Novelty Assessment Report

Paper: FASTer: Toward Powerful and Efficient Autoregressive Vision–Language–Action Models with Learnable Action Tokenizer and Block-wise Decoding

PDF URL: https://openreview.net/pdf?id=k6nTUFoqeT

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Autoregressive vision-language-action (VLA) models have recently demonstrated strong capabilities in robotic manipulation. However, their core process of action tokenization often involves a trade-off between reconstruction fidelity and inference efficiency. We introduce \textbf{FASTer}, a unified framework for efficient and generalizable robot learning that integrates a learnable tokenizer with an autoregressive policy built upon it. FASTerVQ encodes action chunks as single-channel images, capturing global spatio-temporal dependencies while maintaining a high compression ratio. FASTerVLA builds on this tokenizer with block-wise autoregressive decoding and a lightweight action expert, achieving both faster inference and higher task performance. Extensive experiments across simulated and real-world benchmarks show that FASTerVQ delivers superior reconstruction quality, high token utilization, and strong cross-task and cross-embodiment generalization, while FASTerVLA further improves overall capability, surpassing previous state-of-the-art VLA models in both inference speed and task performance.

Core Task Landscape

This paper addresses: **Action Tokenization for Vision-Language-Action Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **24 categories**.

Taxonomy Overview

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

- **Action Tokenization Methods**
- **VLA Architecture and Multimodal Integration**
- **VLA Training and Adaptation**
- **VLA Inference Optimization and Efficiency**
- **VLA Temporal Modeling and Multi-Frame Processing**
- **VLA Evaluation and Benchmarking**
- **VLA Surveys and Methodological Reviews**
- **Related Vision-Language Applications**

Complete Taxonomy Tree

- Action Tokenization for Vision-Language-Action Models Survey Taxonomy
- Action Tokenization Methods
  - Discrete Action Tokenization
  - Vector Quantization-Based Tokenization ★ (3 papers)
    - [0] FASTer: Toward Powerful and Efficient Autoregressive Vision–Language–Action Models with Learnable Action Tokenizer and Block-wise Decoding (Anon et al., 2026) View paper
    - [36] FASTer: Toward Efficient Autoregressive Vision Language Action Modeling via Neural Action Tokenization (Yicheng Liu, 2025) View paper
    - [39] VQ-VLA: Improving Vision-Language-Action Models via Scaling Vector-Quantized Action Tokenizers (Wang Yating, 2025) View paper
  - Adaptive Grid and Spatial Discretization (2 papers)
    - [2] Spatialvla: Exploring spatial representations for visual-language-action model (Qu, 2025) View paper
    - [4] FAST: Efficient Action Tokenization for Vision-Language-Action Models (Karl Pertsch, 2025) View paper
  - B-Spline and Trajectory Encoding (1 papers)
    - [33] BEAST: Efficient Tokenization of B-Splines Encoded Action Sequences for Imitation Learning (Zhou Hong-Yi, 2025) View paper
  - Latent and Continuous Action Representations
  - Latent Action Modeling (2 papers)
    - [14] Behavior Generation with Latent Actions (Lee Seung-Jae, 2024) View paper
    - [47] From Observation to Action: Latent Action-based Primitive Segmentation for VLA Pre-training in Industrial Settings (Jiajie Zhang, 2025) View paper
  - Diffusion-Based Action Generation (3 papers)
    - [16] AsyncVLA: Asynchronous Flow Matching for Vision-Language-Action Models (Yuhua Jiang, 2025) View paper
    - [31] LLaDA-VLA: Vision Language Diffusion Action Models (Wen Yu-qing, 2025) View paper
    - [32] Discrete Diffusion VLA: Bringing Discrete Diffusion to Action Decoding in Vision-Language-Action Policies (Liang Zhixuan, 2025) View paper
  - Text-as-Action Representation (2 papers)

- [43] VLA-0: Building State-of-the-Art VLAs with Zero Modification (Goyal Ankit, 2025) View paper
- [46] Actions as Language: Fine-Tuning VLMs into VLAs Without Catastrophic Forgetting (Wu, 2025) View paper

