Novelty Assessment Report

Paper: Unified Vision-Language-Action Model

PDF URL: https://openreview.net/pdf?id=PklMD8PwUy

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Vision-language-action models (VLAs) have garnered significant attention for their potential in advancing robotic manipulation. However, previous approaches predominantly rely on the general comprehension capabilities of vision-language models (VLMs) to generate action signals, often overlooking the rich temporal and causal structure embedded in visual observations. In this paper, we present UniVLA, a unified and native multimodal VLA model that autoregressively models vision, language, and action signals as discrete token sequences. This tokenized formulation naturally supports flexible multimodal task learning, particularly from large-scale video data, and further demonstrates that generative vision supervision can significantly enhance visual understanding. By incorporating world modeling during post-training, UniVLA captures causal dynamics from videos, facilitating effective transfer to downstream policy learning—especially for long-horizon tasks. Our approach sets new state-of-the-art results across several widely used simulation benchmarks, including CALVIN, LIBERO, and Simplenv-Bridge, substantially outperforming prior methods. For example, UniVLA achieves 95.5% average success rate on LIBERO benchmark, surpassing π_0-FAST's 85.5%. We further demonstrate its broad applicability through experiments on real-world ALOHA manipulation tasks and autonomous driving scenarios.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Unified Vision-Language-Action Modeling Through Autoregressive Token Sequences**

A total of **38 papers** were analyzed and organized into a taxonomy with **27 categories**.

Taxonomy Overview

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

- **Action Tokenization and Representation**
- **Reasoning and Chain-of-Thought Integration**
- **Hybrid Action Generation Paradigms**
- **World Modeling and Predictive Dynamics**
- **Domain-Specific VLA Applications**
- **Unified Multimodal Autoregressive Modeling**
- **Temporal and Multi-Frame Modeling**
- **Architectural Optimization for VLA**
- **Inference Efficiency and Acceleration**
- **Vision-Language Tracking and Localization**
- ... and 2 more categories

Complete Taxonomy Tree

- Unified Vision-Language-Action Modeling Through Autoregressive Token Sequences Survey Taxonomy
- Action Tokenization and Representation
  - Frequency-Domain and Compressed Action Encoding (2 papers)
  - [1] Fast: Efficient action tokenization for vision-language-action models (Karl Pertsch, 2025) View paper
  - [38] Stable-FAST: Stabilizing Inference of Autoregressive Vision-Language-Action Models (X Luo, n.d.) View paper
  - Vector Quantization for Action Tokenization (2 papers)
  - [20] FASTer: Toward Efficient Autoregressive Vision Language Action Modeling via Neural Action Tokenization (Yicheng Liu, 2025) View paper
  - [22] VQ-VLA: Improving Vision-Language-Action Models via Scaling Vector-Quantized Action Tokenizers (Wang Yating, 2025) View paper
  - Universal Latent Action Learning (1 papers)
  - [26] LatBot: Distilling Universal Latent Actions for Vision-Language-Action Models (Zuolei Li, 2025) View paper
  - Neural Action Tokenization with Spatio-Temporal Modeling (1 papers)
  - [37] From Language to Action Streams: Bridging LLM Autoregression for Long-Horizon Robot Action Prediction (Z Wang, n.d.) View paper
- Reasoning and Chain-of-Thought Integration
  - Visual Chain-of-Thought Reasoning (3 papers)
  - [2] Coa-vla: Improving vision-language-action models via visual-text chain-of-affordance (J Li, 2025) View paper
  - [4] CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models (Qing-Qing Zhao, 2025) View paper
  - [14] Flowvla: Visual chain of thought-based motion reasoning for vision-language-action models (Yan Hao-dong, 2025) View paper
  - Integrated Reasoning-Action Architecture (1 papers)
  - [11] DeepThinkVLA: Enhancing Reasoning Capability of Vision-Language-Action Models (Cheng Yin, 2025) View paper
  - Spatio-Temporal Visual Reasoning (1 papers)

