Novelty Assessment Report

Paper: VLM4VLA: Revisiting Vision-Language-Models in Vision-Language-Action Models
PDF URL: https://openreview.net/pdf?id=tc2UsBeODW
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Vision-Language-Action (VLA) models, which integrate pretrained large Vision-Language Models (VLMs) into their policy backbone, are gaining significant attention for their promising generalization capabilities. This paper revisits a fundamental yet seldom systematically studied question: how the choice and specific capabilities of the underlying VLM affect the performance of VLA policies? We introduce \textbf{VLM4VLA}, a minimal adaptation pipeline that converts general-purpose VLMs into VLA policies using only a small set of new learnable parameters for fair and efficient comparison. Our pipeline, though simple, proves surprisingly competitive with more sophisticated network designs. Through extensive empirical studies on various downstream tasks across three benchmarks, we find that a VLM's general capabilities are poor predictors of its downstream task performance, contrary to common assumptions. Inconsistencies across benchmarks suggest that VLA policies require capabilities beyond what current VLMs pursue. We further investigate the impact of specific embodied capabilities by fine-tuning VLMs on seven auxiliary embodied tasks (e.g., embodied QA, visual pointing, depth estimation). Contrary to intuition, improving a VLM's performance on specific embodied skills does not guarantee better downstream control performance. Lastly, our analysis also reveals that the vision encoder is a critical bottleneck, and the ability to fine-tune it is crucial for strong performance. These results highlight a significant gap between current VLM pretraining paradigms and the specific demands of embodied tasks. We will release our code, models, and evaluation logs at \href{https://sites.google.com/view/vlm4vla}{our anonymous website} to encourage further research and help better understanding in this direction.

Core Task Landscape

This paper addresses: **Vision-Language Model Capabilities for Robotic Manipulation Tasks**

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

Taxonomy Overview

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

- **VLA Model Architectures and Design**
- **VLA Applications and Task-Specific Adaptations**
- **High-Level Planning and Reasoning**
- **Multimodal and Sensory Integration**
- **Data Generation and Pretraining**
- **Testing and Evaluation Frameworks**
- **Surveys and Reviews**
- **Specialized Applications and Domains**

Complete Taxonomy Tree

- Vision-Language Model Capabilities for Robotic Manipulation Tasks Survey Taxonomy
- VLA Model Architectures and Design
  - Compact and Efficient VLA Models (4 papers)
  - [1] Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation (Wen Junjie, 2025) View paper
  - [10] TinyVLA: Toward Fast, Data-Efficient Vision-Language-Action Models for Robotic Manipulation (Junjie Wen, 2024) View paper
  - [11] Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation (Pengju An, 2024) View paper
  - [33] Evo-1: Lightweight vision-language-action model with preserved semantic alignment (Lin Tao, 2025) View paper
  - Full-Scale VLA Architectures (5 papers)
  - [4] Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation (Li Qixiu, 2024) View paper
  - [16] Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation (Shi, 2025) View paper
  - [22] Mla: A multisensory language-action model for multimodal understanding and forecasting in robotic manipulation (Liu, 2025) View paper
  - [30] Chatvla: Unified multimodal understanding and robot control with vision-language-action model (Zhou Zhongyi, 2025) View paper
  - [34] Instructvla: Vision-language-action instruction tuning from understanding to manipulation (Yang, 2025) View paper
  - VLM-to-VLA Adaptation Frameworks ★ (3 papers)
  - [0] VLM4VLA: Revisiting Vision-Language-Models in Vision-Language-Action Models (Anon et al., 2026) View paper
  - [9] Robouniview: Visual-language model with unified view representation for robotic manipulation (Liu Fan-fan, 2024) View paper
  - [12] Vision-language foundation models as effective robot imitators (LI Xinghang, 2023) View paper
  - Federated and Distributed VLA Learning (1 papers)
  - [27] FedVLA: Federated Vision-Language-Action Learning with Dual Gating Mixture-of-Experts for Robotic Manipulation (Miao Cui, 2025) View paper

