Novelty Assessment Report

Paper: From Spatial to Actions: Grounding Vision-Language-Action Model in Spatial Foundation Priors
PDF URL: https://openreview.net/pdf?id=fzmittHfq3
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-04

Abstract

Existing vision-language-action (VLA) models act in 3D real-world but are typically built on 2D encoders, leaving a spatial reasoning gap that limits generalization and adaptability. Recent 3D integration techniques for VLAs either require specialized sensors and transfer poorly across modalities, or inject weak cues that lack geometry and degrade vision-language alignment. In this work, we introduce **FALCON (From Spatial to Action)**, a novel paradigm that injects rich 3D spatial tokens into the action head. FALCON leverages spatial foundation models to deliver strong geometric priors from RGB alone, and includes an Embodied Spatial Model that can optionally fuse depth, or pose for higher fidelity when available, without retraining or architectural changes. To preserve language reasoning, spatial tokens are consumed by a Spatial-Enhanced Action Head rather than being concatenated into the vision-language backbone. These designs enable FALCON to address limitations in spatial representation, modality transferability, and alignment. In comprehensive evaluations across three simulation benchmarks and eleven real-world tasks, our proposed FALCON achieves state-of-the-art performance, consistently surpasses competitive baselines, and remains robust under clutter, spatial-prompt conditioning, and variations in object scale and height. Code will be released publicly.

Core Task Landscape

This paper addresses: **Integrating Spatial Foundation Priors into Vision-Language-Action Models for Robotic Manipulation**
A total of **45 papers** were analyzed and organized into a taxonomy with **23 categories**.

Taxonomy Overview

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

- **Spatial Representation and Encoding Methods**
- **Spatial Reasoning and Grounding Mechanisms**
- **Action Generation and Prediction**
- **Multimodal Integration and Perception**
- **Training Paradigms and Data Utilization**
- **Model Efficiency and Optimization**
- **Benchmarking and Evaluation Frameworks**
- **Specialized Applications and Task Domains**

Complete Taxonomy Tree

- Integrating Spatial Foundation Priors into Vision-Language-Action Models for Robotic Manipulation Survey Taxonomy
- Spatial Representation and Encoding Methods
  - Explicit 3D Input Integration (3 papers)
  - [2] 3ds-vla: A 3d spatial-aware vision language action model for robust multi-task manipulation (X Li, 2025) View paper
  - [17] Geovla: Empowering 3d representations in vision-language-action models (Sun Lin, 2025) View paper
  - [28] DepthVLA: Enhancing Vision-Language-Action Models with Depth-Aware Spatial Reasoning (YUAN Tianyuan, 2025) View paper
  - Implicit Spatial Understanding from 2D (3 papers)
  - [3] Evo-0: Vision-language-action model with implicit spatial understanding (Lin Tao, 2025) View paper
  - [34] GeoAware-VLA: Implicit Geometry Aware Vision-Language-Action Model (Abouzeid Ali, 2025) View paper
  - [40] Spatial Forcing: Implicit Spatial Representation Alignment for Vision-language-action Model (Li Fuhao, 2025) View paper
  - Spatial Foundation Model Integration ★ (3 papers)
  - [0] From Spatial to Actions: Grounding Vision-Language-Action Model in Spatial Foundation Priors (Anon et al., 2026) View paper
  - [37] VGGT-DP: Generalizable Robot Control via Vision Foundation Models (Ge Shijia, 2025) View paper
  - [43] Spatial-Aware VLA Pretraining through Visual-Physical Alignment from Human Videos (Yicheng Feng, 2025) View paper
  - Ego-Centric and Position Encoding (2 papers)
  - [1] Spatialvla: Exploring spatial representations for visual-language-action model (Qu, 2025) View paper
  - [4] Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy (Chen Xin-yi, 2025) View paper
- Spatial Reasoning and Grounding Mechanisms
  - Spatial Grounding and Localization (3 papers)
  - [29] RoboGround: Robotic Manipulation with Grounded Vision-Language Priors (Haifeng Huang, 2025) View paper
  - [42] Languageâ□□Guided Robot Grasping Based on Basic Geometric Shape Fitting (Q Niu, 2025) View paper
  - [44] SegGrasp: Zero-Shot Task-Oriented Grasping via Semantic and Geometric Guided Segmentation (Li Haosheng, 2024) View paper
  - Geometric and Tool-Integrated Reasoning (3 papers)
  - [13] TIGeR: Tool-Integrated Geometric Reasoning in Vision-Language Models for Robotics (Han Yi, 2025) View paper
  - [14] VisionCube: 3D-Aware Vision-Language Model for Multi-Step Spatial Reasoning (Feiyang Wang, 2025) View paper