    - [30] FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving (Zeng, 2025) View paper
- Hybrid Action Generation Paradigms
    - Autoregressive-Diffusion Hybrid Models (3 papers)
    - [12] HybridVLA: Collaborative Autoregression and Diffusion in a Unified Vision-Language-Action Model (J Liu, 2025) View paper
    - [15] DiffusionVLA: Scaling Robot Foundation Models via Unified Diffusion and Autoregression (Wen Junjie, 2025) View paper
    - [28] HybridVLA: Collaborative Diffusion and Autoregression in a Unified Vision-Language-Action Model (Liu, 2025) View paper
    - Diffusion-Based Vision-Language-Action Models (1 papers)
    - [25] Dream-VL & Dream-VLA: Open Vision-Language and Vision-Language-Action Models with Diffusion Language Model Backbone (Jiacheng Ye, 2025) View paper
- World Modeling and Predictive Dynamics
    - Autoregressive World Models for Action (2 papers)
    - [5] Physical autoregressive model for robotic manipulation without action pretraining (Song Zi-jian, 2025) View paper
    - [6] WorldVLA: Towards Autoregressive Action World Model (Cen Jun, 2025) View paper
    - World Model-Enhanced Data Scaling (1 papers)
    - [21] DriveVLA-W0: World Models Amplify Data Scaling Law in Autonomous Driving (Li, 2025) View paper
    - Occupancy-Based World Modeling (1 papers)
    - [34] OccLLaMA: An Occupancy-Language-Action Generative World Model for Autonomous Driving (Julong Wei, 2024) View paper
- Domain-Specific VLA Applications
    - Autonomous Driving VLA Models (3 papers)
    - [7] OpenDriveVLA: Towards End-to-end Autonomous Driving with Large Vision Language Action Model (Zhou Xingcheng, 2025) View paper
    - [9] Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers (Chen Yuntao, 2025) View paper
    - [13] AutoVLA: A Vision-Language-Action Model for End-to-End Autonomous Driving with Adaptive Reasoning and Reinforcement Fine-Tuning (Zhou Zewei, 2025) View paper
    - Open-World Instruction Following Agents (1 papers)
    - [8] OmniJARVIS: Unified Vision-Language-Action Tokenization Enables Open-World Instruction Following Agents (Shaofei Cai, 2024) View paper
- Unified Multimodal Autoregressive Modeling
    - Cross-Modal Unified Tokenization (1 papers)
    - [3] Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action (Jiasen LÃ¼, 2024) View paper
    - Native Multimodal VLA Architectures ★ (1 papers)
    - [0] Unified Vision-Language-Action Model (Anon et al., 2026) View paper
    - Instruction-Conditioned Sequence Generation (1 papers)
    - [23] InstructSeq: Unifying Vision Tasks with Instruction-conditioned Multi-modal Sequence Generation (Fang, 2023) View paper
- Temporal and Multi-Frame Modeling
    - Multi-Frame History Integration (1 papers)
    - [17] CronusVLA: Towards Efficient and Robust Manipulation via Multi-Frame Vision-Language-Action Modeling (Li Hao, 2025) View paper
    - Long-Term Video Action Recognition (1 papers)
    - [27] VT-LVLM-AR: A Video-Temporal Large Vision-Language Model Adapter for Fine-Grained Action Recognition in Long-Term Videos (He Shuwei, 2025) View paper
- Architectural Optimization for VLA
    - Trajectory-Aware Attention Mechanisms (1 papers)
    - [29] Actra: Optimized Transformer Architecture for Vision-Language-Action Models in Robot Learning (Ma, 2024) View paper
    - Active Visual Perception Systems (1 papers)
    - [31] Look, Zoom, Understand: The Robotic Eyeball for Embodied Perception (Jiashu Yang, 2025) View paper
- Inference Efficiency and Acceleration
    - Speculative and Parallel Decoding (2 papers)
    - [10] Spec-VLA: Speculative Decoding for Vision-Language-Action Models with Relaxed Acceptance (Songsheng Wang, 2025) View paper
    - [19] PD-VLA: Accelerating Vision-Language-Action Model Integrated with Action Chunking via Parallel Decoding (Song Wen-xuan, 2025) View paper
    - Token and Action Redundancy Reduction (1 papers)
    - [16] Think Twice, Act Once: Token-Aware Compression and Action Reuse for Efficient Inference in Vision-Language-Action Models (Tan Xudong, 2025) View paper
- Vision-Language Tracking and Localization (2 papers)
    - [32] Towards Unified Token Learning for Vision-Language Tracking (Zheng Yao-Zong, 2023) View paper
    - [33] Toward Unified Token Learning for Vision-Language Tracking (Yaozong Zheng, 2023) View paper
- Theoretical Foundations and Surveys (2 papers)
    - [18] Pure vision language action (vla) models: A comprehensive survey (Zhang DaPeng, 2025) View paper
    - [24] A Survey on Vision-Language-Action Models: An Action Tokenization Perspective (Zhong, 2025) View paper
- Semantic Representation and Meaning Extraction
    - Distribution-Based Meaning Representations (1 papers)
    - [35] Meaning Representations from Trajectories in Autoregressive Models (Liu, 2023) View paper
    - Reinforcement Learning for Visual Token Generation (1 papers)
    - [36] Selftok-Zero: Reinforcement Learning for Visual Generation via Discrete and Autoregressive Visual Tokens (B Wang, n.d.) View paper

Narrative

Core task: unified vision-language-action modeling through autoregressive token sequences. This emerging field seeks to unify perception, language understanding, and action generation within a single autoregressive framework, treating all modalities as discrete token sequences. The taxonomy reveals a rich landscape organized around several key themes. Action Tokenization and Representation explores how continuous control signals are discretized for autoregressive modeling, while Reasoning and Chain-of-Thought Integration examines methods that interleave explicit reasoning steps with action prediction. Hybrid Action Generation Paradigms investigates architectures that blend discrete token prediction with continuous action heads or diffusion processes, as seen in works like HybridVLA[12] and DiffusionVLA[15]. World Modeling and Predictive Dynamics focuses on learning forward models that predict future states, enabling planning and simulation. Domain-Specific VLA Applications addresses specialized deployments in robotics, autonomous driving, and embodied AI, with examples like OpenDriveVLA[7] and DrivingGPT[9]. Unified Multimodal Autoregressive Modeling, the branch housing the original paper, emphasizes native architectures that process vision, language, and action tokens within a single transformer backbone, exemplified by Unified-IO 2[3] and related systems.