- VLA Applications and Task-Specific Adaptations
  - Bimanual and Dexterous Manipulation (2 papers)
  - [23] Bi-VLA: Vision-Language-Action Model-Based System for Bimanual Robotic Dexterous Manipulations (Koffivi FidÃ¨le Gbagbe, 2024) View paper
  - [25] Robodexvlm: Visual language model-enabled task planning and motion control for dexterous robot manipulation (Liu Hai-Chao, 2025) View paper
  - Spatial Affordance and Keypoint Prediction (4 papers)
  - [5] Robopoint: A vision-language model for spatial affordance prediction for robotics (Yuan, 2024) View paper
  - [29] Object-Centric Prompt-Driven Vision-Language-Action Model for Robotic Manipulation (Xiao-qi Li, 2025) View paper
  - [35] CrayonRobo: Object-Centric Prompt-Driven Vision-Language-Action Model for Robotic Manipulation (Li Xiaoqi, 2025) View paper
  - [39] SKT: Integrating State-Aware Keypoint Trajectories with Vision-Language Models for Robotic Garment Manipulation (LI Xin, 2024) View paper
  - Grasping with Vision-Language Models (3 papers)
  - [36] Language reasoning in vision-language-action model for robotic grasping (Lingling Fan, 2024) View paper
  - [42] Integrating With Multimodal Information for Enhancing Robotic Grasping With Vision-Language Models (Zhou Zhao, 2025) View paper
  - [50] A Vision-Language Model Approach for Object Segmentation and Robotic Grasping (Mirco Polonara, 2025) View paper
  - General Manipulation Task Execution (4 papers)
  - [7] Vision-language model-driven scene understanding and robotic object manipulation (Sichao Liu, 2024) View paper
  - [20] ManipLLM: Embodied Multimodal Large Language Model for Object-Centric Robotic Manipulation (Xiao-qi Li, 2023) View paper
  - [31] Vision-Language Models Enabled Robot Manipulation (Li, 2025) View paper
  - [48] GPTArm: An Autonomous Task Planning Manipulator Grasping System Based on Visionâ□□Language Models (Jiaqi Zhang, 2025) View paper
- High-Level Planning and Reasoning
  - Multi-Stage and Long-Horizon Planning (2 papers)
  - [15] Reflective Planning: Vision-Language Models for Multi-Stage Long-Horizon Robotic Manipulation (Feng Yunhai, 2025) View paper
  - [32] GraphCoT-VLA: A 3D Spatial-Aware Reasoning Vision-Language-Action Model for Robotic Manipulation with Ambiguous Instructions (Huang He-long, 2025) View paper
  - Failure Detection and Error Recovery (2 papers)
  - [19] AHA: A Vision-Language-Model for Detecting and Reasoning Over Failures in Robotic Manipulation (Duan, 2024) View paper
  - [41] Interactive robot action replanning using multimodal llm trained from human demonstration videos (Chiori Hori, 2025) View paper
  - Model Predictive Control with VLMs (1 papers)
  - [49] VLMPC: Vision-Language Model Predictive Control for Robotic Manipulation (Zhao Wentao, 2024) View paper
- Multimodal and Sensory Integration
  - Speech and Audio Integration (1 papers)
  - [14] Vlas: Vision-language-action model with speech instructions for customized robot manipulation (Zhao Wei, 2025) View paper
  - Tactile and Multisensory Perception (1 papers)
  - [46] Multi-Modal Perception With Vision, Language, and Touch for Robot Manipulation (Huang, 2025) View paper
  - 3D Spatial and Depth Reasoning (1 papers)
  - [6] Multimodal spatial language maps for robot navigation and manipulation (Huang Chenguang, 2025) View paper
- Data Generation and Pretraining
  - Human Activity Video Pretraining (1 papers)
  - [17] Scalable Vision-Language-Action Model Pretraining for Robotic Manipulation with Real-Life Human Activity Videos (Li Qixiu, 2025) View paper
  - Synthetic and Automated Data Generation (2 papers)
  - [40] Manipulate-anything: Automating real-world robots using vision-language models (Duan, 2024) View paper
  - [47] KALIE: Fine-Tuning Vision-Language Models for Open-World Manipulation Without Robot Data (Grace Tang, 2025) View paper
- Testing and Evaluation Frameworks (2 papers)
  - [37] LADEV: A Language-Driven Testing and Evaluation Platform for Vision-Language-Action Models in Robotic Manipulation (Wang Zhijie, 2024) View paper
  - [38] Vision-Language Models for Robot Success Detection (Fiona Luo, 2024) View paper
- Surveys and Reviews (9 papers)
  - [2] Vision-language-action models for robotics: A review towards real-world applications (Kento KAWAHARAZUKA, 2025) View paper
  - [3] Vision language action models in robotic manipulation: A systematic review (Akram, 2025) View paper
  - [8] Recipe for Vision-Language-Action Models in Robotic Manipulation: A Survey (Tomohiro Motoda, 2025) View paper
  - [18] Language-conditioned learning for robotic manipulation: A survey (Hongkuan Zhou, 2023) View paper
  - [21] A Review of Advances in Large Language and Vision Models for Robotic Manipulation: Techniques, Integrations, and Challenges (Sajjad Hussain, 2025) View paper
  - [24] Foundation models in robotics: Applications, challenges, and the future (Roya Firoozi, 2025) View paper
  - [28] Multimodal fusion and vision-language models: A survey for robot vision (Xiaofeng Han, 2025) View paper
  - [43] Embodied AI with Foundation Models for Mobile Service Robots: A Systematic Review (Benhabib, 2025) View paper
  - [44] Large VLM-based Vision-Language-Action Models for Robotic Manipulation: A Survey (Shao Rui, 2025) View paper
- Specialized Applications and Domains (3 papers)
  - [13] Visual language models agent applications in robotic manipulation tasks (Zhao, 2025) View paper
  - [26] Vision-Language Models in Industrial Robotics (NyyssÃ¶nen, 2024) View paper
  - [45] Vima: Robot manipulation with multimodal prompts (Y Jiang, 2023) View paper

Narrative

Core task: vision-language model capabilities for robotic manipulation tasks. The field has organized itself around several major branches that reflect different stages and aspects of building vision-language action (VLA) systems. VLA Model Architectures and Design focuses on how to construct or adapt pretrained vision-language models into action-generating policies, including frameworks that bridge VLMs to VLAs and efficient architectures like TinyVLA[1] or RoboMamba[11]. High-Level Planning and Reasoning addresses how these models can perform task decomposition and multi-step reasoning, while Multimodal and Sensory Integration explores incorporating additional modalities such as tactile or audio signals. Data Generation and Pretraining examines scalable data collection and pretraining strategies, and Testing and Evaluation Frameworks provides benchmarks and metrics for assessing VLA performance. Surveys and Reviews, including VLA Systematic Review[3] and VLA Recipe Survey[8], synthesize emerging best practices, while Specialized Applications and Domains target specific use cases such as industrial robotics or garment manipulation.

Within this landscape, a particularly active line of work centers on VLM-to-VLA adaptation frameworks, which seek efficient pathways to convert large pretrained vision-language models into robotic controllers without prohibitive retraining costs. VLM4VLA[0] sits squarely in this branch, proposing methods to leverage existing VLM representations for action prediction. Nearby efforts like RoboUniView[9] and VLM Robot Imitators[12] similarly explore how to repurpose VLM encoders or align language-conditioned features with low-level control, though they may differ in whether they emphasize unified architectures or imitation-based fine-tuning. A contrasting theme appears in works that prioritize efficiency and deployment constraints, such as TinyVLA[1] and TinyVLA Fast Efficient[10], which focus on model compression and real-time inference. The central tension across these directions involves balancing the rich semantic understanding of large VLMs against the need for sample-efficient adaptation, computational feasibility, and robust generalization to novel manipulation scenarios.

