYJ20250604145700001) and Shenzhen Science and Technology Program (KQTD20240729102207002).

References

Black, K.; Brown, N.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Groom, L.; Hausman, K.; Ichter, B.; et al. 2024. π_0: A vision-language-action flow model for general robot control, 2024. *URL https://arxiv. org/abs/2410.24164*.

Bu, Q.; Yang, Y.; Cai, J.; Gao, S.; Ren, G.; Yao, M.; Luo, P.; and Li, H. 2025. Learning to Act Anywhere with Task-centric Latent Actions. *arXiv preprint arXiv:2502.14420*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *ICCV*, 9650–9660.

Chen, G.; Shen, L.; Shao, R.; Deng, X.; and Nie, L. 2024a. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26540–26550.

Chen, G.; Zhou, X.; Shao, R.; Lyu, Y.; Zhou, K.; Wang, S.; Li, W.; Li, Y.; Qi, Z.; and Nie, L. 2025. Less is More: Empowering GUI Agent with Context-Aware Simplification. In *ICCV*.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024b. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, 19–35. Springer.

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

Fu, Z.; Zhao, T. Z.; and Finn, C. 2024. Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation. In *CoRL*.

Hao, L.; Qi, L.; Rui, S.; Xiang, D.; Yinchuan, L.; Jianye, H.; and Liqiang, N. 2025. STAR: Learning Diverse Robot Skill Abstractions through Rotation-Augmented Vector Quantization. *ICML*.

Hung, C.-Y.; Sun, Q.; Hong, P.; Zadeh, A.; Li, C.; Tan, U.; Majumder, N.; Poria, S.; et al. 2025. Nora: A small open-sourced generalist vision language action model for embodied tasks. *arXiv preprint arXiv:2504.19854*.

Intelligence, P.; Black, K.; Brown, N.; Darpinian, J.; Dhabalia, K.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; et al. 2025. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*.

Kim, M. J.; Kim, M. J.; Finn, C.; and Liang, P. 2025. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*.

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

Lee, S.; Wang, Y.; Etukuru, H.; Kim, H. J.; Shafiullah, N. M. M.; and Pinto, L. 2024. Behavior generation with latent actions. *arXiv preprint arXiv:2403.03181*.

Li, Q.; Liang, Y.; Wang, Z.; Luo, L.; Chen, X.; Liao, M.; Wei, F.; Deng, Y.; Xu, S.; Zhang, Y.; et al. 2024a. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*.

Li, W.; Hu, B.; Shao, R.; Shen, L.; and Nie, L. 2025a. Lion-fs: Fast & slow video-language thinker as online video assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3240–3251.

Li, W.; Zhang, R.; Shao, R.; He, J.; and Nie, L. 2025b. Cogvla: Cognition-aligned vision-language-action model via instruction-driven routing & sparsification. *arXiv preprint arXiv:2508.21046*.

Li, Z.; Xie, Y.; Shao, R.; Chen, G.; Jiang, D.; and Nie, L. 2024b. Optimus-1: Hybrid Multimodal Memory Empowered Agents Excel in Long-Horizon Tasks. In *NeurIPS*, volume 37, 49881–49913.

Li, Z.; Xie, Y.; Shao, R.; Chen, G.; Jiang, D.; and Nie, L. 2025c. Optimus-2: Multimodal minecraft agent with goal-observation-action conditioned policy. In *CVPR*, 9039–9049.

Liu, B.; Zhu, Y.; Gao, C.; Feng, Y.; Liu, Q.; Zhu, Y.; and Stone, P. 2023a. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. *arXiv preprint arXiv:2306.03310*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, J.; Liu, M.; Wang, Z.; An, P.; Li, X.; Zhou, K.; Yang, S.; Zhang, R.; Guo, Y.; and Zhang, S. 2024. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. *Advances in Neural Information Processing Systems*, 37: 40085–40110.

Lv, Q.; Li, H.; Deng, X.; Shao, R.; Li, Y.; Hao, J.; Gao, L.; Wang, M. Y.; and Nie, L. 2025. Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation. In *CVPR*, 17394–17404.

Lyu, Y.; Shao, R.; Chen, G.; Zhu, Y.; Guan, W.; and Nie, L. 2025. PUMA: Layer-Pruned Language Model for Efficient Unified Multimodal Retrieval with Modality-Adaptive Learning. In *ACM MM*.

Mete, A.; Xue, H.; Wilcox, A.; Chen, Y.; and Garg, A. 2024. QueST: Self-Supervised Skill Abstractions for Learning Continuous Control. arXiv:2407.15840.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Pertsch, K.; Stachowicz, K.; Ichter, B.; Driess, D.; Nair, S.; Vuong, Q.; Mees, O.; Finn, C.; and Levine, S. 2025. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*.

Qu, D.; Song, H.; Chen, Q.; Yao, Y.; Ye, X.; Ding, Y.; Wang, Z.; Gu, J.; Zhao, B.; Wang, D.; et al. 2025. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Model. *arXiv preprint arXiv:2501.15830*.

Shao, R.; Li, W.; Zhang, L.; Zhang, R.; Liu, Z.; Chen, R.; and Nie, L. 2025. Large vlm-based vision-language-action models for robotic manipulation: A survey. *arXiv preprint arXiv:2508.13073*.