Several active research directions reveal key trade-offs and open questions. One line explores whether pure autoregressive token prediction suffices or whether hybrid approaches combining discrete and continuous representations yield better control precision and sample efficiency. Another examines the role of explicit reasoning: CoT-VLA[4] and DeepThinkVLA[11] demonstrate that chain-of-thought prompting can improve decision quality, yet questions remain about computational overhead and generalization. Unified VLA[0] sits within the Native Multimodal VLA Architectures cluster, emphasizing end-to-end autoregressive modeling without hybrid components. Compared to Unified-IO 2[3], which pioneered broad multimodal unification, Unified VLA[0] likely refines architectural choices or training strategies for tighter vision-language-action integration. Relative to reasoning-augmented approaches like CoT-VLA[4], it appears to prioritize streamlined token-level prediction, trading explicit intermediate reasoning for architectural simplicity and potentially faster inference.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses specifically on vision-language-action (VLA) models that natively generate action tokens autoregressively alongside vision and language, with generative vision supervision. The sibling subtopics represent broader multimodal unification efforts: one addresses tokenization across many modalities (including audio) beyond VLA, while the other focuses on instruction-conditioned vision task generation without action modeling. All three share autoregressive sequence modeling as a core mechanism but differ in modality scope and end objectives.

Similarities: - All employ autoregressive token sequence generation as the fundamental modeling paradigm - All unify multiple modalities (vision, language, and potentially others) into shared representational frameworks - All leverage transformer-based architectures for cross-modal reasoning - All aim to handle diverse tasks through unified tokenization rather than task-specific architectures

Differences: - Original leaf specifically models action tokens for embodied control, while siblings focus on perception/generation tasks without action - Original leaf emphasizes generative vision supervision as a training signal, while siblings may use different supervision strategies - Cross-Modal Unified Tokenization extends to audio and broader modality sets beyond VLA's vision-language-action scope - Instruction-Conditioned Sequence Generation targets vision task diversity through instructions, whereas the original leaf targets robot control policies - Original leaf excludes adapted pretrained VLMs, suggesting native end-to-end training, while siblings' adaptation strategies are unspecified

Suggested Search Directions: - Investigate whether native VLA architectures could benefit from audio tokenization for richer embodied understanding - Explore how instruction-conditioned generation techniques might enhance VLA policy flexibility and generalization - Examine trade-offs between native multimodal training versus adapting pretrained models across all three paradigms

Sibling Subtopics

- **Cross-Modal Unified Tokenization** (leaves: 1, papers: 1)
- Scope: Frameworks tokenizing diverse modalities including image, text, audio, and action into shared semantic spaces.
- Exclude: Excludes vision-language-action only models; those belong under VLA-specific unified modeling.
- **Instruction-Conditioned Sequence Generation** (leaves: 1, papers: 1)
- Scope: Unified frameworks for diverse vision tasks through instruction-conditioned multimodal sequence generation.
- Exclude: Excludes action-specific models; those belong under VLA-specific categories.

Contributions Analysis

Overall novelty summary. UniVLA proposes a unified autoregressive framework that models vision, language, and action as discrete token sequences, incorporating generative vision supervision and world modeling during post-training. The paper resides in the 'Native Multimodal VLA Architectures' leaf, which currently contains no sibling papers in the taxonomy. This positioning suggests the work occupies a relatively sparse research direction within the broader unified multimodal autoregressive modeling branch, distinguishing itself from hybrid paradigms and reasoning-augmented approaches that populate neighboring taxonomy leaves.

The taxonomy reveals several active neighboring directions. The 'Hybrid Action Generation Paradigms' branch explores autoregressive-diffusion combinations and diffusion-based VLA models, while 'Reasoning and Chain-of-Thought Integration' investigates explicit intermediate reasoning steps before action generation. The 'World Modeling and Predictive Dynamics' branch, particularly relevant to UniVLA's post-training approach, contains methods for autoregressive world models and occupancy-based representations. UniVLA's emphasis on native multimodal tokenization and world modeling positions it at the intersection of unified sequence modeling and predictive dynamics, diverging from hybrid architectures that separate discrete and continuous action representations.

Among 30 candidates examined, the contribution-level analysis reveals mixed novelty signals. The core UniVLA architecture shows overlap with prior work: 2 of 10 examined candidates provide potentially refuting evidence for the unified modeling framework itself. The unified sequence modeling contribution appears more distinctive, with 0 refutable candidates among 10 examined, suggesting this aspect may represent a clearer advance. The benchmark performance claims face stronger prior work, with 3 of 10 candidates offering comparable or overlapping results. These statistics reflect a limited semantic search scope rather than exhaustive coverage, indicating that substantial related work exists in this rapidly evolving area.

Based on the top-30 semantic matches and taxonomy structure, UniVLA appears to refine existing unified autoregressive approaches rather than introduce fundamentally new paradigms. The sparse population of its taxonomy leaf suggests either emerging novelty or incomplete literature coverage in this analysis. The world modeling integration during post-training may represent the most distinctive technical contribution, though the limited search scope prevents definitive assessment of how this compares to concurrent developments in predictive dynamics for VLA systems.

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

Contribution 1: UniVLA unified vision-language-action model

Description: The authors introduce UniVLA, a novel architecture that represents vision, language, and action modalities as discrete tokens in a unified vocabulary and models them autoregressively. This design enables tighter cross-modal integration and supports large-scale video-based training, offering an alternative to existing VLA paradigms that rely on separate vision encoders.

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. Multimodal fusion and vision-language models: A survey for robot vision

URL: View paper

Brief Assessment

Multimodal Fusion Survey[44] is a survey paper focused on multimodal fusion and vision-language models for robot vision tasks, not a technical contribution proposing a unified VLA architecture with discrete token representations. It reviews existing methods rather than introducing novel architectures.