    - [16] Lageo: a latent and geometrical framework for path and manipulation planning (Peng, 2022) View paper
    - Graph-Based Spatial Reasoning (3 papers)
    - [18] GraphCoT-VLA: A 3D Spatial-Aware Reasoning Vision-Language-Action Model for Robotic Manipulation with Ambiguous Instructions (Huang He-long, 2025) View paper
    - [31] Learning Spatial-Aware Manipulation Ordering (Yan Yu-xiang, 2025) View paper
    - [36] Toward Accurate Long-Horizon Robotic Manipulation: Language-to-Action with Foundation Models via Scene Graphs (Park, 2025) View paper
    - Spatial-Aware Policy and Planning (2 papers)
    - [15] Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints (Mingjie Pan, 2025) View paper
    - [22] Spatial policy: Guiding visuomotor robotic manipulation with spatial-aware modeling and reasoning (Liu Yi-jun, 2025) View paper
- Action Generation and Prediction
    - Predictive Action and Trajectory Modeling (3 papers)
    - [25] DreamVLA: A Vision-Language-Action Model Dreamed with Comprehensive World Knowledge (Zhang Wenyao, 2025) View paper
    - [30] VLA-4D: Embedding 4D Awareness into Vision-Language-Action Models for SpatioTemporally Coherent Robotic Manipulation (Hanyu Zhou, 2025) View paper
    - [32] GeoPredict: Leveraging Predictive Kinematics and 3D Gaussian Geometry for Precise VLA Manipulation (Jingjing Qian, 2025) View paper
    - Affordance and Intervention-Based Action (2 papers)
    - [11] Improving vision-language-action models via chain-of-affordance (Jinming Li, 2024) View paper
    - [33] Affordance Field Intervention: Enabling VLAs to Escape Memory Traps in Robotic Manipulation (Siyu Xu, 2025) View paper
    - Long-Horizon and Sequential Action Learning (2 papers)
    - [23] EchoVLA: Robotic Vision-Language-Action Model with Synergistic Declarative Memory for Mobile Manipulation (Min Lin, 2025) View paper
    - [27] LoLA: Long Horizon Latent Action Learning for General Robot Manipulation (Xiaofan Wang, 2025) View paper
- Multimodal Integration and Perception
    - Multisensory and Proprioceptive Integration (2 papers)
    - [9] Mla: A multisensory language-action model for multimodal understanding and forecasting in robotic manipulation (Liu, 2025) View paper
    - [26] RGMP: Recurrent Geometric-prior Multimodal Policy for Generalizable Humanoid Robot Manipulation (Xuetao Li, 2025) View paper
    - Multi-View and Cross-View Alignment (2 papers)
    - [35] TrackVLA++: Unleashing Reasoning and Memory Capabilities in VLA Models for Embodied Visual Tracking (Liu Jiahang, 2025) View paper
    - [41] Uni-Sight: An E2E Vision-Language-Action System Unifying Multi-View Alignment and Multi-Modal Fusion (Daixun Li, 2025) View paper
    - Semantic and Visual Feature Enhancement (2 papers)
    - [6] Magma: A Foundation Model for Multimodal AI Agents (Jianwei Yang, 2025) View paper
    - [8] Semanticvla: Semantic-aligned sparsification and enhancement for efficient robotic manipulation (Wei Li, 2025) View paper
- Training Paradigms and Data Utilization
    - Human Video and In-the-Wild Pretraining (1 papers)
    - [5] Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos (Li Qixiu, 2025) View paper
    - Embodied Prior and Knowledge Distillation (1 papers)
    - [19] Era: Transforming vlms into embodied agents via embodied prior learning and online reinforcement learning (Chen Han-yang, 2025) View paper
    - Reinforcement Learning Integration (2 papers)
    - [7] A Survey on Reinforcement Learning of Vision-Language-Action Models for Robotic Manipulation (H Deng, 2025) View paper
    - [20] Large language model-driven dynamic trajectory planning for human-guided robot assembly (Zhao Boya, 2025) View paper
- Model Efficiency and Optimization
    - Layer-Level and Structural Sparsification (1 papers)
    - [10] MoLe-VLA: Dynamic Layer-skipping Vision Language Action Model via Mixture-of-Layers for Efficient Robot Manipulation (Zhang Rong-yu, 2025) View paper
    - State Space Models for Efficiency (1 papers)
    - [38] RoboMamba: Efficient Vision-Language-Action Model for Robotic Reasoning and Manipulation (Pengju An, 2024) View paper
- Benchmarking and Evaluation Frameworks
    - Long-Horizon and Complex Task Benchmarks (1 papers)
    - [12] VLABench: A Large-Scale Benchmark for Language-Conditioned Robotics Manipulation with Long-Horizon Reasoning Tasks (Zhang Shi-duo, 2024) View paper
    - Survey and Taxonomy Studies (2 papers)
    - [21] Integrating World Models into Vision Language Action and Navigation: A Comprehensive Survey (J Sun, 2025) View paper
    - [24] Vision-Language Models Enabled Robot Manipulation (Li, 2025) View paper
- Specialized Applications and Task Domains
    - Navigation and Mobile Manipulation (1 papers)
    - [39] SoraNav: Adaptive UAV Task-Centric Navigation via Zeroshot VLM Reasoning (Song, 2025) View paper
    - Visual Tracking and Target Following (1 papers)
    - [45] Autonomously Learning to Visually Detect Where Manipulation Will Succeed (Nguyen Hai, 2012) View paper

Narrative

Core task: Integrating spatial foundation priors into vision-language-action models for robotic manipulation. The field has organized itself around several complementary branches that address different facets of this integration challenge. Spatial Representation and Encoding Methods explore how to capture and embed geometric information—ranging from depth maps and point clouds to scene graphs and affordance fields—into model architectures. Spatial Reasoning and Grounding Mechanisms focus on connecting language instructions to

physical locations and object relationships, enabling robots to understand where and how to act. Action Generation and Prediction branches develop techniques for translating multimodal inputs into executable motor commands, while Multimodal Integration and Perception addresses the fusion of vision, language, and sometimes tactile or proprioceptive signals. Training Paradigms and Data Utilization examines strategies for leveraging large-scale datasets and foundation models, Model Efficiency and Optimization tackles computational constraints, Benchmarking and Evaluation Frameworks provide standardized testbeds, and Specialized Applications and Task Domains target specific manipulation scenarios such as grasping or assembly.

