UNIFIED VISION-LANGUAGE-ACTION MODEL

Anonymous authors
Paper under double-blind review

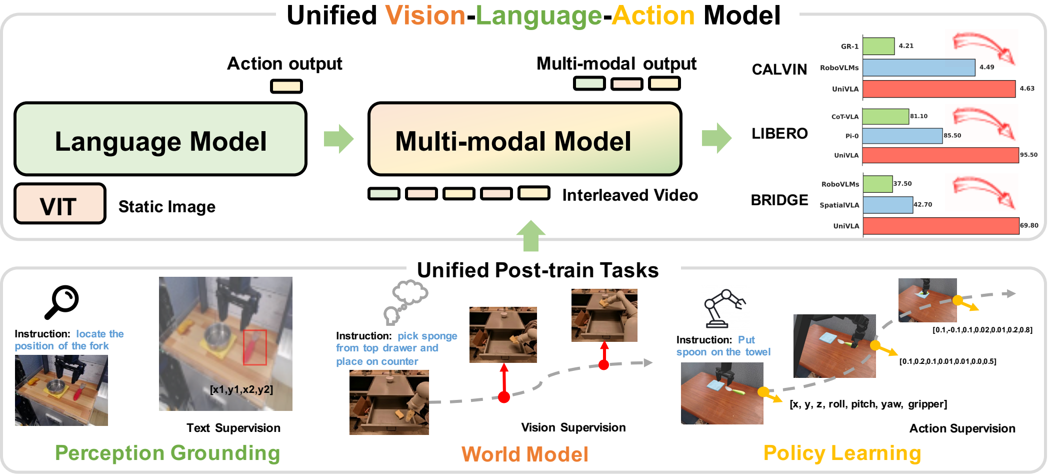


Figure 1: **We present UniVLA, a unified vision-language-action model.** Unlike prior VLA approaches that typically rely on an extra vision encoder to extract image features and generate only action outputs, UniVLA represents vision, language, and action as discrete tokens within a unified autoregressive framework. This unified modeling paradigm enables multi-modal outputs and supports a wide range of tasks—such as text-supervised perception grounding, vision-supervised world modeling, and action-supervised policy learning—within a single architecture. The unified token-based design further allows UniVLA to effectively leverage large-scale multimodal data, particularly video, for scalable and generalizable learning. UniVLA achieves new state-of-the-art results on CALVIN, LIBERO, and SimplerEnv-Bridge, significantly surpassing existing methods.

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.

1 Introduction

Developing agents capable of perceiving, reasoning, and acting in the physical world has long been a central objective of artificial intelligence. Recent advances in vision-language-action (VLA) models Brohan et al. (2023); Octo Model Team et al. (2024); Kim et al. (2024); Black et al. (2024), grounded in the powerful generalization capabilities of vision-language models (VLMs) Peng et al. (2023); Jaech et al. (2024); Beyer et al. (2024); Wang et al. (2024a); Guo et al. (2025), have demonstrated impressive performance across a wide range of robotic manipulation tasks, and are increasingly being adapted to generalist humanoid robots Bjorck et al. (2025); Ding et al. (2025) that demand broader embodied intelligence. However, most existing VLA approaches Kim et al. (2024); Black et al. (2024) follow a language-centric paradigm: visual observations are first projected into a semantic space, and action policies are subsequently derived based on these representations. This late-fusion strategy, while beneficial for semantic understanding and generalization, limits the formation of deeply coupled cross-modal representations and impedes the learning of temporal and causal dependencies across the perception-action loop. This raises a central question: *Can vision, language, and action be jointly modeled within a unified representation space to facilitate tighter cross-modal integration and more effective policy learning?*

While appealing in theory, unified modeling presents two key challenges. First, vision, language, and action are inherently heterogeneous modalities: vision comprises high-dimensional, continuous spatial signals; language conveys abstract, discrete semantics; and actions involve temporally ordered sequences with causal dependencies. Second, the perception-to-action pipeline is inherently dynamic and causal, yet existing VLA models Brohan et al. (2023); Kim et al. (2024); Black et al. (2024) often adopt static, language-centric paradigms that merely learn the mapping from static image to action. These models fail to capture the dynamic nature of real-world interactions, thereby limiting their ability to leverage the rich temporal information from videos for training.