Related Works in Same Category

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

1. Robouniview: Visual-language model with unified view representation for robotic manipulation

Authors: Liu Fan-fan, Yan Feng, Zheng Liming, Feng, Chengjian, et al. (7 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Utilizing Vision-Language Models (VLMs) for robotic manipulation represents a novel paradigm, aiming to enhance the model's ability to generalize to new objects and instructions. However, due to variations in camera specifications and mounting positions, existing methods exhibit significant performance disparities across different robotic platforms. To address this challenge, we propose RoboUniView in this paper, an innovative approach that decouples visual feature extraction from action learnin...

Relationship Analysis

Both papers belong to the VLM-to-VLA Adaptation Frameworks category, focusing on methods for converting general-purpose VLMs into VLA policies for robotic control. While the original paper (VLM4VLA) systematically evaluates how different VLM backbones and their capabilities affect downstream VLA performance using a minimal adaptation pipeline with learnable action queries and MLP heads, the candidate paper (RoboUniView) proposes a specific adaptation approach that decouples visual feature extraction from action learning by introducing a unified view representation (UVFormer) pre-trained on 3D occupancy tasks to handle multi-camera perspectives and achieve camera-parameter generalization.

---

2. Vision-language foundation models as effective robot imitators

Authors: LI Xinghang, Xinghang Li, Liu, Minghuan, Minghuan Liu, et al. (29 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Recent progress in vision language foundation models has shown their ability to understand multimodal data and resolve complicated vision language tasks, including robotics manipulation. We seek a straightforward way of making use of existing vision-language models (VLMs) with simple fine-tuning on robotics data. To this end, we derive a simple and novel vision-language manipulation framework, dubbed RoboFlamingo, built upon the open-source VLMs, OpenFlamingo. Unlike prior works, RoboFlamingo ut...

Relationship Analysis

Both papers belong to the VLM-to-VLA Adaptation Frameworks category, focusing on methods for converting general-purpose VLMs into robotic control policies. They share overlapping approaches in fine-tuning VLMs on manipulation datasets and evaluating on similar benchmarks (CALVIN). The key difference is that the original paper (VLM4VLA) provides a systematic comparative study across 17 different VLMs with minimal architectural additions (<1% parameters) to isolate VLM capabilities, while the candidate paper (RoboFlamingo) proposes a specific adaptation framework built on OpenFlamingo with an explicit policy head (LSTM/Transformer) for sequential modeling and focuses on demonstrating competitive performance rather than systematic VLM comparison.

Contributions Analysis

Overall novelty summary. The paper introduces VLM4VLA, a minimal adaptation pipeline that converts general-purpose vision-language models into vision-language-action policies using a small set of learnable parameters. It resides in the VLM-to-VLA Adaptation Frameworks leaf, which contains only three papers total including this work. This is a relatively sparse research direction within the broader taxonomy of 50 papers across 19 leaf nodes, suggesting the specific question of systematic VLM-to-VLA conversion remains underexplored compared to more crowded areas like full-scale VLA architectures or general manipulation task execution.

The taxonomy tree reveals that VLM-to-VLA adaptation sits within the larger VLA Model Architectures and Design branch, which also includes compact/efficient models and full-scale architectures. Neighboring leaves address efficiency-focused designs like TinyVLA and advanced reasoning systems with memory or multimodal integration. The scope note for this leaf explicitly excludes end-to-end VLA designs, positioning the work as a bridge between pretrained VLMs and robotic control rather than a novel architecture from scratch. Related directions in data generation and evaluation frameworks provide complementary infrastructure, but the core adaptation methodology remains distinct.

Among 30 candidates examined, the minimal adaptation pipeline contribution shows overlap with 3 out of 10 candidates reviewed, while the systematic empirical study of VLM capabilities found no clear refutations across 10 candidates. The analysis of vision encoders as bottlenecks encountered 3 potentially overlapping works among 10 examined. The empirical study component appears more novel within this limited search scope, whereas the adaptation pipeline and encoder analysis face more substantial prior work. These statistics reflect top-K semantic matches and citation expansion, not exhaustive coverage of the field.

Based on the limited search scope of 30 candidates, the work's novelty appears mixed: the systematic empirical investigation of how VLM capabilities transfer to embodied control seems less explored, while the minimal adaptation approach and encoder bottleneck analysis encounter more prior work. The sparse population of the VLM-to-VLA adaptation leaf suggests room for contributions, though the analysis cannot rule out relevant work outside the top-30 semantic matches examined.

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

Contribution 1: VLM4VLA minimal adaptation pipeline

Description: The authors propose a lightweight framework that adapts Vision-Language Models into Vision-Language-Action policies by adding fewer than 1% new parameters. This design enables fair comparison across different VLMs while maintaining competitive performance with more sophisticated architectures.

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. What Matters in Employing Vision Language Models for Tokenizing Actions in Robot Control?

URL: View paper

Brief Assessment

Tokenizing Actions[54] focuses on action tokenization methods and training different VLM architectures for robot control, but does not propose a minimal adaptation pipeline with <1% new parameters as a core contribution. The candidate explores which VLM components matter for action tokenization, while the original paper's contribution centers on a lightweight framework design for fair VLM comparison.

---

2. Vla-adapter: An effective paradigm for tiny-scale vision-language-action model

URL: View paper

Prior Art Analysis

VLA-Adapter[56] demonstrates that a similar minimal adaptation approach was developed independently. Both papers propose lightweight frameworks that adapt VLMs to VLA policies with minimal additional parameters (VLM4VLA uses <1% new parameters, VLA-Adapter uses a 97M policy network with 0.5B backbone). Both employ learnable action query tokens and simple MLP-based policy heads to avoid complex architectures. The candidate paper explicitly states their method 'introduces vla-adapter, a novel paradigm designed to reduce the reliance of vla models on large-scale vlms and extensive pre-training' and achieves 'high performance using only a 0.5b-parameter backbone, without any robotic data pre-training,' which directly parallels the original paper's claim of enabling 'fair comparison across different VLMs while maintaining competitive performance.'