Within this landscape, a particularly active line of work centers on directly incorporating spatial foundation models—such as depth estimators, segmentation networks, or geometric reasoning modules—into vision-language-action architectures. Papers like SpatialVLA[1], 3DS-VLA[2], and DepthVLA[28] exemplify efforts to enrich visual encoders with explicit 3D or depth cues, while GeoVLA[17] and GeoAware-VLA[34] emphasize geometric awareness for more precise spatial reasoning. Spatial to Actions[0] situates itself within this cluster by proposing a framework that systematically integrates spatial foundation priors into the action-generation pipeline, aiming to bridge the gap between high-level semantic understanding and low-level geometric control. Compared to works like Evo-0[3] or InternVLA[4], which prioritize scaling and generalization across diverse tasks, Spatial to Actions[0] places stronger emphasis on leveraging pre-trained spatial representations to improve manipulation accuracy and robustness in geometrically demanding scenarios.

Related Works in Same Category

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

1. VGGT-DP: Generalizable Robot Control via Vision Foundation Models

Authors: Ge Shijia, Zhang Yin-xin, Shijia Ge, Xie, Shuzhao, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Visual imitation learning frameworks allow robots to learn manipulation skills from expert demonstrations. While existing approaches mainly focus on policy design, they often neglect the structure and capacity of visual encoders, limiting spatial understanding and generalization. Inspired by biological vision systems, which rely on both visual and proprioceptive cues for robust control, we propose VGGT-DP, a visuomotor policy framework that integrates geometric priors from a pretrained 3D percep...

Relationship Analysis

Both papers belong to the Spatial Foundation Model Integration category, leveraging pretrained spatial or geometric foundation models to enhance VLAs with 3D spatial priors. They overlap in using spatial foundation models (FALCON uses VGGT for spatial tokens, VGGT-DP uses VGGT as visual encoder) to improve robotic manipulation through geometry-aware representations. However, FALCON focuses on integrating spatial tokens into a Spatial-Enhanced Action Head within a VLA framework with optional depth/pose fusion and modality transferability, while VGGT-DP emphasizes a language-free diffusion policy approach with frame-wise token reuse and proprioception-guided visual learning for efficient inference.

---

2. Spatial-Aware VLA Pretraining through Visual-Physical Alignment from Human Videos

Authors: Yicheng Feng, Wanpeng Zhang, Ye Wang, Hao Luo, Haoqi Yuan, et al. (7 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Vision-Language-Action (VLA) models provide a promising paradigm for robot learning by integrating visual perception with language-guided policy learning. However, most existing approaches rely on 2D visual inputs to perform actions in 3D physical environments, creating a significant gap between perception and action grounding. To bridge this gap, we propose a Spatial-Aware VLA Pretraining paradigm that performs explicit alignment between visual space and physical space during pretraining, enabl...

Relationship Analysis

Both papers belong to the Spatial Foundation Model Integration category, leveraging pretrained spatial or geometric foundation models to inject spatial priors into VLAs. They overlap in addressing the gap between 2D visual perception and 3D physical action by incorporating spatial foundation models (FALCON uses VGGT/ESM for spatial tokens, VIPA-VLA uses Cut3R for 3D visual encoding). The key difference is that FALCON focuses on injecting rich spatial tokens into a Spatial-Enhanced Action Head while preserving VLM alignment, whereas VIPA-VLA emphasizes a dual-encoder architecture with visual-physical alignment pretraining from large-scale human demonstration videos with 3D annotations.

Contributions Analysis

Overall novelty summary. The paper proposes FALCON, a paradigm that injects 3D spatial tokens from foundation models into the action head of vision-language-action models, aiming to bridge the spatial reasoning gap in existing 2D-encoder-based VLAs. Within the taxonomy, it resides in the 'Spatial Foundation Model Integration' leaf under 'Spatial Representation and Encoding Methods', alongside two sibling papers. This leaf represents a focused research direction within a broader taxonomy of 45 papers across multiple branches, suggesting a moderately active but not overcrowded subfield dedicated to leveraging pretrained spatial models for VLA enhancement.

The taxonomy reveals neighboring leaves addressing related spatial challenges: 'Explicit 3D Input Integration' (3 papers) handles depth sensors and point clouds, 'Implicit Spatial Understanding from 2D' (3 papers) learns geometry without explicit sensors, and 'Ego-Centric and Position Encoding' (2 papers) focuses on position-based representations. FALCON's approach diverges by emphasizing foundation model priors over sensor-specific architectures or learned-from-scratch encoders. The taxonomy's scope notes clarify that methods training spatial encoders from scratch or using only vision-language models belong elsewhere, positioning FALCON's foundation-model-centric design as a distinct strategy within the spatial encoding landscape.