To address the above challenges, we introduce UniVLA, a novel framework for unified vision–language–action learning. As illustrated in Figure 1, we propose a unified framework that supports both *multimodal* and *multi-task* learning. At the modality level, vision, language, and action signals are all transformed into discrete tokens and modeled using a shared vocabulary. This unified token representation allows for joint learning across modalities, fostering deeper cross-modal understanding and integration. Building upon the unified framework, we adopt an autoregressive, Markov chain-based sequence modeling approach, where observations and actions are interleaved. This structure naturally incorporates causal dependencies, enabling the model to reason over temporal dynamics rather than treating perception and action as isolated tasks. By integrating the world model paradigm during training, we leverage large-scale robotics videos for self-supervised learning, allowing the model to capture environment dynamics in a temporally consistent and causally grounded manner. Remarkably, we find that post-training with world models significantly enhances policy learning, particularly for long-horizon and out-of-distribution tasks.

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. Our model incorporates world model learning during post-training, enabling it to effectively capture visual dynamics from large-scale videos. This strategy significantly enhances both data efficiency and training efficiency in downstream policy learning, and allows for rapid adaptation to novel robotic tasks. Beyond policy learning, we demonstrate the model's multimodal output capabilities, including spatial reasoning and visual prediction, highlighting its versatility. Furthermore, we extend our approach to autonomous driving scenarios for broader applicability. These results underscore the potential of our unified VLA model as an alternative and promising direction for generalist embodied intelligence.

Our contributions are summarized as follows:

- 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. This approach offers a novel architecture alternative to the existing VLA paradigm, facilitating more integrated cross-modal modeling and enabling large-scale video-based training.

- Our unified sequence modeling framework supports a broad range of multimodal tasks. Through extensive experiments with various post-training strategies, we show that world models can effectively capture temporal dynamics from video data, substantially boosting performance and improving both data and training efficiency in downstream policy learning—spanning simulation benchmarks, real-world robotic platforms, and even driving domains, where world model post-training consistently benefits policy learning.
- Our model achieves state-of-the-art performance on several simulated benchmarks (CALVIN, LIBERO, and SimplerEnv-Bridge) and introduces an open-source VLA method supporting large-scale video training. We further explore its capabilities across various modalities, including spatial reasoning and video prediction, and demonstrate its effective transfer to real ALOHA platform and driving scenarios, highlighting its potential for generalist embodied intelligence.

2 Related Works

2.1 Vision-Language-Action Models

Recent vision-language-action (VLA) models have demonstrated strong task performance across diverse robots and tasks Brohan et al. (2023); Vuong et al. (2023); Driess et al. (2023); Kim et al. (2024); Zhen et al. (2024); Cheang et al. (2024); Black et al. (2024); Zheng et al. (2024); Liu et al. (2025); Kim et al. (2025); Intelligence et al. (2025). These models leverage pre-trained vision-language models (VLMs) to enhance understanding and generalization, further fine-tuned on large-scale robotic datasets for low-level control. Currently, VLA models can be categorized into two paradigms based on their output space: *pure action prediction* and *visual-guided action prediction*.

Pure action prediction. Recent efforts have extended vision-language models (VLMs) to incorporate action modalities, enabling direct action prediction from visual and language inputs. A prominent example is RT-2Brohan et al. (2023), which learns from both internet-scale and robotic data to generate discrete actions autoregressively, showcasing strong generalization and semantic grounding. Building upon this, RT-HBelkhale et al. (2024) introduces hierarchical actions to facilitate data sharing across tasks. OpenVLAKim et al. (2024) scales this paradigm with a 7B-parameter open-source model trained on 970k real-world demonstrations spanning diverse datasets. To enhance spatial reasoning, SpatialVLAQu et al. (2025) integrates spatial representations into the action modeling process. Beyond architecture scaling, new action modeling techniques have also emerged. π_0 Black et al. (2024) incorporates flow matching to improve action learning efficiency, while FAST Pertsch et al. (2025) introduces a unified frequency-domain formulation for discretizing actions.