Evidence

Evidence 1 - **Rationale**: Both papers propose minimal adaptation pipelines that convert VLMs to VLA policies with small parameter additions and without requiring robotic pre-training, demonstrating similar core contributions. - **Original**: we introduce vlm4vla, a minimal adaptation pipeline that converts general-purpose vlms into vla policies using only a small set of new learnable parameters for fair and efficient comparison. our pipeline, though simple, proves surprisingly competitive with more sophisticated network designs. - **Candidate**: we introduce vla-adapter, a novel paradigm designed to reduce the reliance of vla models on large-scale vlms and extensive pre-training. to this end, we first systematically analyze the effectiveness of various vl conditions and present key findings on which conditions are essential for bridging per...

Evidence 2 - **Rationale**: Both papers emphasize minimal parameter additions (VLM4VLA <1%, VLA-Adapter 97M with 0.5B backbone) and use simple architectures (MLP heads) to ensure stability and reduce complexity. - **Original**: vlm4vla is a carefully designed network plug-in, introducing fewer than 1% new parameters. to enhance the stability of inference and the robustness of evaluation, we use a simple mlp head rather than a diffusion-based approach, thus controlling stochasticity and reducing tuning complexity. - **Candidate**: the policy parameters are only 97m when the backbone is qwen2.5-0.5b. each-layerc r t andc aq t are integrated in bridge attention with the corresponding-layer action latent. bridge attention maps vl to action to the greatest extent. the degree ofc r t injection is learnable, ensuring the performanc...

Evidence 3 - **Rationale**: Both papers use learnable action query tokens as the core mechanism to extract and decode action-relevant information from VLMs, demonstrating the same fundamental architectural approach. - **Original**: we introduce a learnable action query token to extract embodiment-related knowledge from the vlm. the representation of this token is then decoded into an action chunk. to align with the pre-training input format of each model, we adapt a unique token concatenation scheme for each vlm4vla instance. - **Candidate**: at timestep, the input into vlm consists of{x v t ,x g t ,lt,aqt}: the 3rd-view imagex v t , the gripper imagex g t , the instructionl t, and additional actionqueryaq t. after inputtingx v t andx g t , the dinov2 (oquab et al., 2024) and siglip (zhai et al., 2023) extract vision embeddings.l t is t...

---

3. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation

URL: View paper

Brief Assessment

RoboMamba[11] focuses on integrating Mamba (a state space model) with vision encoders for robotic manipulation, using a different architectural approach than the minimal VLM adaptation pipeline described in the original paper. The candidate does not challenge the novelty of adapting VLMs with fewer than 1% new parameters.

---

4. A survey on efficient vision-language-action models

URL: View paper

Brief Assessment

Efficient VLA Survey[55] is a survey paper that reviews existing efficient VLA methods but does not propose a specific minimal adaptation pipeline itself. The original paper's VLM4VLA framework is a concrete implementation with <1% new parameters, while the survey categorizes various efficiency approaches without claiming priority on this specific design.

---

5. 3ds-vla: A 3d spatial-aware vision language action model for robust multi-task manipulation

URL: View paper

Brief Assessment

3DS-VLA[53] focuses on enhancing 2D VLMs with 3D spatial awareness for robotic manipulation, using a different architectural approach (2D-to-3D positional alignment, 3D tokenizers) rather than the minimal parameter adaptation pipeline described in the original paper. The candidate addresses 3D spatial reasoning challenges, not the systematic comparison of VLM capabilities with minimal architectural changes.

---

6. Large VLM-based Vision-Language-Action Models for Robotic Manipulation: A Survey

URL: View paper

Brief Assessment

Large VLM Survey[44] discusses various VLA architectures but does not present a specific minimal adaptation pipeline with fewer than 1% new parameters as a novel contribution. The survey categorizes existing methods rather than proposing new frameworks.

---

7. Vision-language foundation models as effective robot imitators

URL: View paper

Prior Art Analysis

VLM Robot Imitators[12] demonstrates that a similar minimal adaptation approach was proposed earlier. Both papers introduce lightweight frameworks that adapt Vision-Language Models for robotic control with minimal additional parameters. VLM Robot Imitators[12] explicitly describes adding a simple policy head with fewer trainable parameters to convert VLMs into manipulation policies, using only imitation learning on robotics data. The architectural decomposition and training approach are nearly identical to the ORIGINAL paper's claimed contribution.

Evidence

Evidence 1 - **Rationale**: Both papers describe minimal adaptation frameworks that convert VLMs into manipulation policies using simple additional components and imitation learning only on robotics data, demonstrating that VLM Robot Imitators[12] proposed this approach earlier. - **Original**: we introduce vlm4vla, a minimal adaptation pipeline that converts general-purpose vlms into vla policies using only a small set of new learnable parameters for fair and efficient comparison. our pipeline, though simple, proves surprisingly competitive with more sophisticated network designs. - **Candidate**: we derive a simple and novel vision-language manipulation framework, dubbed roboflamingo, built upon the open-source vlms, openflamingo. unlike prior works,roboflamingo utilizes pre-trained vlms for single-step visionlanguage comprehension, models sequential history information with an explicit poli...

Evidence 2 - **Rationale**: Both papers emphasize adding minimal trainable parameters through a simple policy head to adapt VLMs for robotic control, showing VLM Robot Imitators[12] established this minimal adaptation approach first. - **Original**: vlm4vla is a carefully designed network plug-in, introducing fewer than 1% new parameters. to enhance the stability of inference and the robustness of evaluation, we use a simple mlp head rather than a diffusion-based approach, thus controlling stochasticity and reducing tuning complexity. - **Candidate**: to adapt large-scale vision-language models to robotic manipulation, roboflamingo simply adds a policy head for end-to-end finetuning. it addresses three main challenges: 1) it adapts vision-language models with static image inputs to video observations; 2) it generates robot control signals instead...