- VLA Architecture and Multimodal Integration
  - Unified Multimodal VLA Architectures (3 papers)
  - [1] Unified Vision-Language-Action Model (Wang, 2025) View paper
  - [26] Omnijarvis: Unified vision-language-action tokenization enables open-world instruction following agents (Shaofei Cai, 2024) View paper
  - [30] How to Build a Pre-trained Multimodal model for Simultaneously Chatting and Decision-making? (Zuojin Tang, 2024) View paper
  - Spatial and 3D-Enhanced VLA Models (3 papers)
  - [6] 3D-VLA: A 3D Vision-Language-Action Generative World Model (Qiu Xiaowen, 2024) View paper
  - [9] From Spatial to Actions: Grounding Vision-Language-Action Model in Spatial Foundation Priors (Zhang, 2025) View paper
  - [48] Rethinking 3D Robotic Perception: Elastic Voxel Representation with Splatting Distillation (Shaohui Pan, 2025) View paper
  - Object-Centric and Attention-Based Tokenization (2 papers)
  - [3] Focusing on What Matters: Object-Agent-centric Tokenization for Vision Language Action models (Dijkman, 2025) View paper
  - [20] ShowUI: One Vision-Language-Action Model for GUI Visual Agent (Kevin Qinghong Lin, 2024) View paper
  - Hierarchical and Multi-Level VLA Frameworks (3 papers)
  - [27] LoHoVLA: A Unified Vision-Language-Action Model for Long-Horizon Embodied Tasks (Yang Yi, 2025) View paper
  - [34] Asynchronous Fast-Slow Vision-Language-Action Policies for Whole-Body Robotic Manipulation (Teqiang Zou, 2025) View paper
  - [50] Deep Learning-Powered Natural Language Interface in Intelligent Pest Extermination Robotics (Wenhua Gan, 2025) View paper
  - Reasoning-Enhanced VLA Models (3 papers)
  - [5] CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models (Zhao Qingqing, 2025) View paper
  - [19] Coa-vla: Improving vision-language-action models via visual-text chain-of-affordance (J Li, 2025) View paper
  - [44] Robotic VLA Benefits from Joint Learning with Motion Image Diffusion (Yu Fang, 2025) View paper
- VLA Training and Adaptation
  - Pre-training and Foundation Model Adaptation (3 papers)
  - [7] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control (Brohan, 2023) View paper
  - [10] Being-h0: vision-language-action pretraining from large-scale human videos (Luo Hao, 2025) View paper
  - [18] Enhancing generalization in vision-language-action models by preserving pretrained representations (Gopalkrishnan, 2025) View paper
  - Fine-Tuning and Task Adaptation (2 papers)
  - [8] Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success (Moo Kim, 2025) View paper
  - [35] VITA-VLA: Efficiently Teaching Vision-Language Models to Act via Action Expert Distillation (Fu, 2025) View paper
  - Cross-Embodiment and Multi-Robot Transfer (1 papers)
  - [12] Embodiment Transfer Learning for Vision-Language-Action Models (Li Chengmeng, 2025) View paper
  - Reinforcement Learning and Reward Modeling (2 papers)
  - [21] A vision-language-action-critic model for robotic real-world reinforcement learning (Zhai Shaopeng, 2025) View paper
  - [49] Enhancing Vision-Language Model Training with Reinforcement Learning in Synthetic Worlds for Real-World Success (Dereka, 2025) View paper
- VLA Inference Optimization and Efficiency
  - Token Pruning and Compression (3 papers)
  - [11] Think Twice, Act Once: Token-Aware Compression and Action Reuse for Efficient Inference in Vision-Language-Action Models (Tan Xudong, 2025) View paper
  - [37] VLA-Pruner: Temporal-Aware Dual-Level Visual Token Pruning for Efficient Vision-Language-Action Inference (Ziyan Liu, 2025) View paper
  - [38] Token Expand-Merge: Training-Free Token Compression for Vision-Language-Action Models (Yifan Ye, 2025) View paper
  - Token Caching and Reuse (1 papers)
  - [13] Vla-cache: Towards efficient vision-language-action model via adaptive token caching in robotic manipulation (Siyu Xu, 2025) View paper
  - Speculative and Parallel Decoding (1 papers)
  - [24] Spec-VLA: Speculative Decoding for Vision-Language-Action Models with Relaxed Acceptance (Songsheng Wang, 2025) View paper
  - Lightweight and Efficient VLA Architectures (2 papers)
  - [22] CogVLA: Cognition-Aligned Vision-Language-Action Model via Instruction-Driven Routing & Sparsification (Li Wei, 2025) View paper
  - [40] SwiftVLA: Unlocking Spatiotemporal Dynamics for Lightweight VLA Models at Minimal Overhead (Chaojun Ni, 2025) View paper
- VLA Temporal Modeling and Multi-Frame Processing (2 papers)
  - [28] CronusVLA: Towards Efficient and Robust Manipulation via Multi-Frame Vision-Language-Action Modeling (Li Hao, 2025) View paper
- VLA Evaluation and Benchmarking
  - Cross-Platform and Scaling Analysis (1 papers)
  - [29] Cross-Platform Scaling of Vision-Language-Action Models from Edge to Cloud GPUs (Amir Taherin, 2025) View paper
  - Test-Time Scaling and Sampling (1 papers)
  - [25] Verifier-free Test-Time Sampling for Vision Language Action Models (Kim Dong-Young, 2025) View paper
- VLA Surveys and Methodological Reviews (4 papers)
  - [15] What Matters in Employing Vision Language Models for Tokenizing Actions in Robot Control? (N Dorka, 2024) View paper
  - [17] Recipe for Vision-Language-Action Models in Robotic Manipulation: A Survey (Tomohiro Motoda, 2025) View paper
  - [23] Vision-Language-Action Models: Foundations, Techniques and Applications (Yan Li, 2025) View paper
  - [42] A Survey on Vision-Language-Action Models: An Action Tokenization Perspective (Zhong, 2025) View paper