2. LoHoVLA: A Unified Vision-Language-Action Model for Long-Horizon Embodied Tasks

URL: View paper

Brief Assessment

LoHoVLA[43] focuses on hierarchical long-horizon task decomposition with joint language-action token generation, while the original paper emphasizes unified autoregressive modeling of all modalities as discrete tokens with world model post-training for scalable video-based learning.

3. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation

URL: View paper

Brief Assessment

CogAct[45] uses a componentized architecture with separate vision, language, and action modules where actions are predicted by a specialized diffusion transformer conditioned on VLM output, rather than representing all modalities as discrete tokens in a unified autoregressive framework as UniVLA does.

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

URL: View paper

Brief Assessment

XR-1[40] focuses on learning unified vision-motion codes (UVMC) as an intermediate representation between observations and actions using a dual-branch VQ-VAE, rather than directly modeling vision, language, and action as discrete tokens in a unified autoregressive framework as UniVLA does.

5. Unified Diffusion VLA: Vision-Language-Action Model via Joint Discrete Denoising Diffusion Process

URL: View paper

Brief Assessment

Unified Diffusion VLA[47] uses a joint diffusion process for synchronous denoising across modalities, fundamentally different from UniVLA's autoregressive discrete token modeling approach.

6. ShowUI: One Vision-Language-Action Model for GUI Visual Agent

URL: View paper

Brief Assessment

ShowUI[46] focuses on GUI visual agents with UI-specific token selection mechanisms, not a general unified VLA framework for robotic manipulation. The architectural approaches differ fundamentally in their application domains and design principles.

7. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

URL: View paper

Prior Art Analysis

RT-2[41] demonstrates that vision-language-action models were proposed prior to UniVLA. RT-2[41] introduces the concept of vision-language-action (VLA) models that directly integrate vision-language models into robotic control by representing actions as text tokens within a unified framework. The paper explicitly states 'we refer to such category of models as vision-language-action models (vla) and instantiate an example of such a model, which we call rt-2' and describes how 'we express the actions as text tokens and incorporate them directly into the training set of the model in the same way as natural language tokens.' This predates UniVLA's claim of being 'the first unified vision-language-action (vla) model that encodes vision, language, and action as discrete tokens within a shared vocabulary.'

Evidence

Evidence 1 - **Rationale**: RT-2[41] explicitly introduces the VLA model category and instantiates RT-2 as an example, predating UniVLA's claim of being 'the first' unified VLA model. - **Original**: we propose univla, the first unified vision-language-action (vla) model that encodes vision, language, and action as discrete tokens within a shared vocabulary, jointly modeling them through autoregressive sequence learning. - **Candidate**: we refer to such category of models as vision-language-action models (vla) and instantiate an example of such a model, which we call rt-2. our extensive evaluation (6k evaluation trials) shows that our approach leads to performant robotic policies and enables rt-2 to obtain a range of emergent capab...

Evidence 2 - **Rationale**: RT-2[41] describes representing actions as text tokens in a unified format with language, which is the core architectural principle claimed as novel by UniVLA. - **Original**: unlike prior vla approaches that typically rely on an extra vision encoder to extract image features and generate only action outputs, univla represents vision, language, and action as discrete tokens within a unified autoregressive framework. - **Candidate**: in order to fit both natural language responses and robotic actions into the same format, we express the actions as text tokens and incorporate them directly into the training set of the model in the same way as natural language tokens.

Evidence 3 - **Rationale**: RT-2[41] describes the same unified approach of transforming actions into discrete tokens treated identically to language tokens, which UniVLA claims as its novel contribution. - **Original**: we introduce univla, a novel framework for unified vision-language-action learning. as illustrated in figure 1, we propose a unified framework that supports bothmultimodalandmulti-tasklearning. at the modality level, vision, language, and action signals are all transformed into discrete tokens and m... - **Candidate**: we take a direct approach to this problem, representing actions as tokens in the model's output, which are treated in the same way as language tokens. we base our action encoding on the discretization proposed by brohan et al. (2022) for the rt-1 model. the action space consists of 6-dof positional ...

Evidence 4 - **Rationale**: RT-2[41] demonstrates the unified token approach for joint learning across modalities, directly challenging UniVLA's novelty claim about unified token representation enabling cross-modal integration. - **Original**: this unified token representation allows for joint learning across modalities, fostering deeper cross-modal understanding and integration. - **Candidate**: in this way, vision-language models can be directly trained to act as instruction following robotic policies. this simple approach is in contrast with prior alternatives for incorporating vlms into robot policies (shridhar et al., 2022a) or designing new vision-languageaction architectures from scra...

8. Palm-e: An embodied multimodal language model

URL: View paper

Brief Assessment

PaLM-E[39] focuses on embodied multimodal language models that inject continuous sensor modalities into language models for robotic reasoning, but uses a different architecture where visual observations are encoded and projected into language embedding space rather than representing all modalities as discrete tokens in a unified autoregressive framework as UniVLA does.

9. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action

URL: View paper

Brief Assessment

Unified-IO 2[3] focuses on a general-purpose multimodal model spanning vision, language, audio, and action across diverse tasks (image editing, generation, audio synthesis), whereas UniVLA specifically targets robotic manipulation with autoregressive discrete token modeling optimized for perception-action loops and world model learning.