Visual-guided action prediction. These studies leverage the power of visual pretraining, typically based on a policy-as-video formulation, by predicting future visual signals and subsequently decoding them into actions. SuSIE Black et al. (2023) predicts key future frames and derives actions through inverse dynamics. UniPi Du et al. (2023) generates videos from text instructions, extracting actions from the frames. GR series Wu et al. (2024); Cheang et al. (2024); Li et al. (2025a) leverages video pretraining for general policy learning. PAD Guo et al. (2024) uses diffusion models to simultaneously learn future images and actions. LAPA Ye et al. (2025) proposes to learn latent actions between images with VQ-VAE from action-free internet-scale videos. Track2Act Bharadhwaj et al. (2024) extracts point tracks from diverse web videos to guide the learning of interaction plans.

Both approaches have their strengths and weaknesses. The first, focused on action prediction, integrates well with Vision-Language Models (VLMs) but lacks spatial understanding and visual prediction capabilities. The second, incorporating visual generation, requires separating generative and action prediction models, limiting the full potential of VLMs. Our work unifies these directions, combining video generation pretraining with the strengths of VLMs to propose a native multimodal model with significant future potential.

Regarding comparisons with current VLA architectures such as $\pi_{0.5}$ Intelligence et al. (2025), our framework prioritizes video-centric modeling during the pretraining stage. While tokenizing actions may introduce a marginal trade-off in low-level control precision compared to continuous heads, this unified architecture offers significant advantages in pretraining scalability and multimodal alignment. Furthermore, this foundation is inherently extensible, accommodating specialized action experts during fine-tuning for scenarios requiring high-precision actuation.

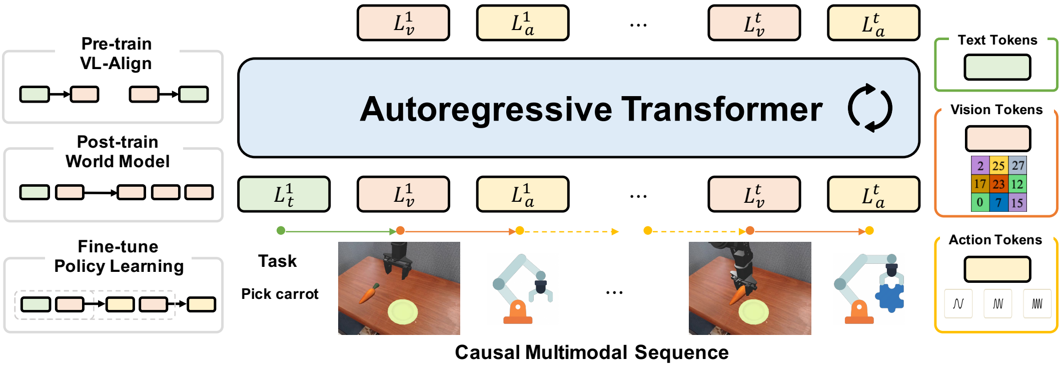


Figure 2: **Overview of the UniVLA framework.** Our model unifies information from different modalities into a discrete interleaved sequence, which is modeled using an autoregressive Transformer. To enable unified modeling, images are discretized using vector-quantized (VQ) encoders, while actions are transformed into the frequency domain and discretized via Discrete Cosine Transform (DCT) encoding. This causal multimodal sequence naturally preserves the temporal dynamics and causality required for real-world tasks. The model builds upon a pretrained vision-language model and follows a two-stage training strategy: (1) a post-training phase that adopts world-model training on large-scale datasets without requiring actions, and (2) a fine-tuning phase that interleaves actions into the sequence, enabling policy learning on downstream tasks.