Shao, R.; Wu, T.; and Liu, Z. 2023. Detecting and grounding multi-modal media manipulation. In *CVPR*, 6904–6913.

Shao, R.; Wu, T.; Wu, J.; Nie, L.; and Liu, Z. 2024. Detecting and grounding multi-modal media manipulation and beyond. *TPAMI*.

Shukor, M.; Aubakirova, D.; Capuano, F.; Kooijmans, P.; Palma, S.; Zouitine, A.; Aractingi, M.; Pascal, C.; Russi, M.; Marafioti, A.; et al. 2025. SmolVLA: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*.

Song, W.; Chen, J.; Ding, P.; Zhao, H.; Zhao, W.; Zhong, Z.; Ge, Z.; Ma, J.; and Li, H. 2025. Accelerating Vision-Language-Action Model Integrated with Action Chunking via Parallel Decoding. *arXiv preprint arXiv:2503.02310*.

Sun, Y.; Li, Y.; Sun, R.; Liu, C.; Zhou, F.; Jin, Z.; Wang, L.; Shen, X.; Hao, Z.; and Xiong, H. 2025. Audio-enhanced vision-language modeling with latent space broadening for high quality data expansion. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 4872–4881.

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

Wang, H.; Xiong, C.; Wang, R.; and Chen, X. 2025a. BitVLA: 1-bit Vision-Language-Action Models for Robotics Manipulation. *arXiv preprint arXiv:2506.07530*.

Wang, Z.; Sun, Y.; Wang, H.; Jing, B.; Shen, X.; Dong, X. L.; Hao, Z.; Xiong, H.; and Song, Y. 2025b. Reasoning-Enhanced Domain-Adaptive Pretraining of Multimodal Large Language Models for Short Video Content Governance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 1104–1112.

Xie, M.; Zeng, S.; Chang, X.; Liu, X.; Pan, Z.; Xu, M.; and Wei, X. 2025. SeqGrowGraph: Learning Lane Topology as a Chain of Graph Expansions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 27166–27175.

Yue, Y.; Wang, Y.; Kang, B.; Han, Y.; Wang, S.; Song, S.; Feng, J.; and Huang, G. 2024. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. *NeurIPS*, 37: 56619–56643.

Zeng, S.; Chang, X.; Liu, X.; Pan, Z.; and Wei, X. 2024. Driving with prior maps: Unified vector prior encoding for autonomous vehicle mapping. *arXiv preprint arXiv:2409.05352*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *ICCV*, 11975–11986.

Zhang, F.; Chen, G.; Wang, H.; and Zhang, C. 2024a. CF-DAN: Facial-expression recognition based on cross-fusion dual-attention network. *Computational Visual Media*, 10(3): 593–608.

Zhang, J.; Cai, K.; Fan, Y.; Wang, J.; and Wang, K. 2025a. CF-VLM:CounterFactual Vision-Language Fine-tuning. arXiv:2506.17267.

Zhang, J.; Fan, Y.; Lin, W.; Chen, R.; Jiang, H.; Chai, W.; Wang, J.; and Wang, K. 2025b. GAM-Agent: Game-Theoretic and Uncertainty-Aware Collaboration for Complex Visual Reasoning. arXiv:2505.23399.

Zhang, J.; Huang, Z.; Fan, Y.; Liu, N.; Li, M.; Yang, Z.; Yao, J.; Wang, J.; and Wang, K. 2025c. KABB: Knowledge-Aware Bayesian Bandits for Dynamic Expert Coordination in Multi-Agent Systems. arXiv:2502.07350.

Zhang, R.; Dong, M.; Zhang, Y.; Heng, L.; Chi, X.; Dai, G.; Du, L.; Wang, D.; Du, Y.; and Zhang, S. 2025d. MoLe-VLA: Dynamic Layer-skipping Vision Language Action Model via Mixture-of-Layers for Efficient Robot Manipulation. *arXiv preprint arXiv:2503.20384*.

Zhang, R.; Shao, R.; Chen, G.; Zhang, M.; Zhou, K.; Guan, W.; and Nie, L. 2025e. Falcon: Resolving visual redundancy and fragmentation in high-resolution multimodal large language models via visual registers. *arXiv preprint arXiv:2501.16297*.

Zhang, Y.-F.; Wen, Q.; Fu, C.; Wang, X.; Zhang, Z.; Wang, L.; and Jin, R. 2024b. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*.

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

Zhou, S.; Li, L.; Zhang, X.; Zhang, B.; Bai, S.; Sun, M.; Zhao, Z.; Lu, X.; and Chu, X. 2024. Lidar-ptq: Post-training quantization for point cloud 3d object detection. *arXiv preprint arXiv:2401.15865*.

Zhou, S.; Yuan, Z.; Yang, D.; Hu, X.; Qian, J.; and Zhao, Z. 2025. Pillarhist: A quantization-aware pillar feature encoder based on height-aware histogram. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 27336–27345.

Zhu, Y.; Lyu, Y.; Yu, Z.; Shao, R.; Zhou, K.; and Nie, L. 2025. Emosym: A symbiotic framework for unified emotional understanding and generation via latent reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 5451–5460.