- Related Vision-Language Applications (1 papers)
  - [45] Leveraging Vision-Language Large Models for Interpretable Video Action Recognition with Semantic Tokenization (Peng Jingwei, 2025) View paper

Narrative

Core task: action tokenization for vision-language-action models. Vision-language-action (VLA) models integrate visual perception, language understanding, and robotic control by converting continuous action spaces into discrete or learned representations that can be processed alongside text and image tokens. The field's taxonomy reveals several major branches: Action Tokenization Methods explore how to represent actions—whether through discrete vector quantization (VQ-VLA[39], FAST[4], FASTer[0]), latent embeddings (Latent Actions[14]), or semantic abstractions (Semantic Tokenization[45])—while VLA Architecture and Multimodal Integration addresses how to fuse vision and language backbones with action prediction heads (RT-2[7], 3D-VLA[6], SpatialVLA[2]). Training and Adaptation branches cover fine-tuning strategies (Fine-Tuning VLA[8], Preserving Pretrained[18]) and cross-embodiment transfer (Embodiment Transfer[12]), whereas Inference Optimization focuses on efficiency gains through caching (VLA-Cache[13]), pruning (VLA-Pruner[37]), and asynchronous processing (AsyncVLA[16]). Temporal Modeling examines multi-frame reasoning, Evaluation establishes benchmarks, and Surveys (VLA Recipe Survey[17], Action Tokenization Survey[42]) synthesize methodological insights.

A particularly active line of work centers on discrete action tokenization via vector quantization, where methods like FAST[4] and VQ-VLA[39] learn compact codebooks to represent continuous actions as discrete tokens compatible with language model architectures. FASTer[0] sits squarely within this cluster, extending vector quantization-based tokenization to improve codebook utilization and reconstruction fidelity. Nearby, FASTer Neural[36] explores neural variants of the same approach, while Object-Agent Tokenization[3] emphasizes object-centric representations that complement action discretization. These discrete tokenization strategies contrast with latent action approaches (Latent Actions[14], CronusVLA Latent[41]) that embed actions in continuous spaces, and with methods that treat actions as natural language sequences (Actions as Language[46]). The trade-offs revolve around expressiveness versus compatibility with pretrained language models, with discrete tokenization offering seamless integration at the cost of potential quantization error, a challenge FASTer[0] addresses through refined codebook learning.

Related Works in Same Category

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

1. FASTer: Toward Efficient Autoregressive Vision Language Action Modeling via Neural Action Tokenization

Authors: Yicheng Liu, Shiduo Zhang, Zibin Dong, Baijun Ye, Tianyuan Yuan, et al. (15 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Autoregressive vision-language-action (VLA) models have recently demonstrated strong capabilities in robotic manipulation. However, their core process of action tokenization often involves a trade-off between reconstruction fidelity and inference efficiency. We introduce FASTer, a unified framework for efficient and generalizable robot learning that integrates a learnable tokenizer with an autoregressive policy built upon it. FASTerVQ encodes action chunks as single-channel images, capturing glo...

⚠ Similarity Notice

This paper appears to be the same work as the original paper. Both share an identical title ('FASTer: Toward Efficient Autoregressive Vision Language Action Modeling via Neural Action Tokenization'), describe the same FASTerVQ tokenizer and FASTerVLA model architecture, and present identical technical contributions including residual vector quantization, block-wise autoregressive decoding, and the same experimental results across multiple benchmarks.

---

2. VQ-VLA: Improving Vision-Language-Action Models via Scaling Vector-Quantized Action Tokenizers

Authors: Wang Yating, Zhu, Haoyi, Yating Wang, Liu, et al. (15 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

In this paper, we introduce an innovative vector quantization based action tokenizer built upon the largest-scale action trajectory dataset to date, leveraging over 100 times more data than previous approaches. This extensive dataset enables our tokenizer to capture rich spatiotemporal dynamics, resulting in a model that not only accelerates inference but also generates smoother and more coherent action outputs. Once trained, the tokenizer can be seamlessly adapted to a wide range of downstream ...