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

URL: View paper

Prior Art Analysis

VLA Embodied Survey[42] demonstrates that prior work exists on unified vision-language-action models with discrete token representations. The survey explicitly discusses models like Gato, RoboCat, and WorldVLA that represent vision, language, and action as discrete tokens within unified vocabularies and model them autoregressively—the same core paradigm claimed as novel by the original paper. The survey describes how Gato 'enables the simultaneous training of different tasks' through 'a unified tokenization scheme, harmonizing the input and output across diverse tasks and domains,' and how WorldVLA 'quantizes multimodal data into discrete tokens, forming a shared vocabulary of quantized multimodal tokens' where 'all modalities can be modeled autoregressively.' These descriptions directly parallel the original paper's claimed novelty of representing 'vision, language, and action as discrete tokens within a unified autoregressive framework.'

Evidence

Evidence 1 - **Rationale**: This evidence pair shows that Gato, published before the original paper, already implemented a unified tokenization scheme that harmonizes different modalities (vision, language, action) within a shared vocabulary—directly challenging the claim of being 'the first' to do so. - **Original**: we propose univla, the first unified vision-language-action (vla) model that encodes vision, language, and action as discrete tokens within a shared vocabulary, jointly modeling them through autoregressive sequence learning. - **Candidate**: gato proposes a model that can play atari games, caption images, and stack blocks, all with a single set of model parameters. this achievement is facilitated by a unified tokenization scheme, harmonizing the input and output across diverse tasks and domains. consequently, gato enables the simultaneo...

Evidence 2 - **Rationale**: This pair demonstrates that WorldVLA (and another model also named UniVLA) were already known to use discrete token quantization with autoregressive modeling across all modalities, refuting the novelty claim of the unified discrete token representation approach. - **Original**: univla represents vision, language, and action as discrete tokens within a unified autoregressive framework. this unified modeling paradigm enables multi-modal outputs and supports a wide range of tasks - **Candidate**: worldvla and univla advance this direction by integrating vlas with world models. it quantizes multimodal data into discrete tokens, forming a shared vocabulary of quantized multimodal tokens. consequently, all modalities can be modeled autoregressively, enabling not only action and text generation ...

Evidence 3 - **Rationale**: RoboCat, building on Gato, uses VQ-GAN for discrete image tokenization and predicts both actions and future observations autoregressively, demonstrating that the paradigm of discrete token-based unified modeling existed in prior work. - **Original**: unlike prior vla approaches that typically rely on an extra vision encoder to extract image features and generate only action outputs, univla represents vision, language, and action as discrete tokens within a unified autoregressive framework. - **Candidate**: robocat proposes a self-improvement process designed to enable an agent to rapidly adapt to new tasks with as few as 100 demonstrations. this self-improvement process iteratively finetunes the model and self-generates new data with the finetuned model. built upon the gato model, robocat incorporates...

Evidence 4 - **Rationale**: The survey explicitly identifies that unifying the three modalities (vision, language, action) under an autoregressive paradigm with quantized tokens was an established direction before the original paper, with WorldVLA and another UniVLA model already advancing this approach. - **Original**: we present univla, a unified and native multimodal vla model that autoregressively models vision, language, and action signals as discrete token sequences. - **Candidate**: visual autoregressive modeling (v ar) with quantized visual tokens demonstrates improved performance over diffusion models in image generation. this suggests that the three modalities of vlas can be unified under the autoregressive paradigm. worldvla and univla advance this direction by integrating ...

Contribution 2: Unified sequence modeling framework supporting multimodal tasks

Description: The framework enables diverse multimodal tasks including text-supervised perception grounding, vision-supervised world modeling, and action-supervised policy learning within a single architecture. The authors demonstrate that world model post-training substantially improves performance and efficiency in downstream policy learning across simulation benchmarks, real-world robots, and driving scenarios.

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. A step toward world models: A survey on robotic manipulation

URL: View paper

Brief Assessment

World Models Survey[52] is a survey paper that reviews existing world model approaches in robotic manipulation, rather than proposing a novel unified framework. It discusses various paradigms (implicit modeling, latent dynamics, video generation) but does not claim to be the first to propose a unified multimodal sequence modeling framework for policy learning.

2. Imagine-2-Drive: Leveraging High-Fidelity World Models via Multi-Modal Diffusion Policies

URL: View paper

Brief Assessment

Imagine-2-Drive[48] focuses on world models for autonomous driving using diffusion-based policies, not a unified vision-language-action framework for robotic manipulation. The candidate addresses trajectory generation in driving scenarios rather than the multimodal task learning (perception grounding, world modeling, policy learning) proposed in the original paper.

3. 3D-VLA: A 3D Vision-Language-Action Generative World Model

URL: View paper

Brief Assessment

3D-VLA[49] focuses on 3D perception and world models for spatial reasoning in embodied environments, while the original paper emphasizes unified token-based sequence modeling across vision-language-action modalities with video-based temporal dynamics learning. The candidate's 3D-centric approach and diffusion-based generation differ fundamentally from the original's autoregressive token framework.

4. Spatial-Temporal Aware Visuomotor Diffusion Policy Learning

URL: View paper

Brief Assessment

Spatial-Temporal Visuomotor[50] focuses on diffusion-based policy learning with 3D Gaussian world models for spatial-temporal awareness in robotic manipulation, not on unified sequence modeling across text-supervised perception grounding, vision-supervised world modeling, and action-supervised policy learning within a single autoregressive architecture.

5. Pre-training contextualized world models with in-the-wild videos for reinforcement learning

URL: View paper

Brief Assessment

Contextualized World Models[51] focuses on separating context and dynamics modeling for video pre-training in RL, not on unified multimodal sequence modeling with text-supervised perception grounding and action-supervised policy learning within a single architecture.