Among 23 candidates examined, the core FALCON paradigm (Contribution 1) shows substantial prior work: 10 candidates examined, 5 potentially refutable. The Embodied Spatial Model (Contribution 2) appears more novel, with 6 candidates examined and none clearly refutable. The Spatial-Enhanced Action Head (Contribution 3) examined 7 candidates with 1 refutable. These statistics reflect a limited semantic search scope, not exhaustive coverage. The paradigm's core idea of injecting spatial tokens into action heads has recognizable precedents among the examined candidates, while the flexible modality integration mechanism appears less explored within this search window.

Given the limited search scope of 23 candidates, the analysis suggests FALCON operates in a moderately explored area where spatial foundation model integration is an active concern, but specific architectural choices around action-head injection and modality flexibility may offer incremental distinctions. The taxonomy context indicates this is one approach among several competing strategies for spatial enhancement, with the field still exploring optimal integration points and architectural patterns for combining geometric priors with vision-language reasoning.

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

Contribution 1: FALCON paradigm for injecting 3D spatial tokens into VLA action head

Description: The authors propose a new architecture that integrates spatial tokens from foundation models directly into the action prediction component rather than the vision-language backbone. This design preserves language reasoning while providing robust geometric priors from RGB inputs alone.

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

---

1. Spatialvla: Exploring spatial representations for visual-language-action model

URL: View paper

Prior Art Analysis

SpatialVLA[1] demonstrates prior work that injects spatial representations into the action prediction component rather than the vision-language backbone. The candidate paper explicitly describes integrating ego3d position encoding with adaptive action grids to represent spatial information for action generation, and emphasizes that spatial tokens are processed separately from the VLM backbone to preserve language reasoning capabilities. This architectural approach of decoupling spatial processing from the VLM while injecting it into the action head predates the original paper's FALCON paradigm.

Evidence

Evidence 1 - **Rationale**: Both papers describe augmenting VLM backbones with spatial representations specifically for action generation, demonstrating that the concept of injecting spatial information into the action component existed prior to FALCON. - **Original**: To overcome limitation (3) of alignment challenges, we draw inspiration from the brain's division of labor. the vlm (cerebrum) handles high-level reasoning and semantics, while the action head (cerebellum) manages fine-grained motor control and sensorimotor integration (rochefort et al., 2011; figur... - **Candidate**: spatialvla is developed based on a vision-language model to inherit the general world knowledge. formally, spatialvla takes image observations ot = {i1 t , ..,in t } and a natural language task instruction l as inputs, and then learns a mapping function τ(·) to generate a sequence of robot actions a...

Evidence 2 - **Rationale**: Both architectures separate spatial feature extraction from the VLM and integrate spatial information specifically at the action prediction stage, showing the same fundamental design pattern. - **Original**: as illustrated in fig. 2, falcon is an end-to-end vla consists of three core components: (1) a 2d vlm for multimodal semantic representation, (2) an esm for extracting 3d structural features, and (3) a spatial-enhanced action head that combines both streams to generate precise robot actions. - **Candidate**: the model comprises three key components: (1) siglip vision encoder extracts 2d semantic features, which are then infused with 3d spatial context via ego3d position encoding; (2) continuous 7d actions Δt, Δr, g are translated to 3 spatial action tokens by querying adaptive action grids and auto-regr...

---

2. 3ds-vla: A 3d spatial-aware vision language action model for robust multi-task manipulation

URL: View paper

Prior Art Analysis

3DS-VLA[2] demonstrates prior work that integrates 3D spatial information into vision-language-action models through a similar architectural approach. Both papers propose injecting spatial tokens into the action prediction component rather than the vision-language backbone. The candidate paper explicitly describes using a '2d-to-3d positional alignment mechanism' to enable 2D vision encoders to process 3D spatial observations, and then feeding these spatial representations to an action generation module. This architectural pattern—where spatial tokens are processed separately and then integrated at the action prediction stage—directly parallels the ORIGINAL paper's claimed novelty of injecting spatial tokens into the action head rather than the VLM backbone.

Evidence

Evidence 1 - **Rationale**: Both papers describe mechanisms for integrating 3D spatial information into 2D pretrained models while preserving the pretrained knowledge, with the candidate demonstrating this approach was already established. - **Original**: falcon adopts broader and richer tokens from these foundation models, and delivers comprehensive spatial information from rgb signals alone, improving robustness when explicit 3d inputs are absent... we design a spatial-enhanced action head that directly incorporates spatial tokens into action decis... - **Candidate**: we introduce a 2d-to-3d positional alignment mechanism that geometrically aligns each 3d token with the pretrained 2d positional embedding (pe) pointing to the same spatial region. Through this alignment, the 3d tokens are spatially encoded using pretrained 2d pes with corresponding spatial and sema...

Evidence 2 - **Rationale**: The candidate paper demonstrates a complete implementation of injecting spatial tokens into action prediction while preserving VLM integrity through parameter-efficient fine-tuning, showing this architectural pattern existed before the ORIGINAL paper's submission. - **Original**: This departs from prior approaches that forcibly align spatial and text tokens within vlms (fan et al., 2025; wu et al., 2025). in this way, falcon provides (i) robust spatial reasoning, (ii) strong modality transferability, and (iii) principled integration of 3d priors into vlas. - **Candidate**: as shown in fig. 2, we enable parameter-efficient fine-tuning (peft) to adapt a pretrained visionlanguage model (e.g., llama-adapter) into a policy model. The model π consists of a 2d visual encoder, llm (llama), a cross-modality projection module, and lora adapters... During imitation learning, we ...