Relationship Analysis

Both papers belong to the Vector Quantization-Based Tokenization category, employing VQ/codebook learning to discretize action sequences for vision-language-action models. They overlap in using residual vector quantization (RVQ) and transformer-based architectures to achieve efficient action tokenization with high compression ratios. However, FASTer focuses on a unified framework combining a learnable tokenizer (FASTerVQ) with block-wise autoregressive decoding and action experts for efficient inference, while VQ-VLA emphasizes scaling the tokenizer training on over 100x more data (including synthetic trajectories) to improve zero-shot generalization and demonstrates that synthetic-real domain gaps are minimal for action tokenization.

Contributions Analysis

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

Contribution 1: FASTerVQ: Learnable Action Tokenizer with Transformer-based RVQ

Description: FASTerVQ is a neural action tokenizer that encodes action chunks as single-channel images using transformer-based residual vector quantization. It achieves high compression ratios while maintaining reconstruction fidelity by capturing global spatio-temporal dependencies and modeling actions in both temporal and frequency domains.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. OmniSAT: Compact Action Token, Faster Auto Regression

URL: View paper

Brief Assessment

OmniSAT[59] focuses on B-spline encoding with multi-stage residual quantization applied to position/rotation/gripper subspaces separately, whereas FASTerVQ uses transformer-based residual vector quantization with action patchification and dual-domain (temporal + frequency) reconstruction losses. The architectural approaches and design philosophies differ substantially.

---

2. Towards Generally Intelligent Robots That Simply Work Everywhere

URL: View paper

Brief Assessment

Generally Intelligent Robots[62] mentions using residual VQ-VAE for action tokenization but provides insufficient technical detail about the architecture, training methodology, or transformer-based design to challenge FASTerVQ's novelty claims regarding its specific transformer-based RVQ implementation with hybrid encoders and dual-domain reconstruction.

---

3. Causal Motion Tokenizer for Streaming Motion Generation

URL: View paper

Brief Assessment

Causal Motion Tokenizer[58] focuses on human motion generation from text for animation/robotics, using RVQ-VAE with causal convolutions for streaming motion synthesis. FASTerVQ targets robotic action tokenization for vision-language-action models with different architectural choices and application domains.

---

4. Baku: An efficient transformer for multi-task policy learning

URL: View paper

Brief Assessment

Baku[57] focuses on multi-task policy learning architecture for robotics manipulation, not on action tokenization methods. The paper does not discuss vector quantization or residual quantization approaches for action encoding.

---

5. Behavior Generation with Latent Actions

URL: View paper

Prior Art Analysis

Latent Actions[14] demonstrates that prior work exists on using vector quantization for action tokenization in behavior modeling. The candidate paper presents VQ-BET, which tokenizes continuous actions using hierarchical vector quantization for behavior generation. Both papers address the same core problem of discretizing continuous action spaces using vector quantization techniques, with the candidate explicitly describing a 'hierarchical vector quantization module' for 'tokenizing continuous actions.' This establishes that the concept of using VQ-based methods for action tokenization predates the original paper's FASTerVQ contribution.

Evidence

Evidence 1 - **Rationale**: Both papers describe using vector quantization to tokenize/encode actions. The candidate's hierarchical VQ module serves the same fundamental purpose as FASTerVQ's encoding of action chunks, demonstrating prior work on VQ-based action tokenization. - **Original**: FASTerVQ encodes action chunks as single-channel images, capturing global spatio-temporal dependencies while maintaining a high compression ratio. - **Candidate**: vq-bet augments bet by tokenizing continuous actions with a hierarchical vector quantization module.

Evidence 2 - **Rationale**: The candidate paper explicitly discusses the problem of discretizing actions and presents VQ as a solution to limitations of prior discretization methods, establishing that action discretization/tokenization was an active research area before the original paper. - **Original**: we propose fastervq, a compact and high-compression-ratio action tokenizer that combines transformer-based residual vector quantization (rvq) with a lightweight mixture mechanism. - **Candidate**: a recent class of models called behavior transformers (bet) addresses this by discretizing actions using k-means clustering to capture different modes. however, k-means struggles to scale for high-dimensional action spaces or long sequences, and lacks gradient information

Evidence 3 - **Rationale**: The candidate demonstrates that VQ-based action tokenization was already being compared against other state-of-the-art methods, indicating this was an established approach in the field rather than a novel contribution. - **Original**: fastervq applies dct and l1 reconstruction losses and adopts rvq, encoding actions into nc code levels; each level can be reshaped into a ch x ca tensor. - **Candidate**: vq-bet improves on state-of-the-art models such as bet and diffusion policies. importantly, we demonstrate vq-bet's improved ability to capture behavior modes while accelerating inference speed 5x over diffusion policies.