6. GenRL: Multimodal-foundation world models for generalization in embodied agents

URL: View paper

Brief Assessment

GenRL[54] focuses on connecting foundation VLMs with world models for RL using video-language alignment, not on unified vision-language-action token sequences. The original paper's autoregressive token-based framework for perception grounding, world modeling, and policy learning differs fundamentally from GenRL's approach of aligning separate representation spaces.

7. Multimodal foundation world models for generalist embodied agents

URL: View paper

Brief Assessment

Multimodal Foundation Agents[56] focuses on world models for grounding vision-language prompts in embodied RL domains without language annotations, using a connector-aligner architecture. The original paper's unified token-based autoregressive framework for text-supervised perception, vision-supervised world modeling, and action-supervised policy learning represents a different architectural approach to multimodal integration.

8. Merlot: Multimodal neural script knowledge models

URL: View paper

Brief Assessment

Merlot[57] focuses on learning multimodal script knowledge from YouTube videos for video QA and visual commonsense reasoning tasks, using contrastive frame-caption matching and temporal ordering objectives. The original paper's contribution centers on a unified vision-language-action model for robotic policy learning with world model post-training, which is a fundamentally different application domain and architectural approach.

9. Can World Models Benefit VLMs for World Dynamics?

URL: View paper

Brief Assessment

World Models VLMs[55] focuses on using video diffusion models as generative encoders for vision-language understanding tasks, not on unified vision-language-action modeling for robotic policy learning. The candidate addresses visual question answering and spatial reasoning, while the original paper targets robotic manipulation with action prediction.

10. Learning to model the world with language

URL: View paper

Brief Assessment

Language World Models[53] focuses on world model learning for predicting future text and image representations in embodied agents, not on unified vision-language-action architectures for robotic manipulation with discrete token representations across all modalities.

Contribution 3: State-of-the-art performance on robotic manipulation benchmarks

Description: UniVLA achieves new state-of-the-art results on multiple simulation benchmarks including CALVIN, LIBERO, and SimplerEnv-Bridge, significantly outperforming prior methods. The model also demonstrates effective transfer to real ALOHA platform and autonomous driving scenarios, highlighting its potential for generalist embodied intelligence.

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. Vita-vla: Efficiently teaching vision-language models to act via action expert distillation

URL: View paper

Brief Assessment

Vita-VLA[63] focuses on distilling action knowledge from small action models into VLMs, achieving strong results on LIBERO and CALVIN. However, it does not refute UniVLA's novelty claims about unified vision-language-action modeling with world model post-training, as Vita-VLA[63] uses a different architectural approach (distillation-based) rather than unified token-based autoregressive modeling with video-based world model learning.

2. Predictive inverse dynamics models are scalable learners for robotic manipulation

URL: View paper

Prior Art Analysis

Predictive Inverse Dynamics[64] demonstrates that prior work achieved state-of-the-art results on the same benchmarks (CALVIN, LIBERO) before UniVLA. The candidate paper reports achieving an average length of 4.28 on CALVIN ABC-D and 87.7% average success rate on LIBERO-long. These results were published in December 2024 (arxiv:2412.15109v1), establishing that similar or superior performance on these specific benchmarks was achieved by other methods prior to UniVLA's claims. The candidate's results directly challenge UniVLA's novelty claim of being first to achieve state-of-the-art on these benchmarks.

Evidence

Evidence 1 - **Rationale**: This shows Predictive Inverse Dynamics[64] achieved 87.7% on LIBERO-long (78.7% + 9%), demonstrating competitive state-of-the-art performance on LIBERO before UniVLA's publication, though UniVLA's 95.5% is higher. - **Original**: our approach sets new state-of-the-art results across several widely used simulation benchmarks, including calvin, libero, and simplenv-bridge, substantially outperforming prior methods. for example, univla achieves 95.5% average success rate on libero benchmark, surpassing π0-fast's 85.5%. - **Candidate**: as presented in table 1, our policy achieved an average success rate of 78.7% without pre-training. after pre-training, the success rate increases by 9%, significantly outperforming the baselines.

Evidence 2 - **Rationale**: Both papers target the same benchmarks (CALVIN and LIBERO) for evaluation, with Predictive Inverse Dynamics[64] published in December 2024, demonstrating that achieving strong results on these specific benchmarks was not novel to UniVLA. - **Original**: experiments across multiple simulation benchmarks, including calvin mees et al. (2022b), libero liu et al. (2023), and simplerenv li et al. (2024d), demonstrating clear performance improvements over existing methods. - **Candidate**: we conduct experiments on two simulation benchmarks libero-long (liu et al., 2024), calvin abc-d (mees et al., 2022). our aim is to answer: 1) how does our method perform on challenging simulation benchmarks?

3. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge

URL: View paper

Prior Art Analysis

DreamVLA[59] demonstrates that it achieves state-of-the-art performance on the same benchmarks (CALVIN, LIBERO) that the original paper claims as novel achievements. Specifically, DreamVLA[59] reports 4.44 average length on CALVIN ABC-D (compared to UniVLA's 4.41) and 92.6% average on LIBERO (compared to UniVLA's 95.5%). While UniVLA performs slightly better on LIBERO, DreamVLA[59] outperforms on CALVIN and was published as a preprint with arxiv timestamp 2025, suggesting concurrent or prior work. The candidate paper's comprehensive comparison table directly includes UniVLA's results, demonstrating that similar state-of-the-art claims on these benchmarks were being made by multiple concurrent works, thereby challenging the novelty of UniVLA's achievement claim.