Evidence 3 - **Rationale**: Both papers describe training VLMs directly on robotics data with minimal architectural additions, achieving competitive performance despite simplicity, confirming VLM Robot Imitators[12] demonstrated this approach earlier. - **Original**: this network allows us to train the modified vlms directly using downstream robot data, facilitating alignment between the vlm's capabilities and the demands of robotic tasks. despite its simplicity, vlm4vla proves effective, demonstrating competitive performance against more advanced network design... - **Candidate**: roboflamingo is grounded upon the open-source vlm, openflamingo (awadalla et al., 2023), and resolves the challenge by decoupling visual-language understanding and decision-making. unlike previous works, roboflamingo takes advantage of pre-trained vlms mainly for understanding vision observations an...

---

8. Smolvla: A vision-language-action model for affordable and efficient robotics

URL: View paper

Prior Art Analysis

SmolVLA[52] demonstrates prior work on minimal-parameter adaptation of VLMs for robotic control. Both papers propose lightweight frameworks that add minimal new parameters to pretrained VLMs for action prediction. SmolVLA[52] explicitly states it 'introduces fewer than 1% new parameters' through its action expert design, directly paralleling the ORIGINAL paper's claim of 'fewer than 1% new parameters.' Both use learnable query tokens and small MLPs for action decoding. The architectural similarities and shared design philosophy of minimal adaptation suggest the ORIGINAL paper's novelty claim regarding this lightweight framework approach can be refuted by SmolVLA[52]'s prior publication.

Evidence

Evidence 1 - **Rationale**: Both papers emphasize creating efficient, minimal adaptation frameworks for converting VLMs into VLAs with focus on computational efficiency and accessibility. - **Original**: we introduce vlm4vla, a minimal adaptation pipeline that converts general-purpose vlms into vla policies using only a small set of new learnable parameters for fair and efficient comparison. - **Candidate**: we present smolvla, a small, efficient, and community-driven vla that drastically reduces both training and inference costs, while retaining competitive performance. smolvla is designed to be trained on a single gpu and deployed on consumer-grade gpus or even cpus.

Evidence 2 - **Rationale**: Both describe minimal architectural additions to pretrained VLMs. SmolVLA's action expert represents a small addition to the base VLM, similar to VLM4VLA's approach. - **Original**: vlm4vla is a carefully designed network plug-in, introducing fewer than 1% new parameters. - **Candidate**: smolvla consists of a compact pretrained vision-language model, discarding the lastl -n layers (scissors icon). the remaining layers embed three inputs: (i) language instruction, (ii) rgb image(s), and (iii) robot sensorimotor state. their merged tokens feed an action expert of alternating cross-att...

Evidence 3 - **Rationale**: Both papers use small additional modules (learnable tokens + MLP in ORIGINAL; action expert in SmolVLA) to decode VLM features into action chunks, demonstrating the same minimal adaptation strategy. - **Original**: we introduce a learnable action query token to extract embodiment-related knowledge from the vlm. the representation of this token is then decoded into an action chunk using a small mlp-based policy head. - **Candidate**: the action expertvθ is trained to predict an action chunkat =(at, . . . , at+n) from vlm features. in keeping with prior work, our implementation ofvθ relies on the transformer architecture

Evidence 4 - **Rationale**: Both papers describe using simple projection layers as part of their minimal adaptation approach, showing similar architectural design principles for efficient VLM-to-VLA conversion. - **Original**: during training, we finetune all parameters of the vlm, including the llm, the vision encoder, and the word embeddings. - **Candidate**: we use linear projection layers in various points inside of smolvla. in particular, we use linear projection layers to(i) project the states to match the vlm dimension(ii) project the actions to match the action expert dimensions and(iii) to adapt the vlm features to align with the action expert's d...

---

9. BitVLA: 1-bit Vision-Language-Action Models for Robotics Manipulation

URL: View paper

Brief Assessment

BitVLA[51] focuses on 1-bit quantization of VLA models for memory efficiency, not on minimal-parameter adaptation pipelines for fair VLM comparison. The architectural approaches are fundamentally different.

---

10. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation

URL: View paper

Brief Assessment

TinyVLA[1] focuses on combining lightweight VLMs with diffusion policy heads for fast inference in robotic manipulation, rather than providing a minimal adaptation framework for fair comparison across different VLMs. The architectural approaches and research objectives differ fundamentally.

Contribution 2: Systematic empirical study of VLM capabilities for embodied control

Description: The authors conduct large-scale experiments evaluating 17 VLMs across three benchmarks (Calvin, SimplerEnv, Libero) to investigate how general VLM capabilities, embodied-specific fine-tuning, and vision encoder training strategies affect downstream manipulation performance. They reveal inconsistencies and gaps between VLM pretraining paradigms and embodied task demands.

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. ManipLLM: Embodied Multimodal Large Language Model for Object-Centric Robotic Manipulation

URL: View paper

Brief Assessment

ManipLLM[20] focuses on developing a specific manipulation system using MLLMs for object-centric robotic tasks, rather than conducting a systematic empirical study evaluating multiple VLMs across benchmarks to understand how VLM capabilities transfer to embodied control.

---

2. Embodiedgpt: Vision-language pre-training via embodied chain of thought

URL: View paper

Brief Assessment

EmbodiedGPT[67] focuses on building an end-to-end multi-modal foundation model for embodied AI with chain-of-thought planning capabilities, rather than conducting a systematic empirical study comparing multiple VLMs across benchmarks for manipulation performance.

---

3. Momanipvla: Transferring vision-language-action models for general mobile manipulation

URL: View paper

Brief Assessment

MoManipVLA[69] focuses on transferring fixed-base VLA models to mobile manipulation through motion planning, not on systematically evaluating VLM capabilities across benchmarks for embodied control.