---

6. VersatileMotion: A Unified Framework for Motion Synthesis and Comprehension

URL: View paper

Brief Assessment

VersatileMotion[63] focuses on human motion synthesis and understanding using VQ-VAE with flow matching for motion tokenization, not robotic action sequences. The domain (human motion vs. robotic manipulation), application (motion generation/understanding vs. robot control), and technical approach differ fundamentally from FASTerVQ's robotics-specific action tokenization.

---

7. Grounding multimodal large language models in actions

URL: View paper

Brief Assessment

Grounding Actions[60] focuses on adapting multimodal LLMs for action generation across different embodiments using various action space adapters, including learned tokenization methods. However, it does not specifically propose FASTerVQ's unique architecture combining transformer-based residual vector quantization with action patchification for robotics, nor does it address the specific compression-fidelity trade-offs and cross-embodiment generalization claims made by the original paper.

---

8. HOIGPT: Learning Long Sequence Hand-Object Interaction with Language Models

URL: View paper

Brief Assessment

HOIGPT[61] focuses on hand-object interaction tokenization for 3D mesh sequences using a hand-object decomposed VQ-VAE, not robotic action sequences. The domain (hand-object interaction vs. robotic manipulation) and application context differ fundamentally from FASTerVQ's robotics focus.

Contribution 2: Block-wise Autoregressive Decoding with Lightweight Action Expert

Description: The method introduces block-wise autoregressive decoding that predicts multiple tokens in parallel within each block, reducing inference steps. A lightweight action expert module is added to bridge the modality gap between linguistic reasoning and continuous control while maintaining parameter efficiency.

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. WorldVLA: Towards Autoregressive Action World Model

URL: View paper

Brief Assessment

WorldVLA[51] focuses on unifying action and world models for joint image-action generation, not on block-wise autoregressive decoding or lightweight action experts for efficient token prediction as in the original paper.

---

2. FASTer: Toward Efficient Autoregressive Vision Language Action Modeling via Neural Action Tokenization

URL: View paper

Brief Assessment

FASTer Neural[36] focuses on vision-language-action models for robotic manipulation with block-wise decoding and action experts, not general RL frameworks for agent training across diverse environments.

---

3. LLaDA-VLA: Vision Language Diffusion Action Models

URL: View paper

Brief Assessment

LLaDA-VLA[31] uses masked diffusion models with hierarchical action-structured decoding, not block-wise autoregressive decoding. The candidate's approach is fundamentally different from the original paper's autoregressive framework with parallel token prediction.

---

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

URL: View paper

Brief Assessment

DreamVLA[54] uses a block-wise structured attention mechanism for masking mutual attention between dynamic, spatial and semantic information during training, and employs a diffusion-based transformer for action modeling. This differs fundamentally from the original paper's block-wise autoregressive decoding that predicts multiple action tokens in parallel to reduce inference steps, combined with a lightweight action expert for bridging modality gaps in continuous control.

---

5. Carp: Visuomotor policy learning via coarse-to-fine autoregressive prediction

URL: View paper

Prior Art Analysis

Carp[53] demonstrates that block-wise autoregressive decoding for action prediction was proposed prior to the ORIGINAL paper. Both papers employ a coarse-to-fine autoregressive prediction mechanism that predicts multiple tokens in parallel within blocks, reducing inference steps. Carp[53] explicitly describes a 'coarse-to-fine autoregressive prediction' approach where 'all distributions over the lk tokens in rk will be generated in parallel, with the coarse-to-fine dependency ensured by a block-wise causal attention mask.' This directly parallels the ORIGINAL paper's claim of introducing 'block-wise autoregressive decoding that predicts multiple tokens in parallel within each block.' Additionally, both papers address the modality gap between vision-language understanding and continuous control through specialized architectural components.

Evidence

Evidence 1 - **Rationale**: Both papers claim to introduce a novel autoregressive prediction mechanism that refines sequences progressively while maintaining efficiency. Carp[53] explicitly describes this as a core contribution predating the ORIGINAL paper. - **Original**: we introduce block-wise autoregressive decoding for efficient action-token modeling, together with a share structured action expert that aligns with the vlm backbone. these components under review as a conference paper at iclr 2026 jointly unleash the capability of autoregressive vlas, allowing them... - **Candidate**: coarse-to-fine autoregressive prediction: this mechanism refines action sequences in the latent space using cross-entropy loss with relaxed markovian assumptions during iterations, achieving dm-like performance with high efficiency and comparable multi-modal behavior.