Evidence

Evidence 1 - **Rationale**: DreamVLA[59] reports achieving 92.6% average on LIBERO (from Table 2 context), demonstrating that multiple methods were achieving high performance on this benchmark concurrently, challenging the uniqueness of UniVLA's state-of-the-art claim. - **Original**: univla achieves 95.5% average success rate on libero benchmark - **Candidate**: for the libero benchmark [122], dreamvla exhibits better or comparable ability across all tracks compared to previous approaches by future world knowledge prediction as shown in table 2.

Evidence 2 - **Rationale**: DreamVLA[59] explicitly claims to set 'a new state of the art' on CALVIN with 4.44 average length, which exceeds UniVLA's reported 4.41 on the same ABC-D benchmark, demonstrating that the state-of-the-art claim was not uniquely held by UniVLA. - **Original**: our approach sets new state-of-the-art results across several widely used simulation benchmarks, including calvin, libero, and simplenv-bridge, substantially outperforming prior methods. - **Candidate**: dreamvla sets a new state of the art on the calvin abc-d benchmark (4.44 average task length), outperforming prior methods by up to 3.5% on the simulation platform

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

URL: View paper

Brief Assessment

Fine-tuning VLA[58] focuses on optimizing fine-tuning strategies for vision-language-action models on specific benchmarks (LIBERO, ALOHA), while the original paper presents a unified architecture for vision-language-action modeling with world model pretraining. The candidate addresses fine-tuning methodology rather than the unified token-based architecture and world model approach that are central to the original contribution.

5. Unified Diffusion VLA: Vision-Language-Action Model via Joint Discrete Denoising Diffusion Process

URL: View paper

Prior Art Analysis

Unified Diffusion VLA[47] demonstrates that prior work achieved state-of-the-art results on the same benchmarks (CALVIN, LIBERO, SimplerEnv) before the original paper. The candidate explicitly states achieving 'state-of-the-art performance on benchmarks such as calvin, libero, and simplerenv' using their diffusion-based approach, which represents a different technical paradigm that was published and achieved these results on the identical benchmark suite.

Evidence

Evidence 1 - **Rationale**: The candidate describes 'recent work' on unified VLAs that jointly understand, generate, and act, indicating that unified multimodal VLA approaches with strong benchmark performance existed as prior work in the field. - **Original**: our model achieves state-of-the-art performance on several simulated benchmarks (calvin, libero, and simplerenv-bridge) - **Candidate**: vision-language-action (vla) models aim to understand natural language instructions and visual observations and to execute corresponding actions as an embodied agent. recent work integrates future images into the understanding-acting loop, yielding unified vlas that jointly understand, generate, and...

6. Molmoact: Action reasoning models that can reason in space

URL: View paper

Brief Assessment

MolmoAct[61] focuses on action reasoning models with spatial reasoning capabilities (depth perception tokens, visual reasoning traces) rather than unified vision-language-action modeling. While both achieve strong results on CALVIN and LIBERO, they represent fundamentally different architectural paradigms and cannot refute each other's novelty claims.

7. RoboBERT: An End-to-end Multimodal Robotic Manipulation Model

URL: View paper

Brief Assessment

RoboBERT[62] focuses on a two-stage training paradigm with frozen vision encoders and CNN-based diffusion policies, achieving strong results on CALVIN but not addressing LIBERO or SimplerEnv-Bridge benchmarks mentioned in the original contribution.

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

URL: View paper

Brief Assessment

BitVLA[66] focuses on 1-bit quantization for efficient deployment on resource-constrained devices, not on achieving state-of-the-art performance through unified vision-language-action modeling or world model training. The candidate addresses a different research problem (model compression) rather than challenging the novelty of UniVLA's architectural contributions or training methodology.

9. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals

URL: View paper

Brief Assessment

Multimodal Diffusion Transformer[65] focuses on learning from multimodal goals (language and images) with sparse annotations, while UniVLA emphasizes unified vision-language-action modeling with world model post-training. The candidate does not challenge UniVLA's novelty claims about state-of-the-art performance on CALVIN, LIBERO, and SimplerEnv-Bridge benchmarks.

10. Robomm: All-in-one multimodal large model for robotic manipulation

URL: View paper

Brief Assessment

RoboMM[60] focuses on multi-dataset joint training with 3D spatial alignment and occupancy supervision, while UniVLA emphasizes unified vision-language-action token modeling with world model post-training. The architectural approaches and training paradigms differ fundamentally.

Appendix: Text Similarity Detection

Textual similarity detection checked 29 papers and found 3 similarity segment(s) across 2 paper(s).