---

4. Large VLM-based Vision-Language-Action Models for Robotic Manipulation: A Survey

URL: View paper

Brief Assessment

Large VLM Survey[44] provides a comprehensive review of existing VLA models and their characteristics, but does not conduct original empirical experiments evaluating 17 VLMs across benchmarks. It synthesizes prior work rather than presenting new experimental findings.

---

5. Survey of vision-language-action models for embodied manipulation

URL: View paper

Brief Assessment

VLA Embodied Survey[66] is a comprehensive survey paper that reviews existing VLA models and their evaluation across multiple dimensions. It does not present original empirical experiments comparing 17 VLMs across benchmarks, but rather synthesizes existing research. The original paper conducts novel large-scale experiments with a unified VLM4VLA pipeline to systematically evaluate how VLM capabilities transfer to embodied tasks.

---

6. Robopoint: A vision-language model for spatial affordance prediction for robotics

URL: View paper

Brief Assessment

RoboPoint[5] focuses on spatial affordance prediction via point-based action spaces for robotics, not on systematic evaluation of multiple VLMs across embodied manipulation benchmarks. The candidate does not challenge the novelty of conducting large-scale comparative studies of VLM architectures for embodied control tasks.

---

7. VLP: Vision-Language Preference Learning for Embodied Manipulation

URL: View paper

Brief Assessment

VLP Preference Learning[68] focuses on learning preference models from vision-language alignment for reward learning in manipulation tasks, not on systematically evaluating how different VLM capabilities affect downstream control performance across multiple benchmarks.

---

8. Physically grounded vision-language models for robotic manipulation

URL: View paper

Brief Assessment

Physically Grounded VLM[65] focuses on physical concept reasoning (material, fragility) for robotic manipulation, not on systematic evaluation of VLM architectures across manipulation benchmarks like the original paper.

---

9. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation

URL: View paper

Brief Assessment

TinyVLA[1] does not conduct a systematic study evaluating multiple VLMs across benchmarks to investigate how VLM capabilities affect manipulation performance. Instead, it proposes a specific architecture combining small VLMs with diffusion models and evaluates this approach.

---

10. A survey on vision-language-action models for embodied ai

URL: View paper

Brief Assessment

VLA Embodied AI Survey[70] is a comprehensive survey paper that reviews existing VLA models and their components, rather than conducting original empirical experiments evaluating VLM capabilities for manipulation tasks. The survey discusses various VLA approaches but does not present systematic experimental comparisons of VLMs across benchmarks like the original paper does.

Contribution 3: Analysis of vision encoder as critical bottleneck

Description: The authors identify through ablation studies that fine-tuning the vision encoder is essential for VLA performance, showing significant degradation when frozen. This finding highlights the importance of visual adaptation over simply scaling language model parameters for embodied tasks.

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. The future of action recognition: are multi-modal visual language models the key?

URL: View paper

Brief Assessment

Action Recognition Future[60] focuses on action recognition tasks using multi-modal visual language models, not on vision-language-action policies for robotic manipulation. The technical domains and applications are fundamentally different.

---

2. Openvla: An open-source vision-language-action model

URL: View paper

Prior Art Analysis

OpenVLA[57] demonstrates through ablation studies that fine-tuning the vision encoder is essential for VLA performance, showing significant degradation when frozen. The paper explicitly states that 'we found fine-tuning the vision encoder during vla training to be crucial for good vla performance' and provides empirical evidence showing substantial performance drops when the vision encoder is frozen. This directly addresses the same research question about the importance of vision encoder fine-tuning in vision-language-action models, demonstrating that this finding was established prior to or contemporaneously with the original paper's contribution.

Evidence

Evidence 1 - **Rationale**: Both papers identify the vision encoder as a critical component requiring fine-tuning. OpenVLA[57] explicitly states this finding in their design decisions section, demonstrating prior work on this exact contribution. - **Original**: lastly, our analysis also reveals that the vision encoder is a critical bottleneck, and the ability to fine-tune it is crucial for strong performance. - **Candidate**: fine-tuning vision encoder. prior work on vlms found that freezing vision encoders during vlm training typically leads to higher performance [44]. intuitively, a frozen vision encoder may better preserve the robust features learned from its internet-scale pretraining. however, we found fine-tuning t...

Evidence 2 - **Rationale**: Both papers conduct ablation studies comparing frozen versus fine-tuned vision encoders. OpenVLA[57] systematically evaluates this design choice and reports performance differences. - **Original**: table 3 shows the performance of three models when the vision encoder is frozen during vlm4vla training. we observe a significant performance degradation for all models on both the calvin and simpler benchmarks after freezing the vision encoder - **Candidate**: we compare the following fine-tuning approaches: full finetuning updates all weights during finetuning, as described in section 5.2; last layer only fine-tunes only the last layer of openvla's transformer backbone and the token embedding matrix; frozen vision freezes the vision encoder but fine-tune...

Evidence 3 - **Rationale**: Both papers reach the same conclusion that freezing the vision encoder leads to poor performance and that fine-tuning it is crucial for VLA adaptation, with OpenVLA[57] providing empirical evidence supporting this finding. - **Original**: this finding strongly suggests that finetuning the vision encoder is crucial when adapting a vlm into a vla, and that the impact of this module can be more significant than merely increasing the number of trainable parameters in the language model. - **Candidate**: we find that only fine-tuning the network's last layer or freezing the vision encoder leads to poor performance, suggesting that further adaptation of the visual features to the target scene is crucial.

---

3. Coa-vla: Improving vision-language-action models via visual-text chain-of-affordance

URL: View paper

Brief Assessment

COA-VLA[63] focuses on affordance-based reasoning for VLA models rather than analyzing vision encoder fine-tuning importance. The paper does not investigate or discuss the vision encoder as a bottleneck in VLA performance.

---

4. Visual instruction tuning towards general-purpose multimodal model: A survey

URL: View paper

Brief Assessment

Visual Instruction Survey[61] is a comprehensive survey paper on visual instruction tuning methods across various vision tasks. It does not present empirical ablation studies on vision encoder fine-tuning for VLA policies, focusing instead on cataloging existing visual instruction tuning approaches.