---

3. PointVLA: Injecting the 3D World into Vision-Language-Action Models

URL: View paper

Prior Art Analysis

PointVLA[46] demonstrates prior work that injects 3D features into action prediction components of VLA models. Both papers propose architectures that integrate spatial information directly into the action generation pathway rather than the vision-language backbone. PointVLA[46] explicitly describes freezing the vanilla action expert and injecting 3D features via a lightweight modular block, which is conceptually similar to FALCON's approach of using a spatial-enhanced action head that consumes spatial tokens. The candidate paper's method of identifying less useful blocks for 3D feature injection and the original paper's design of a spatial-enhanced action head both aim to preserve pre-trained representations while adding geometric understanding to action prediction.

Evidence

Evidence 1 - **Rationale**: Both papers describe injecting 3D spatial information into the action prediction component while preserving pre-trained representations. PointVLA[46] explicitly freezes the action expert and uses modular blocks, while FALCON uses a spatial-enhanced action head, but both avoid disrupting the vision-language backbone. - **Original**: To overcome limitation (3) of alignment challenges, we draw inspiration from the brain's division of labor. the vlm (cerebrum) handles high-level reasoning and semantics, while the action head (cerebellum) manages fine-grained motor control and sensorimotor integration - **Candidate**: our method freezes the vanilla action expert and injects 3d features via a lightweight modular block. To identify the most effective way of integrating point cloud representations, we conduct a skip-block analysis to pinpoint less useful blocks in the vanilla action expert, ensuring that 3d features...

Evidence 2 - **Rationale**: Both approaches inject 3D information into the action prediction pathway rather than the VLM backbone. PointVLA[46]'s framework of enhancing pre-trained VLAs with 3D inputs via modular blocks demonstrates the same core architectural principle as FALCON's spatial-enhanced action head. - **Original**: we design a spatial-enhanced action head that directly incorporates spatial tokens into action decisions, which is a more natural fit since precise control depends on detailed spatial cues. This departs from prior approaches that forcibly align spatial and text tokens within vlms - **Candidate**: we propose pointvla, a framework that enhances pre-trained vlas with point cloud inputs without requiring retraining. our method freezes the vanilla action expert and injects 3d features via a lightweight modular block.

Evidence 3 - **Rationale**: Both papers describe lightweight mechanisms for integrating 3D spatial features into action prediction. PointVLA[46]'s lightweight modular block approach and FALCON's lightweight fusion mechanism in the action head represent similar architectural strategies for combining spatial and semantic information. - **Original**: The extracted semantic action token ˆtact and spatial

tokens tspl are then integrated in the spatialenhanced action head, collectively guide action generation. We introduce a lightweight fusion mechanism that aligns and combines these complementary representations - **Candidate**: our method freezes the vanilla action expert and injects 3d features via a lightweight modular block. To identify the most effective way of integrating point cloud representations, we conduct a skip-block analysis to pinpoint less useful blocks in the vanilla action expert

---

4. mindmap: Spatial Memory in Deep Feature Maps for 3D Action Policies

URL: View paper

Brief Assessment

mindmap[50] focuses on spatial memory through 3D reconstruction for manipulation policies, not on injecting spatial tokens from foundation models into VLA action heads. The architectural approaches differ fundamentally.

---

5. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Models

URL: View paper

Brief Assessment

SpatialVLA Representations[48] full text is not available (marked as 'n/a'), making comparison impossible. Without access to the candidate paper's methodology, we cannot assess whether it demonstrates prior work on injecting spatial tokens into action heads.

---

6. RetoVLA: Reusing Register Tokens for Spatial Reasoning in Vision-Language-Action Models

URL: View paper

Prior Art Analysis

RetoVLA[47] demonstrates prior work that injects spatial tokens directly into the action prediction component rather than the vision-language backbone. The candidate paper explicitly states that register tokens (containing spatial information) are 'injected into the action expert' to enhance spatial reasoning while maintaining a lightweight structure. This directly parallels the original paper's claim of injecting spatial tokens into the action head to preserve language reasoning. Both papers share the core architectural principle of decoupling spatial processing from the vision-language model and integrating it at the action prediction stage.

Evidence

Evidence 1 - **Rationale**: The candidate paper's approach of injecting spatial tokens into the action expert demonstrates the same design principle as the original paper's spatial-enhanced action head, both avoiding integration into the VLM backbone. - **Original**: we design a spatial-enhanced action head that directly incorporates spatial tokens into action decisions, which is a more natural fit since precise control depends on detailed spatial cues. This departs from prior approaches that forcibly align spatial and text tokens within vlms - **Candidate**: we suppose that these tokens contain essential spatial information and propose retovla, a novel architecture that reuses them directly by injecting them into the action expert.

Evidence 2 - **Rationale**: Both papers describe integrating spatial information at the action prediction stage to enhance spatial reasoning capabilities, demonstrating the same fundamental approach to spatial token injection. - **Original**: the proposed spatial-enhanced action head integrates geometric representations tspl from the esm with semantic features ˆtact from the vlm, enabling more accurate and spatially-aware policy learning. - **Candidate**: retovla maintains a lightweight structure while leveraging this repurposed spatial context to enhance reasoning.