2.2 World Models for Robotics

World models Ha & Schmidhuber (2018); Hafner et al. (2019a); LeCun (2022) have gained widespread attention for their ability to capture and reason about the dynamics of the physical world. They have emerged as a cornerstone in a range of domains, including interactive video generation Bruce et al. (2024); Che et al. (2024), autonomous driving Hu et al. (2023a); Wang et al. (2024d;b); Gao et al. (2024), and robotics Du et al. (2023); Wu et al. (2023); Yang et al. (2023). Recent advances in robotics increasingly focus on general-purpose controllable video generation to simulate realistic and diverse robot-environment interactions. Visual Foresight Finn & Levine (2017) leverages action-conditioned video prediction with model-predictive control, enabling robots to plan manipulation tasks by forecasting future visual observations. UniSim Yang et al. (2023) builds a "universal simulator" trained on diverse visual datasets, capable of visualizing the effects of both high-level instructions (e.g., "open the drawer") and low-level controls in novel scenes. RoboDreamer Zhou et al. (2024) learns a compositional world model by factorizing video generation, facilitating the synthesis of novel action sequences. DREMA Barcellona et al. (2024) replicates scene dynamics and structure by integrating Gaussian Splatting with physics simulation. VLP Du et al. (2024) enables long-horizon visual planning by combining text-to-video generation with vision-language models as heuristic evaluators. DayDreamer Wu et al. (2023) extends Dreamer Hafner et al. (2019b) to real-world robotic platforms, while UVA Li et al. (2025b) proposes a joint video-action latent space to decouple video and action generation, achieving high accuracy and efficiency in policy inference. AdaWorld Gao et al. (2025) extracts latent actions from videos in a self-supervised manner and builds an autoregressive world model conditioned on these latent actions.

3 Unified Vision-Language-Action Model

In this section, we present the design of UniVLA, as illustrated in Figure 2. Unlike previous VLA models Kim et al. (2024); Black et al. (2024) that rely on ViT Dosovitskiy et al. (2021) for image encoding, our approach adopts an encoder-free architecture, converting all modalities into discrete tokens and learning them autoregressively. The overall design is simple yet effective, demonstrating strong scalability.

Our *unified paradigm* has two key aspects: first, it *unifies the learning of multiple modalities*, integrating various modality tokens into a shared representation space and employing a transformer for autoregressive learning; Second, it *unifies sequence modeling across tasks* through the natural interleaving of modalities, facilitating the seamless combination of tasks such as video generation, visual grounding, and action learning. In the following sections, we will introduce the method from the perspectives of *Unified Multimodal Model* and *Unified Multimodal Sequence Modeling*.

3.1 Unified Multimodal Model

As illustrated in Figure 2, our method unifies language, vision, and action modalities by converting each into discrete tokens and concatenating them into a single multimodal sequence L. Specifically, L_t, L_v, and L_a denote the discrete token sequences for language, vision, and action, respectively, all drawn from a shared vocabulary. The superscript indicates the temporal step, with tokens interleaved across modalities to preserve temporal alignment.

For example, in the robotic manipulation task, a textual instruction is provided only at the beginning, followed by a naturally interleaved sequence of visual observations and actions. The language and vision tokenizers adopt the same design as Emu3 Wang et al. (2024c); visual observations are discretized using a VQ tokenizer Zheng et al. (2022), while actions are encoded using FAST Pertsch et al. (2025). To clearly demarcate modality boundaries, we employ special tokens—`boi` (begin of image), `eoi` (end of image), `boa` (begin of action), and `eoa` (end of action)—to encapsulate image and action tokens, respectively.

Action Modeling We follow FAST Pertsch et al. (2025) and apply the Discrete Cosine Transform (DCT) to convert continuous action sequences into discrete action tokens. Specifically, given an action sequence within a time window, we define L_a at a given time step as a sequence of action tokens $[T_1, \ldots, T_n]$. The raw action sequence $A_{1:H} = \{a_1, a_2, \ldots, a_H\}$ spans a window of size H, where each action a_t is a d-dimensional vector. The FAST action tokenizer encodes $A_{1:H}$ into a discrete token sequence $[T_1, \ldots, T_n]$, with n tokens drawn from a vocabulary of size $|V|$. Similar to natural language processing, action sequences can vary in token length, resulting in a variable-length (n) discrete representation.

Training Objective Since all modality signals are transformed into discrete tokens, the training objective is simplified to a standard next-token prediction task using cross-entropy loss. To accommodate different task formats, we selectively include specific tokens in the loss computation, ensuring compatibility and flexibility across diverse tasks.

3.2 Unified Multimodal Sequence Modeling

As shown in Figure 2, our multimodal sequence representation naturally captures the temporal dynamics and causal structure inherent in task execution. The embodied planning problem can be formulated as a Markov Decision Process (MDP), a general mathematical framework for decision-making in partially stochastic environments. For example, in the task of picking a carrot, the instruction and current observation inform the action; this action alters the environment, leading to a new observation that subsequently guides the next action. Building on this interleaved Markovian formulation, we unify a variety of tasks within a shared sequence modeling framework, and present the task-specific modeling strategies in the following.