---

5. Instructvla: Vision-language-action instruction tuning from understanding to manipulation

URL: View paper

Prior Art Analysis

InstructVLA[34] demonstrates similar findings about the critical importance of vision encoders in VLA performance. The candidate paper shows that removing the dinov2-based vision encoder from the action expert results in a 50.0% performance drop, and incorporating film enhancement yields a further 15.3% improvement. This evidence, combined with their architectural design that explicitly emphasizes the vision encoder's role in capturing task-relevant visual cues, demonstrates that prior work had already identified and validated the vision encoder as a critical bottleneck for VLA performance before the original paper's submission.

Evidence

Evidence 1 - **Rationale**: Both papers identify the vision encoder as a critical bottleneck. The candidate paper provides quantitative evidence (50.0% performance drop) demonstrating the vision encoder's critical role, which directly parallels the original paper's finding about the vision encoder being crucial for strong performance. - **Original**: lastly, our analysis also reveals that the vision encoder is a critical bottleneck, and the ability to fine-tune it is crucial for strong performance. - **Candidate**: as shown in figure 6(c), while the base vlm offers general visual understanding, fine-grained perception for manipulation tasks demands richer representations. removing the dinov2-based vit encoder from the action expert results in a 50.0% performance drop, highlighting its critical role in capturin...

Evidence 2 - **Rationale**: The candidate paper's ablation study explicitly demonstrates through empirical evidence that the vision encoder is a critical component, with its removal causing severe performance degradation. This finding predates the original paper's similar conclusion. - **Original**: lastly, we identify the vision encoder as a critical bottleneck, with fine-tuning the vision encoder proving crucial for strong control performance. - **Candidate**: ablation on action expert designs. as shown in figure 6(c), while the base vlm offers general visual understanding, fine-grained perception for manipulation tasks demands richer representations. removing the dinov2-based vit encoder from the action expert results in a 50.0% performance drop, highlig...

Evidence 3 - **Rationale**: Both papers conduct systematic ablation studies on vision encoder importance. The candidate paper's finding that removing the vision encoder causes 50.0% performance drop provides similar evidence to the original paper's observation of significant degradation when freezing the vision encoder. - **Original**: table 3 shows the performance of three models when the vision encoder is frozen during vlm4vla training. we observe a significant performance degradation for all models on both the calvin and simpler benchmarks after freezing the vision encoder - **Candidate**: effects of latent action queries. latent action tokens are a key design component for decoupling high-level vlm planning from low-level action generation. as shown in figure 6 (b), we vary the number of tokens from 16 to 128. too few tokens limit behavioral diversity, while too many reduce training ...

---

6. Fine-tuning vision-language-action models: Optimizing speed and success

URL: View paper

Brief Assessment

Fine-tuning Speed Success[58] focuses on VLA fine-tuning strategies (action decoding, representations, objectives) rather than analyzing vision encoder importance. The candidate does not investigate vision encoder freezing vs. fine-tuning as a primary research question.

---

7. Embodiment Transfer Learning for Vision-Language-Action Models

URL: View paper

Brief Assessment

Embodiment Transfer Learning[64] focuses on multi-robot transfer learning challenges and does not systematically analyze vision encoder fine-tuning importance across different VLA architectures.

---

8. Large VLM-based Vision-Language-Action Models for Robotic Manipulation: A Survey

URL: View paper

Prior Art Analysis

Large VLM Survey[44] demonstrates that prior work has already identified the vision encoder as a critical component requiring fine-tuning in VLA models. The survey explicitly discusses how vision encoders are essential for VLA performance and references multiple prior works that address vision encoder training strategies, indicating this finding was established before the original paper's contribution.

Evidence

Evidence 1 - **Rationale**: This shows that prior work (OpenVLA) already recognized the importance of vision encoder selection and fine-tuning for VLA performance, predating the original paper's claimed discovery. - **Original**: our analysis also reveals that the vision encoder is a critical bottleneck, and the ability to fine-tune it is crucial for strong performance. - **Candidate**: openvla [26] replaces the large-parameter vision encoder in rt series with a combination of siglip [92] and dinov2 [93]. through fine-tuning on large-scale real-world robotic manipulation data, it achieves superior performance with fewer model parameters.

Evidence 2 - **Rationale**: This demonstrates that prior work (ReVLA) had already identified vision encoder fine-tuning as critical and developed specific strategies to address it, refuting the novelty of this finding. - **Original**: we identify the vision encoder as a critical bottleneck, with fine-tuning the vision encoder proving crucial for strong control performance. - **Candidate**: revla [96] employs a reversible training strategy that gradually restores the vision encoder to its original pre-trained state, mitigating catastrophic forgetting during fine-tuning and improving out-of-distribution visual generalization.

---

9. Tgrpo: Fine-tuning vision-language-action model via trajectory-wise group relative policy optimization

URL: View paper

Brief Assessment

TGRPO[59] focuses on reinforcement learning-based fine-tuning methods for VLA models and does not investigate the importance of vision encoder fine-tuning versus freezing, which is the core of the original contribution.

---

10. Zeetad: Adapting pretrained vision-language model for zero-shot end-to-end temporal action detection

URL: View paper

Brief Assessment

ZeetAD[62] focuses on zero-shot temporal action detection using CLIP for video understanding, not on vision-language-action policies for robotic manipulation. The technical domains and tasks are fundamentally different.