---

7. Spatial Forcing: Implicit Spatial Representation Alignment for Vision-language-action Model

URL: View paper

Brief Assessment

Spatial Forcing[40] aligns intermediate visual embeddings with 3D foundation models rather than injecting spatial tokens into the action head. The architectural approach differs fundamentally from FALCON's spatial-enhanced action head design.

---

8. Improving vision-language-action models via chain-of-affordance

URL: View paper

Brief Assessment

Chain of Affordance[11] focuses on affordance-based reasoning (object, grasp, spatial, movement affordances) rather than 3D spatial token injection architectures. The candidate does not address the specific architectural innovation of injecting spatial foundation model tokens into the action head versus the VLM backbone.

---

9. Geovla: Empowering 3d representations in vision-language-action models

URL: View paper

Prior Art Analysis

GeoVLA[17] demonstrates prior work that injects 3D geometric information directly into the action head rather than the vision-language backbone. The candidate paper explicitly describes a parallel architecture where point cloud features are processed independently through a Point Embedding Network (PEN) and then integrated with vision-language features in a '3D-Enhanced Action Expert' (3DAE) module. This design preserves the pre-trained VLM alignment while incorporating spatial information at the action generation stage, which is the same core architectural principle claimed as novel in the FALCON paradigm.

Evidence

Evidence 1 - **Rationale**: Both papers describe injecting 3D spatial information into a specialized action head/expert rather than the VLM backbone, demonstrating the same architectural paradigm. - **Original**: we propose falcon (from spatial to action), a novel paradigm that integrates richer and more representative 3d spatial tokens into vlas through an improved injection scheme... we design a spatial-enhanced action head that directly incorporates spatial tokens into action decisions - **Candidate**: geovla incorporates a customized point encoder, point embedding network (pen), and a spatial-aware action expert, 3d-enhanced action expert (3dae), to bridge the gap between 2d and 3d modalities... Subsequently, these embeddings are concatenated and processed by our novel 3dae module

Evidence 2 - **Rationale**: Both papers explicitly motivate their approach by avoiding disruption to VLM alignment and instead injecting spatial information into the action generation component. - **Original**: to overcome limitation (3) of alignment challenges, we draw inspiration from the brain's division of labor... we design a spatial-enhanced action head that directly incorporates spatial tokens into action decisions, which is a more natural fit since precise control depends on detailed spatial cues. ... - **Candidate**: to preserve the alignment of vla models while incorporating 3d information, a complementary line of work explores injecting geometry into action heads rather than modifying visual backbones... our proposed method, geovla, distinguishes itself by employing a specialized point encoder (point embedding...

---

10. Toward Embodiment Equivariant Vision-Language-Action Policy

URL: View paper

Brief Assessment

Embodiment Equivariant[49] focuses on designing action spaces equivariant to embodiment configuration transformations for cross-embodiment generalization, not on injecting 3D spatial tokens from foundation models into action heads. The architectural concerns are fundamentally different.

Contribution 2: Embodied Spatial Model for flexible 3D modality integration

Description: The authors develop a spatial encoding module that can flexibly incorporate additional 3D inputs such as depth maps or camera poses when available, while maintaining strong performance with RGB-only input. This enables modality transferability without requiring model retraining.

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

---

1. A Spatial Pose Detection Method for Scrapers Based on Planar Vision and Laser Range Fusion

URL: View paper

Brief Assessment

Scraper Pose Detection[62] focuses on industrial pose detection using laser displacement sensors and planar vision for metal components, not on embodied AI models with flexible depth/pose fusion for robotic manipulation tasks.

---

2. Embodied VideoAgent: Persistent Memory from Egocentric Videos and Embodied Sensors Enables Dynamic Scene Understanding

URL: View paper

Brief Assessment

Embodied VideoAgent[61] focuses on constructing scene memory from egocentric video and embodied sensors for dynamic scene understanding, not on flexible depth/pose fusion in spatial encoding modules for robot manipulation policies.

---

3. Morefusion: Multi-object reasoning for 6d pose estimation from volumetric fusion

URL: View paper

Brief Assessment

MoreFusion[58] focuses on 6D pose estimation from volumetric fusion for multi-object reasoning in manipulation tasks, not on flexible depth/pose integration in embodied spatial models for vision-language-action frameworks without retraining.

---

4. Cross-Spatial Fusion and Dynamic-Range Particle Filter-Based FPGA-GPU Architecture for 1-ms RGB-Based Object Pose Tracking

URL: View paper

Brief Assessment

Cross-Spatial Fusion[57] focuses on real-time pose tracking in factory automation using FPGA-GPU architecture with cross-space fusion for depth enhancement, not on flexible modality integration in embodied spatial models for vision-language-action tasks without retraining.

---

5. Tracking and Planning with Spatial World Models

URL: View paper

Brief Assessment

Spatial World Models[59] focuses on differentiable rendering for navigation and tracking, not on flexible depth/pose fusion in embodied spatial models. The candidate uses TSDF fusion-based pose estimation for navigation tasks, which is architecturally distinct from the original paper's embodied spatial model that optionally integrates depth and camera poses for manipulation without retraining.

---

6. Joint estimation of depth and motion from a monocular endoscopy image sequence using a multi-loss rebalancing network.