World Model (Post-train) Within the MDP framework, a world model aims to learn the dynamics of the environment by modeling the transition function $P(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$. The learned world model enables agents to simulate future trajectories, plan actions, and reason about consequences without direct interaction with the environment. Specifically, in the context of robotic tasks, we treat the language instructiom as a general form of action. Given the current observation L_v^1 and the instruction L_t^1, the world model need to predict future visual content. In this setting, we use the loss computed solely from the vision tokens as the supervisory signal, enabling the model to generate visual predictions conditioned on the given instruction and observed state. Sequence S_v formulation is as follows:

$$S_v = \{L_t^1, L_v^1, L_v^2, ..., L_v^t\} \tag{1}$$

Policy Learning (Fine-tune) Policy learning enables the agent to determine optimal actions based on both current observations and prior states, thereby effectively guiding task execution. In this setting, we employ a loss function computed solely from the action tokens. The sequence S_a representing the interactions over time is formulated as:

$$S_a = \{L_t^1, L_v^1, L_a^1, L_v^2, L_a^2, \ldots, L_v^t, L_a^t\} \tag{2}$$

As illustrated in Figure 2, in this interleaved format, we adopt a two-stage training paradigm for robotic tasks. The model is initialized with a vision-language (VL) aligned checkpoint, endowing it with basic vision-language capabilities. The post-training stage leverages a world model objective to capture video dynamics, treating world modeling as a general visual learning task. Building upon the learned world model, the fine-tuning stage focuses on action learning to refine task-specific behaviors. We observe that incorporating the world model significantly enhances the efficiency and effectiveness of policy learning.

4 Experiments

4.1 Dataset

CALVIN. *CALVIN* Mees et al. (2022b) is a simulated benchmark tailored for evaluating long-horizon, language-conditioned robotic manipulation. It comprises four simulated environments (A, B, C, and D), each containing demonstration trajectories collected via human teleoperation. The benchmark encompasses 34 distinct manipulation tasks with a total of 1,000 unique language instructions. Performance is measured by the average number of successfully completed sub-tasks within a sequence. Standard evaluation protocols include the *ABC→D* and *ABCD→D* settings, which test a model's ability to generalize to unseen environments and compositions of long-horizon tasks.

LIBERO. The *LIBERO* benchmark Liu et al. (2023) is a comprehensive suite for lifelong robotic manipulation, comprising four task suites with 10 tasks and 50 human demonstrations each. These suites are designed to evaluate different generalization abilities: *LIBERO-Spatial* tests spatial reasoning by varying layouts with fixed objects; *LIBERO-Object* assesses object-level generalization with varying objects in a fixed scene; *LIBERO-Goal* targets goal-conditioned behavior by varying task goals; and *LIBERO-Long* (*LIBERO-10*) features long-horizon, compositional tasks with diverse objects, layouts, and goals, challenging temporal and compositional reasoning.

SimplerEnv. SimplerEnv Li et al. (2024d) serves as a simulation benchmark designed to evaluate the transferability and generalization capabilities of models trained on real-world video data. It incorporates diverse manipulation tasks on WidowX and Google Robot platforms with variations in lighting, object textures, colors, and camera viewpoints.

4.2 Implementation Details

The model is a purely autoregressive Transformer with 8.5B parameters, identical to Emu3 Wang et al. (2024c). Images are tokenized via a VQ-based encoder with 8× spatial compression. Actions are encoded as frame-to-frame differences, normalized by the 1st/99th percentiles, and tokenized with FAST Pertsch et al. (2025), whose 1024-token vocabulary replaces the last 1024 IDs of the language tokenizer.

Post-training Stage. In the post-training stage, we leverage large-scale robot-centric video datasets to study the effects of various post-training strategies on downstream policy learning. The model is initialized with pre-trained weights from the first stage of Emu3 Wang et al. (2024c). We curate a total of 622K videos from existing robotics datasets (details provided in the appendix), and identify the world model as the most effective post-training approach. During training, supervision is applied solely on the vision tokens. The model is trained for 30K steps with a batch size of 64.