Appendix: Text Similarity Detection

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

References

- [0] VLM4VLA: Revisiting Vision-Language-Models in Vision-Language-Action Models View paper
- [1] Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation View paper
- [2] Vision-language-action models for robotics: A review towards real-world applications View paper
- [3] Vision language action models in robotic manipulation: A systematic review View paper
- [4] Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation View paper
- [5] Robopoint: A vision-language model for spatial affordance prediction for robotics View paper
- [6] Multimodal spatial language maps for robot navigation and manipulation View paper
- [7] Vision-language model-driven scene understanding and robotic object manipulation View paper
- [8] Recipe for Vision-Language-Action Models in Robotic Manipulation: A Survey View paper
- [9] Robouniview: Visual-language model with unified view representation for robotic manipulation View paper
- [10] TinyVLA: Toward Fast, Data-Efficient Vision-Language-Action Models for Robotic Manipulation View paper
- [11] Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation View paper
- [12] Vision-language foundation models as effective robot imitators View paper
- [13] Visual language models agent applications in robotic manipulation tasks View paper
- [14] Vlas: Vision-language-action model with speech instructions for customized robot manipulation View paper
- [15] Reflective Planning: Vision-Language Models for Multi-Stage Long-Horizon Robotic Manipulation View paper
- [16] Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation View paper
- [17] Scalable Vision-Language-Action Model Pretraining for Robotic Manipulation with Real-Life Human Activity Videos View paper
- [18] Language-conditioned learning for robotic manipulation: A survey View paper
- [19] AHA: A Vision-Language-Model for Detecting and Reasoning Over Failures in Robotic Manipulation View paper
- [20] ManipLLM: Embodied Multimodal Large Language Model for Object-Centric Robotic Manipulation View paper
- [21] A Review of Advances in Large Language and Vision Models for Robotic Manipulation: Techniques, Integrations, and Challenges View paper
- [22] Mla: A multisensory language-action model for multimodal understanding and forecasting in robotic manipulation View paper
- [23] Bi-VLA: Vision-Language-Action Model-Based System for Bimanual Robotic Dexterous Manipulations View paper
- [24] Foundation models in robotics: Applications, challenges, and the future View paper
- [25] Robodexvlm: Visual language model-enabled task planning and motion control for dexterous robot manipulation View paper

- [26] Vision-Language Models in Industrial Robotics View paper
- [27] FedVLA: Federated Vision-Language-Action Learning with Dual Gating Mixture-of-Experts for Robotic Manipulation View paper
- [28] Multimodal fusion and vision-language models: A survey for robot vision View paper
- [29] Object-Centric Prompt-Driven Vision-Language-Action Model for Robotic Manipulation View paper
- [30] Chatvla: Unified multimodal understanding and robot control with vision-language-action model View paper
- [31] Vision-Language Models Enabled Robot Manipulation View paper
- [32] GraphCoT-VLA: A 3D Spatial-Aware Reasoning Vision-Language-Action Model for Robotic Manipulation with Ambiguous Instructions View paper
- [33] Evo-1: Lightweight vision-language-action model with preserved semantic alignment View paper
- [34] Instructvla: Vision-language-action instruction tuning from understanding to manipulation View paper
- [35] CrayonRobo: Object-Centric Prompt-Driven Vision-Language-Action Model for Robotic Manipulation View paper
- [36] Language reasoning in vision-language-action model for robotic grasping View paper
- [37] LADEV: A Language-Driven Testing and Evaluation Platform for Vision-Language-Action Models in Robotic Manipulation View paper
- [38] Vision-Language Models for Robot Success Detection View paper
- [39] SKT: Integrating State-Aware Keypoint Trajectories with Vision-Language Models for Robotic Garment Manipulation View paper
- [40] Manipulate-anything: Automating real-world robots using vision-language models View paper
- [41] Interactive robot action replanning using multimodal llm trained from human demonstration videos View paper
- [42] Integrating With Multimodal Information for Enhancing Robotic Grasping With Vision-Language Models View paper
- [43] Embodied AI with Foundation Models for Mobile Service Robots: A Systematic Review View paper
- [44] Large VLM-based Vision-Language-Action Models for Robotic Manipulation: A Survey View paper
- [45] Vima: Robot manipulation with multimodal prompts View paper
- [46] Multi-Modal Perception With Vision, Language, and Touch for Robot Manipulation View paper
- [47] KALIE: Fine-Tuning Vision-Language Models for Open-World Manipulation Without Robot Data View paper
- [48] GPTArm: An Autonomous Task Planning Manipulator Grasping System Based on Visionâ□□Language Models View paper
- [49] VLMPC: Vision-Language Model Predictive Control for Robotic Manipulation View paper
- [50] A Vision-Language Model Approach for Object Segmentation and Robotic Grasping View paper
- [51] BitVLA: 1-bit Vision-Language-Action Models for Robotics Manipulation View paper
- [52] Smolvla: A vision-language-action model for affordable and efficient robotics View paper
- [53] 3ds-vla: A 3d spatial-aware vision language action model for robust multi-task manipulation View paper
- [54] What Matters in Employing Vision Language Models for Tokenizing Actions in Robot Control? View paper
- [55] A survey on efficient vision-language-action models View paper
- [56] Vla-adapter: An effective paradigm for tiny-scale vision-language-action model View paper
- [57] Openvla: An open-source vision-language-action model View paper
- [58] Fine-tuning vision-language-action models: Optimizing speed and success View paper
- [59] Tgrpo: Fine-tuning vision-language-action model via trajectory-wise group relative policy optimization View paper
- [60] The future of action recognition: are multi-modal visual language models the key? View paper
- [61] Visual instruction tuning towards general-purpose multimodal model: A survey View paper
- [62] Zeetad: Adapting pretrained vision-language model for zero-shot end-to-end temporal action detection View paper
- [63] Coa-vla: Improving vision-language-action models via visual-text chain-of-affordance View paper
- [64] Embodiment Transfer Learning for Vision-Language-Action Models View paper
- [65] Physically grounded vision-language models for robotic manipulation View paper
- [66] Survey of vision-language-action models for embodied manipulation View paper
- [67] Embodiedgpt: Vision-language pre-training via embodied chain of thought View paper
- [68] VLP: Vision-Language Preference Learning for Embodied Manipulation View paper
- [69] Momanipvla: Transferring vision-language-action models for general mobile manipulation View paper
- [70] A survey on vision-language-action models for embodied ai View paper