URL: View paper

Brief Assessment

Multi-loss Rebalancing[60] focuses on joint depth and motion estimation from monocular endoscopy sequences for medical applications, not on flexible modality integration in embodied spatial models for robotic manipulation without retraining.

Contribution 3: Spatial-Enhanced Action Head for multimodal fusion

Description: The authors introduce a dedicated fusion mechanism that combines spatial tokens with semantic features at the action prediction stage. This approach avoids disrupting the pre-trained vision-language alignment while enabling precise spatial reasoning for robot control.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Spatial integration of multimodal brain images in cerebral infarction.

URL: View paper

Brief Assessment

Spatial Integration[56] focuses on multimodal brain image integration in cerebral infarction (medical imaging), not robotic action prediction or vision-language-action models. The domains are entirely different.

---

2. Transformer RGBT tracking with spatio-temporal multimodal tokens

URL: View paper

Brief Assessment

RGBT Tracking[51] focuses on fusing RGB and thermal infrared modalities for visual object tracking, not on spatial token fusion for robot action prediction. The candidate addresses appearance changes in tracking via spatio-temporal tokens, while the original contribution concerns spatial reasoning for robotic manipulation tasks.

---

3. Mutually beneficial transformer for multimodal data fusion

URL: View paper

Brief Assessment

Multimodal Transformer[52] focuses on HSI-LIDAR fusion for remote sensing classification using spatial constraints and channel diversity transformers, not robot action prediction or vision-language-action models.

---

4. Explainable Action Prediction through Self-Supervision on Scene Graphs

URL: View paper

Brief Assessment

Scene Graphs[55] focuses on scene graph representations for driver action prediction using graph neural networks with attention mechanisms, not on spatial token fusion for vision-language-action models in robotic manipulation.

---

5. Tmformer: Token merging transformer for brain tumor segmentation with missing modalities

URL: View paper

Brief Assessment

TMFormer[54] addresses brain tumor segmentation with missing MRI modalities using token merging strategies, not robot action prediction or vision-language-action models. The technical domains are entirely different.

---

6. Brain harmony: A multimodal foundation model unifying morphology and function into 1D tokens

URL: View paper

Brief Assessment

Brain Harmony[53] focuses on fusing brain morphology (T1-weighted MRI) and functional dynamics (fMRI) into 1D tokens for neuroscience applications, not robot manipulation. The fusion mechanism in Brain Harmony[53] uses learnable 1D brain hub tokens to reconstruct structural and functional latents, which is fundamentally different from the spatial-enhanced action head that combines spatial tokens with semantic features for robot control.

---

7. RetoVLA: Reusing Register Tokens for Spatial Reasoning in Vision-Language-Action Models

URL: View paper

Prior Art Analysis

RetoVLA[47] demonstrates prior work on fusing spatial tokens with semantic features at the action prediction stage. The candidate paper describes injecting register tokens (containing spatial information) into the action expert, which represents the same architectural principle of performing spatial-semantic fusion at the action prediction component rather than within the vision-language backbone. Both papers share the core design of avoiding disruption to the pre-trained vision-language alignment by performing fusion at the action stage.

Evidence

Evidence 1 - **Rationale**: The candidate's injection of spatial tokens into the action expert parallels the original's integration of spatial representations with semantic features at the action head, demonstrating the same fusion mechanism at the action prediction stage. - **Original**: the proposed spatial-enhanced action head integrates geometric representations tspl from the esm with semantic features ˆtact from the vlm, enabling more accurate and spatially-aware policy learning. - **Candidate**: we propose retovla, a novel architecture that reuses them directly by injecting them into the action expert. retovla maintains a lightweight structure while leveraging this repurposed spatial context to enhance reasoning.

Evidence 2 - **Rationale**: Both papers emphasize preserving the pre-trained VLM while adding spatial capabilities through action-stage fusion, demonstrating the same architectural principle of decoupled spatial processing. - **Original**: This dedicated fusion mechanism preserves the pre-trained representation space and generalizable capabilities of the vlm while enriching vla with geometrically grounded structural awareness. - **Candidate**: retovla maintains a lightweight structure while leveraging this repurposed spatial context to enhance reasoning.