The following **2 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

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

Detected in: Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

2. Unified Diffusion VLA: Vision-Language-Action Model via Joint Discrete Denoising Diffusion Process

Detected in: Contribution: contribution_1, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Unified Vision-Language-Action Model View paper
- [1] Fast: Efficient action tokenization for vision-language-action models View paper
- [2] Coa-vla: Improving vision-language-action models via visual-text chain-of-affordance View paper
- [3] Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action View paper
- [4] CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models View paper
- [5] Physical autoregressive model for robotic manipulation without action pretraining View paper
- [6] WorldVLA: Towards Autoregressive Action World Model View paper
- [7] OpenDriveVLA: Towards End-to-end Autonomous Driving with Large Vision Language Action Model View paper
- [8] OmniJARVIS: Unified Vision-Language-Action Tokenization Enables Open-World Instruction Following Agents View paper
- [9] Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers View paper
- [10] Spec-VLA: Speculative Decoding for Vision-Language-Action Models with Relaxed Acceptance View paper
- [11] DeepThinkVLA: Enhancing Reasoning Capability of Vision-Language-Action Models View paper
- [12] HybridVLA: Collaborative Autoregression and Diffusion in a Unified Vision-Language-Action Model View paper
- [13] AutoVLA: A Vision-Language-Action Model for End-to-End Autonomous Driving with Adaptive Reasoning and Reinforcement Fine-Tuning View paper
- [14] Flowvla: Visual chain of thought-based motion reasoning for vision-language-action models View paper
- [15] DiffusionVLA: Scaling Robot Foundation Models via Unified Diffusion and Autoregression View paper
- [16] Think Twice, Act Once: Token-Aware Compression and Action Reuse for Efficient Inference in Vision-Language-Action Models View paper
- [17] CronusVLA: Towards Efficient and Robust Manipulation via Multi-Frame Vision-Language-Action Modeling View paper
- [18] Pure vision language action (vla) models: A comprehensive survey View paper
- [19] PD-VLA: Accelerating Vision-Language-Action Model Integrated with Action Chunking via Parallel Decoding View paper
- [20] FASTer: Toward Efficient Autoregressive Vision Language Action Modeling via Neural Action Tokenization View paper
- [21] DriveVLA-W0: World Models Amplify Data Scaling Law in Autonomous Driving View paper
- [22] VQ-VLA: Improving Vision-Language-Action Models via Scaling Vector-Quantized Action Tokenizers View paper
- [23] InstructSeq: Unifying Vision Tasks with Instruction-conditioned Multi-modal Sequence Generation View paper
- [24] A Survey on Vision-Language-Action Models: An Action Tokenization Perspective View paper
- [25] Dream-VL & Dream-VLA: Open Vision-Language and Vision-Language-Action Models with Diffusion Language Model Backbone View paper
- [26] LatBot: Distilling Universal Latent Actions for Vision-Language-Action Models View paper
- [27] VT-LVLM-AR: A Video-Temporal Large Vision-Language Model Adapter for Fine-Grained Action Recognition in Long-Term Videos View paper
- [28] HybridVLA: Collaborative Diffusion and Autoregression in a Unified Vision-Language-Action Model View paper
- [29] Actra: Optimized Transformer Architecture for Vision-Language-Action Models in Robot Learning View paper
- [30] FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving View paper
- [31] Look, Zoom, Understand: The Robotic Eyeball for Embodied Perception View paper
- [32] Towards Unified Token Learning for Vision-Language Tracking View paper
- [33] Toward Unified Token Learning for Vision-Language Tracking View paper

- [34] OccLLaMA: An Occupancy-Language-Action Generative World Model for Autonomous Driving View paper
- [35] Meaning Representations from Trajectories in Autoregressive Models View paper
- [36] Selftok-Zero: Reinforcement Learning for Visual Generation via Discrete and Autoregressive Visual Tokens View paper
- [37] From Language to Action Streams: Bridging LLM Autoregression for Long-Horizon Robot Action Prediction View paper
- [38] Stable-FAST: Stabilizing Inference of Autoregressive Vision-Language-Action Models View paper
- [39] Palm-e: An embodied multimodal language model View paper
- [40] XR-1: Towards Versatile Vision-Language-Action Models via Learning Unified Vision-Motion Representations View paper
- [41] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control View paper
- [42] A survey on vision-language-action models for embodied ai View paper
- [43] LoHoVLA: A Unified Vision-Language-Action Model for Long-Horizon Embodied Tasks View paper
- [44] Multimodal fusion and vision-language models: A survey for robot vision View paper
- [45] Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation View paper
- [46] ShowUI: One Vision-Language-Action Model for GUI Visual Agent View paper
- [47] Unified Diffusion VLA: Vision-Language-Action Model via Joint Discrete Denoising Diffusion Process View paper
- [48] Imagine-2-Drive: Leveraging High-Fidelity World Models via Multi-Modal Diffusion Policies View paper
- [49] 3D-VLA: A 3D Vision-Language-Action Generative World Model View paper
- [50] Spatial-Temporal Aware Visuomotor Diffusion Policy Learning View paper
- [51] Pre-training contextualized world models with in-the-wild videos for reinforcement learning View paper
- [52] A step toward world models: A survey on robotic manipulation View paper
- [53] Learning to model the world with language View paper
- [54] GenRL: Multimodal-foundation world models for generalization in embodied agents View paper
- [55] Can World Models Benefit VLMs for World Dynamics? View paper
- [56] Multimodal foundation world models for generalist embodied agents View paper
- [57] Merlot: Multimodal neural script knowledge models View paper
- [58] Fine-tuning vision-language-action models: Optimizing speed and success View paper
- [59] Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge View paper
- [60] Robomm: All-in-one multimodal large model for robotic manipulation View paper
- [61] Molmoact: Action reasoning models that can reason in space View paper
- [62] RoboBERT: An End-to-end Multimodal Robotic Manipulation Model View paper
- [63] Vita-vla: Efficiently teaching vision-language models to act via action expert distillation View paper
- [64] Predictive inverse dynamics models are scalable learners for robotic manipulation View paper
- [65] Multimodal diffusion transformer: Learning versatile behavior from multimodal goals View paper
- [66] BitVLA: 1-bit Vision-Language-Action Models for Robotics Manipulation View paper