Fine-tuning Stage. During fine-tuning, the model is initialized with weights from the post-training stage and trained using a two-frame interleaved vision-action sequence with an action chunk size of 10. A cosine annealing learning rate schedule is applied, starting at 8×10^{-5}, and the loss is computed solely over action tokens. For the CALVIN benchmark, RGB observations from both third-person (200×200) and wrist-view (80×80) cameras are used. Training is conducted on A100 GPUs with a batch size of 192 for 8k steps. For the LIBERO benchmark, third-person and wrist-view RGB images (both at 200×200) are used to train a unified model with a batch size of 192 for 8k steps. A single model is evaluated across four task suites. For the SimplerEnv benchmark, single-view RGB observations are used with input resized to 256×256. Training is conducted on the Bridge-WidowX setup using a batch size of 128 for 20k steps, with an action chunk size of 5.

Inference. Our method adopts an interleaved vision–action training scheme, requiring supervision from future frames only during training. During inference, the model generates only action tokens without predicting future frames, conditioning on the current observed images. The action generation stops once the model predicts the `eoa` (end of action) token.

Additional implementation details on the post-training strategy, real-robot fine-tuning procedures, and autonomous driving experiments are provided in the appendix.

Table 1: **Long-horizon robotic manipulation evaluation on the CALVIN benchmark.** * denotes the use of video inputs.

Method	**Task**	**Tasks Completed in a Row**					**Avg. Len ↑**
		1	2	3	4	5	
MCIL Lynch & Sermanet (2020)	ABCD→D	0.373	0.027	0.002	0.000	0.000	0.40
RT-1 Brohan et al. (2022)	ABCD→D	0.844	0.617	0.438	0.323	0.227	2.45
Robo-Flamingo Li et al. (2024c)	ABCD→D	0.964	0.896	0.824	0.740	0.660	4.09
GR-1* Wu et al. (2024)	ABCD→D	0.949	0.896	0.844	0.789	0.731	4.21
UP-VLA Zhang et al. (2025)	ABCD→D	0.962	0.921	0.879	0.842	0.812	4.42
RoboVLMs* Li et al. (2024b)	ABCD→D	0.967	0.930	0.899	0.865	0.826	4.49
UniVLA*	ABCD→D	**0.985**	**0.961**	**0.931**	**0.899**	**0.851**	**4.63**
MCIL Lynch & Sermanet (2020)	ABC→D	0.304	0.013	0.002	0.000	0.000	0.31
Robo-Flamingo Li et al. (2024c)	ABC→D	0.824	0.619	0.466	0.331	0.235	2.47
SuSIE Black et al. (2023)	ABC→D	0.870	0.690	0.490	0.380	0.260	2.69
GR-1* Wu et al. (2024)	ABC→D	0.854	0.712	0.596	0.497	0.401	3.06
MoDE* Reuss et al. (2024)	ABC→D	0.962	0.889	0.811	0.718	0.635	4.01
UP-VLA Zhang et al. (2025)	ABC→D	0.928	0.865	0.815	0.769	0.699	4.08
RoboVLMs* Li et al. (2024b)	ABC→D	0.980	0.936	0.854	0.778	0.704	4.25
Seer-Large Tian et al. (2024)	ABC→D	0.963	0.916	0.861	0.803	0.740	4.28
UniVLA*	ABC→D	**0.989**	**0.948**	**0.890**	**0.828**	**0.751**	**4.41**

4.3 Main Results

In this section, we evaluate our method on three simulation benchmarks: CALVIN (long-horizon tasks), LIBERO (diverse generalization), and SimplerEnv (real-to-sim manipulation). Our approach consistently achieves state-of-the-art performance across all settings.

CALVIN Simulation Evaluation. Table 1 presents the experimental results in the CALVIN benchmark. Our method achieves the highest performance on both the ABC→D and ABCD→D tasks, significantly outperforming previous approaches and demonstrating strong capabilities in multi-task learning and long-horizon planning.

LIBERO Simulation Evaluation. Following Zhao et al. (2025), we report the average success rate over 500 episodes for each task suite (Spatial, Object, Goal, Long). As shown in Table 2, UniVLA achieves the best overall performance across all LIBERO benchmark suites, with particularly significant gains on long-horizon tasks—raising the previous best from 69.0% to 94.0%. Compared to π_0 Pertsch et al. (2025), our method demonstrates superior performance on long-horizon tasks.