Appendix: Text Similarity Detection

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

References

- [0] From Spatial to Actions: Grounding Vision-Language-Action Model in Spatial Foundation Priors View paper
- [1] Spatialvla: Exploring spatial representations for visual-language-action model View paper
- [2] 3ds-vla: A 3d spatial-aware vision language action model for robust multi-task manipulation View paper
- [3] Evo-0: Vision-language-action model with implicit spatial understanding View paper
- [4] Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy View paper
- [5] Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos View paper
- [6] Magma: A Foundation Model for Multimodal AI Agents View paper
- [7] A Survey on Reinforcement Learning of Vision-Language-Action Models for Robotic Manipulation View paper
- [8] Semanticvla: Semantic-aligned sparsification and enhancement for efficient robotic manipulation View paper
- [9] Mla: A multisensory language-action model for multimodal understanding and forecasting in robotic manipulation View paper
- [10] MoLe-VLA: Dynamic Layer-skipping Vision Language Action Model via Mixture-of-Layers for Efficient Robot Manipulation View paper
- [11] Improving vision-language-action models via chain-of-affordance View paper
- [12] VLABench: A Large-Scale Benchmark for Language-Conditioned Robotics Manipulation with Long-Horizon Reasoning Tasks View paper
- [13] TIGeR: Tool-Integrated Geometric Reasoning in Vision-Language Models for Robotics View paper
- [14] VisionCube: 3D-Aware Vision-Language Model for Multi-Step Spatial Reasoning View paper
- [15] Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints View paper
- [16] Lageo: a latent and geometrical framework for path and manipulation planning View paper
- [17] Geovla: Empowering 3d representations in vision-language-action models View paper
- [18] GraphCoT-VLA: A 3D Spatial-Aware Reasoning Vision-Language-Action Model for Robotic Manipulation with Ambiguous Instructions View paper
- [19] Era: Transforming vlms into embodied agents via embodied prior learning and online reinforcement learning View paper
- [20] Large language model-driven dynamic trajectory planning for human-guided robot assembly View paper
- [21] Integrating World Models into Vision Language Action and Navigation: A Comprehensive Survey View paper
- [22] Spatial policy: Guiding visuomotor robotic manipulation with spatial-aware modeling and reasoning View paper
- [23] EchoVLA: Robotic Vision-Language-Action Model with Synergistic Declarative Memory for Mobile Manipulation View paper
- [24] Vision-Language Models Enabled Robot Manipulation View paper
- [25] DreamVLA: A Vision-Language-Action Model Dreamed with Comprehensive World Knowledge View paper
- [26] RGMP: Recurrent Geometric-prior Multimodal Policy for Generalizable Humanoid Robot Manipulation View paper
- [27] LoLA: Long Horizon Latent Action Learning for General Robot Manipulation View paper

- [28] DepthVLA: Enhancing Vision-Language-Action Models with Depth-Aware Spatial Reasoning View paper
- [29] RoboGround: Robotic Manipulation with Grounded Vision-Language Priors View paper
- [30] VLA-4D: Embedding 4D Awareness into Vision-Language-Action Models for SpatioTemporally Coherent Robotic Manipulation View paper
- [31] Learning Spatial-Aware Manipulation Ordering View paper
- [32] GeoPredict: Leveraging Predictive Kinematics and 3D Gaussian Geometry for Precise VLA Manipulation View paper
- [33] Affordance Field Intervention: Enabling VLAs to Escape Memory Traps in Robotic Manipulation View paper
- [34] GeoAware-VLA: Implicit Geometry Aware Vision-Language-Action Model View paper
- [35] TrackVLA++: Unleashing Reasoning and Memory Capabilities in VLA Models for Embodied Visual Tracking View paper
- [36] Toward Accurate Long-Horizon Robotic Manipulation: Language-to-Action with Foundation Models via Scene Graphs View paper
- [37] VGGT-DP: Generalizable Robot Control via Vision Foundation Models View paper
- [38] RoboMamba: Efficient Vision-Language-Action Model for Robotic Reasoning and Manipulation View paper
- [39] SoraNav: Adaptive UAV Task-Centric Navigation via Zeroshot VLM Reasoning View paper
- [40] Spatial Forcing: Implicit Spatial Representation Alignment for Vision-language-action Model View paper
- [41] Uni-Sight: An E2E Vision-Language-Action System Unifying Multi-View Alignment and Multi-Modal Fusion View paper
- [42] Languageâ□□Guided Robot Grasping Based on Basic Geometric Shape Fitting View paper
- [43] Spatial-Aware VLA Pretraining through Visual-Physical Alignment from Human Videos View paper
- [44] SegGrasp: Zero-Shot Task-Oriented Grasping via Semantic and Geometric Guided Segmentation View paper
- [45] Autonomously Learning to Visually Detect Where Manipulation Will Succeed View paper
- [46] PointVLA: Injecting the 3D World into Vision-Language-Action Models View paper
- [47] RetoVLA: Reusing Register Tokens for Spatial Reasoning in Vision-Language-Action Models View paper
- [48] SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Models View paper
- [49] Toward Embodiment Equivariant Vision-Language-Action Policy View paper
- [50] mindmap: Spatial Memory in Deep Feature Maps for 3D Action Policies View paper
- [51] Transformer RGBT tracking with spatio-temporal multimodal tokens View paper
- [52] Mutually beneficial transformer for multimodal data fusion View paper
- [53] Brain harmony: A multimodal foundation model unifying morphology and function into 1D tokens View paper
- [54] Tmformer: Token merging transformer for brain tumor segmentation with missing modalities View paper
- [55] Explainable Action Prediction through Self-Supervision on Scene Graphs View paper
- [56] Spatial integration of multimodal brain images in cerebral infarction. View paper
- [57] Cross-Spatial Fusion and Dynamic-Range Particle Filter-Based FPGA-GPU Architecture for 1-ms RGB-Based Object Pose Tracking View paper
- [58] Morefusion: Multi-object reasoning for 6d pose estimation from volumetric fusion View paper
- [59] Tracking and Planning with Spatial World Models View paper
- [60] Joint estimation of depth and motion from a monocular endoscopy image sequence using a multi-loss rebalancing network. View paper
- [61] Embodied VideoAgent: Persistent Memory from Egocentric Videos and Embodied Sensors Enables Dynamic Scene Understanding View paper
- [62] A Spatial Pose Detection Method for Scrapers Based on Planar Vision and Laser Range Fusion View paper