Evidence 2 - **Rationale**: Both papers describe parallel token prediction within blocks using causal attention masks. Carp[53]'s description of generating 'all distributions over the lk tokens in rk will be generated in parallel' directly corresponds to the ORIGINAL paper's 'efficient parallel prediction within each block.' - **Original**: building upon fastervq, fastervla employs block-wise autoregressive decoding that leverages the structured latent space and the partial independence of tokens along the action dimension, enabling efficient parallel prediction within each block while maintaining coherent spatio-temporal structure. - **Candidate**: during the k-th autoregressive step, all distributions over the lk tokens in rk will be generated in parallel, with the coarse-to-fine dependency ensured by a block-wise causal attention mask

Evidence 3 - **Rationale**: While the architectural details differ, both papers address the fundamental challenge of bridging the gap between high-level representations and continuous action spaces through specialized modules. Carp[53]'s multi-scale tokenization serves a similar bridging function to the ORIGINAL paper's action expert. - **Original**: a lightweight action expert, architecturally aligned with the vlm backbone yet highly parameter-efficient, is introduced to bridge the modality gap between linguistic reasoning and continuous control. - **Candidate**: we propose a novel multi-scale action quantization autoencoder that encodes a sequence of actions into k discrete token maps, r = (r1, r2, . . . ,rk), which are used for training and inference.

Evidence 4 - **Rationale**: Both papers explicitly use block-wise causal attention masks to enable parallel prediction within blocks. This is a key technical mechanism that enables the claimed efficiency improvements in both approaches. - **Original**: we replace the standard causal mask with a block-wise causal mask that allows intra-block attention (fig. 3c). to integrate text and action generation, bar uses two control tokens, (boblk) and (eoblk), to toggle between block-wise and standard ar prediction - **Candidate**: during the k-th autoregressive step, all distributions over the lk tokens in rk will be generated in parallel, with the coarse-to-fine dependency ensured by a block-wise causal attention mask

---

6. Handsonvlm: Vision-language models for hand-object interaction prediction

URL: View paper

Brief Assessment

HandsOnVLM[56] focuses on predicting human hand trajectories for interaction tasks using vision-language models, not on action tokenization or block-wise decoding for robotic control. The technical domains are fundamentally different.

---

7. Pure vision language action (vla) models: A comprehensive survey

URL: View paper

Brief Assessment

Pure VLA Survey[55] is a survey paper that provides a taxonomy and review of VLA methods. It does not present original technical contributions that could refute the novelty of block-wise autoregressive decoding or lightweight action expert modules.

---

8. SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics

URL: View paper

Brief Assessment

SmolVLA[52] focuses on model size reduction and asynchronous inference for affordable robotics, not on block-wise autoregressive decoding or action expert modules for bridging modality gaps.

---

9. FAST: Efficient Action Tokenization for Vision-Language-Action Models

URL: View paper

Brief Assessment

FAST[4] focuses on frequency-space tokenization using discrete cosine transform for action compression, not on block-wise autoregressive decoding mechanisms. The candidate's core contribution is a compression-based tokenization scheme rather than architectural innovations in decoding strategies.

---

10. Unified Vision-Language-Action Model

URL: View paper

Brief Assessment

Unified VLA[1] focuses on unified multimodal token sequences and world modeling from videos, not on block-wise autoregressive decoding or lightweight action expert modules for bridging modality gaps in action prediction.

Contribution 3: Comprehensive Benchmark for Action Tokenization in VLAs

Description: The authors create an extensive evaluation framework spanning multiple real-world robotic platforms and simulated environments to systematically analyze action tokenization methods for vision-language-action models, demonstrating superior performance across diverse embodiments and 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. Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions

URL: View paper

Brief Assessment

Q-Transformer[67] focuses on autoregressive Q-learning for offline RL with action discretization, not on benchmarking action tokenization methods across multiple robotic platforms and simulators as claimed in the original paper.

---

2. Lanmp: A multifaceted mobile manipulation benchmark for robots

URL: View paper

Brief Assessment

LaNMP[68] focuses on benchmarking mobile manipulation tasks with language, navigation, manipulation, and perception capabilities across multiple environments. It does not address action tokenization methods for vision-language-action models, which is the core contribution of the original paper.

---

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

URL: View paper

Brief Assessment

SpatialVLA[2] focuses on spatial representations (ego3D position encoding and adaptive action grids) for robot manipulation rather than benchmarking action tokenization methods. The paper does not systematically evaluate or compare different action tokenization approaches across platforms.

---

4. Perceiver-actor: A multi-task transformer for robotic manipulation

URL: View paper

Brief Assessment

Perceiver-Actor[64] focuses on a voxel-based transformer architecture for multi-task manipulation rather than benchmarking action tokenization methods. The paper does not systematically compare different tokenization approaches across platforms.

---

5. Enhancing generalization in vision-language-action models by preserving pretrained representations

URL: View paper

Brief Assessment

Preserving Pretrained[18] focuses on preserving pretrained VLM representations through dual-encoder design and string-based action tokenization, not on benchmarking action tokenization methods across platforms. The candidate does not establish a comprehensive evaluation framework for comparing different tokenization approaches.

---

6. Behavior Generation with Latent Actions

URL: View paper

Brief Assessment

Latent Actions[14] focuses on behavior generation in decision-making tasks (manipulation, autonomous driving, robotics) rather than systematically benchmarking action tokenization methods across multiple robotic platforms specifically for vision-language-action models.

---

7. VQ-VLA: Improving Vision-Language-Action Models via Scaling Vector-Quantized Action Tokenizers

URL: View paper

Brief Assessment

VQ-VLA[39] focuses on scaling a vector quantization tokenizer with synthetic data for improved performance, rather than establishing a systematic benchmark framework across multiple platforms for evaluating action tokenization methods.

---

8. Action-quantized offline reinforcement learning for robotic skill learning

URL: View paper

Brief Assessment

Action-Quantized Offline[69] focuses on offline RL with action discretization for robotic manipulation, not on benchmarking action tokenization methods across multiple platforms for vision-language-action models. The candidate addresses a different problem domain (offline RL vs. VLA evaluation).

---

9. Discrete policy: Learning disentangled action space for multi-task robotic manipulation

URL: View paper

Brief Assessment

Discrete Policy[66] focuses on vector quantization for multi-task manipulation policies, not on benchmarking action tokenization methods across platforms. The paper does not present a systematic evaluation framework comparing different tokenization approaches.

---

10. The colosseum: A benchmark for evaluating generalization for robotic manipulation

URL: View paper

Brief Assessment

Colosseum[65] focuses on evaluating generalization for robotic manipulation across diverse environments and robots, but does not specifically address action tokenization methods for vision-language-action models, which is the core focus of the original paper's benchmark.

Appendix: Text Similarity Detection

Textual similarity detection checked 27 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 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. FASTer: Toward Efficient Autoregressive Vision Language Action Modeling via Neural Action Tokenization

Detected in: Core Task (sibling), Contribution: contribution_2

⚠ **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] FASTer: Toward Powerful and Efficient Autoregressive Vision–Language–Action Models with Learnable Action Tokenizer and Block-wise Decoding View paper
- [1] Unified Vision-Language-Action Model View paper
- [2] Spatialvla: Exploring spatial representations for visual-language-action model View paper
- [3] Focusing on What Matters: Object-Agent-centric Tokenization for Vision Language Action models View paper
- [4] FAST: Efficient Action Tokenization for Vision-Language-Action Models View paper
- [5] CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models View paper
- [6] 3D-VLA: A 3D Vision-Language-Action Generative World Model View paper
- [7] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control View paper
- [8] Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success View paper
- [9] From Spatial to Actions: Grounding Vision-Language-Action Model in Spatial Foundation Priors View paper
- [10] Being-h0: vision-language-action pretraining from large-scale human videos View paper
- [11] Think Twice, Act Once: Token-Aware Compression and Action Reuse for Efficient Inference in Vision-Language-Action Models View paper
- [12] Embodiment Transfer Learning for Vision-Language-Action Models View paper
- [13] Vla-cache: Towards efficient vision-language-action model via adaptive token caching in robotic manipulation View paper
- [14] Behavior Generation with Latent Actions View paper
- [15] What Matters in Employing Vision Language Models for Tokenizing Actions in Robot Control? View paper
- [16] AsyncVLA: Asynchronous Flow Matching for Vision-Language-Action Models View paper
- [17] Recipe for Vision-Language-Action Models in Robotic Manipulation: A Survey View paper
- [18] Enhancing generalization in vision-language-action models by preserving pretrained representations View paper
- [19] Coa-vla: Improving vision-language-action models via visual-text chain-of-affordance View paper
- [20] ShowUI: One Vision-Language-Action Model for GUI Visual Agent View paper
- [21] A vision-language-action-critic model for robotic real-world reinforcement learning View paper
- [22] CogVLA: Cognition-Aligned Vision-Language-Action Model via Instruction-Driven Routing & Sparsification View paper
- [23] Vision-Language-Action Models: Foundations, Techniques and Applications View paper
- [24] Spec-VLA: Speculative Decoding for Vision-Language-Action Models with Relaxed Acceptance View paper
- [25] Verifier-free Test-Time Sampling for Vision Language Action Models View paper
- [26] Omnijarvis: Unified vision-language-action tokenization enables open-world instruction following agents View paper
- [27] LoHoVLA: A Unified Vision-Language-Action Model for Long-Horizon Embodied Tasks View paper
- [28] CronusVLA: Towards Efficient and Robust Manipulation via Multi-Frame Vision-Language-Action Modeling View paper
- [29] Cross-Platform Scaling of Vision-Language-Action Models from Edge to Cloud GPUs View paper
- [30] How to Build a Pre-trained Multimodal model for Simultaneously Chatting and Decision-making? View paper
- [31] LLaDA-VLA: Vision Language Diffusion Action Models View paper
- [32] Discrete Diffusion VLA: Bringing Discrete Diffusion to Action Decoding in Vision-Language-Action Policies View paper
- [33] BEAST: Efficient Tokenization of B-Splines Encoded Action Sequences for Imitation Learning View paper
- [34] Asynchronous Fast-Slow Vision-Language-Action Policies for Whole-Body Robotic Manipulation View paper
- [35] VITA-VLA: Efficiently Teaching Vision-Language Models to Act via Action Expert Distillation View paper
- [36] FASTer: Toward Efficient Autoregressive Vision Language Action Modeling via Neural Action Tokenization View paper
- [37] VLA-Pruner: Temporal-Aware Dual-Level Visual Token Pruning for Efficient Vision-Language-Action Inference View paper
- [38] Token Expand-Merge: Training-Free Token Compression for Vision-Language-Action Models View paper
- [39] VQ-VLA: Improving Vision-Language-Action Models via Scaling Vector-Quantized Action Tokenizers View paper
- [40] SwiftVLA: Unlocking Spatiotemporal Dynamics for Lightweight VLA Models at Minimal Overhead View paper
- [41] CronusVLA: Transferring Latent Motion Across Time for Multi-Frame Prediction in Manipulation View paper
- [42] A Survey on Vision-Language-Action Models: An Action Tokenization Perspective View paper
- [43] VLA-0: Building State-of-the-Art VLAs with Zero Modification View paper
- [44] Robotic VLA Benefits from Joint Learning with Motion Image Diffusion View paper
- [45] Leveraging Vision-Language Large Models for Interpretable Video Action Recognition with Semantic Tokenization View paper
- [46] Actions as Language: Fine-Tuning VLMs into VLAs Without Catastrophic Forgetting View paper
- [47] From Observation to Action: Latent Action-based Primitive Segmentation for VLA Pre-training in Industrial Settings View paper
- [48] Rethinking 3D Robotic Perception: Elastic Voxel Representation with Splatting Distillation View paper
- [49] Enhancing Vision-Language Model Training with Reinforcement Learning in Synthetic Worlds for Real-World Success View paper
- [50] Deep Learning-Powered Natural Language Interface in Intelligent Pest Extermination Robotics View paper

- [51] WorldVLA: Towards Autoregressive Action World Model View paper
- [52] SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics View paper
- [53] Carp: Visuomotor policy learning via coarse-to-fine autoregressive prediction View paper
- [54] Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge View paper
- [55] Pure vision language action (vla) models: A comprehensive survey View paper
- [56] Handsonvlm: Vision-language models for hand-object interaction prediction View paper
- [57] Baku: An efficient transformer for multi-task policy learning View paper
- [58] Causal Motion Tokenizer for Streaming Motion Generation View paper
- [59] OmniSAT: Compact Action Token, Faster Auto Regression View paper
- [60] Grounding multimodal large language models in actions View paper
- [61] HOIGPT: Learning Long Sequence Hand-Object Interaction with Language Models View paper
- [62] Towards Generally Intelligent Robots That Simply Work Everywhere View paper
- [63] VersatileMotion: A Unified Framework for Motion Synthesis and Comprehension View paper
- [64] Perceiver-actor: A multi-task transformer for robotic manipulation View paper
- [65] The colosseum: A benchmark for evaluating generalization for robotic manipulation View paper
- [66] Discrete policy: Learning disentangled action space for multi-task robotic manipulation View paper
- [67] Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions View paper
- [68] Lanmp: A multifaceted mobile manipulation benchmark for robots View paper
- [69] Action-quantized offline reinforcement learning for robotic skill learning View paper