Table 2: **Comparison of different methods on the LIBERO benchmark.**

Method	**SPATIAL**	**OBJECT**	**GOAL**	**LONG**	**Average**
DP* Chi et al. (2023)	78.3%	92.5%	68.3%	50.5%	72.4%
Octo Team et al. (2024)	78.9%	85.7%	84.6%	51.1%	75.1%
OpenVLA Kim et al. (2024)	84.9%	88.4%	79.2%	53.7%	76.5%
SpatialVLA Qu et al. (2025)	88.2%	89.9%	78.6%	55.5%	78.1%
CoT-VLA Zhao et al. (2025)	87.5%	91.6%	87.6%	69.0%	81.1%
π_0-FAST Pertsch et al. (2025)	96.4%	96.8%	88.6%	60.2%	85.5%
UniVLA (single-frame)	**97.0%**	**99.0%**	92.6%	90.8%	94.8%
UniVLA	95.4%	98.8%	**93.6%**	**94.0%**	**95.5%**

SimplerEnv Simulation Evaluation. Table 3 summarizes the performance across various manipulation policies on the Bridge-WidowX setup. Our approach demonstrates a significant improvement over prior methods, raising the average success rate from 42.7% to 69.8%. In particular, it shows marked improvements on previously difficult tasks, including stack block, put carrot and put spoon.

Table 3: **Evaluation on SimplerEnv-WidowX across various manipulation tasks.**

Model	**Put Spoon on Towel**		**Put Carrot on Plate**		**Stack Green on Yellow Block**		**Put Eggplant in Yellow Basket**		**Overall**
	Grasp	Success	Grasp	Success	Grasp	Success	Grasp	Success	Success
RT-1-X Brohan et al. (2023)	16.7%	0.0%	20.8%	4.2%	8.3%	0.0%	0.0%	0.0%	1.1%
Octo-Base Octo Model Team et al. (2024)	34.7%	12.5%	52.8%	8.3%	31.9%	0.0%	66.7%	43.1%	16.0%
Octo-Small Octo Model Team et al. (2024)	77.8%	47.2%	27.8%	9.7%	40.3%	4.2%	87.5%	56.9%	29.5%
OpenVLA Kim et al. (2024)	4.1%	0.0%	33.3%	0.0%	12.5%	0.0%	8.3%	4.1%	1.0%
RoboVLMs Li et al. (2024b)	70.8%	45.8%	33.3%	20.8%	54.2%	4.2%	91.7%	79.2	37.5%
SpatialVLA Qu et al. (2025)	20.8%	16.7%	29.2%	25.0%	62.5%	29.2%	100%	**100%**	42.7%
CogACT Li et al. (2024a)	-	71.1%	-	50.8%	-	15.0%	-	67.5%	51.3%
UniVLA	**83.3%**	**83.3%**	**74.0%**	**66.7%**	**95.8%**	**33.3%**	**100.0%**	95.8%	**69.8%**

Table 4: **Effectiveness of World Model Post-Training.** We compare different post-training strategies by fine-tuning only with action prediction on the downstream benchmarks.

Post-training Stage			**Generalization**		**Long-horizon**	
Strategy	**Sequence**	**Supervision**	**LIBERO**	**SimplerEnv-WidowX**	**LIBERO-Long**	**CALVIN**
			48.5	0.0	17.4	1.46
action prediction	T, I, A	action	43.9 (-4.6)	0.0	10.6 (-6.8)	0.52(-0.94)
text-to-image	T, I	vision	69.8 (+21.3)	6.3 (+6.3)	55.8 (+38.4)	3.79 (+2.33)
video prediction	$I_1, ..., I_t$	vision	78.9 (+30.4)	17.7 (+17.7)	80.8 (+63.4)	3.59 (+2.13)
world model	$T, I_1, ..., I_t$	vision	**94.2** (+45.7)	**64.6** (+64.6)	**89.2** (+71.8)	**4.61** (+3.15)

4.4 In-Depth Analysis

This section presents an in-depth analysis of our unified framework, deriving critical design insights for future VLA models. We primarily focus on the **pivotal role of the world model**, illustrating how its integration as a post-training strategy fundamentally advances downstream policy learning capabilities (Table 4) and accelerates training efficiency (Table 5). We then turn to an ablation analysis of the module design. We find that even without post-training stage, incorporating visual prediction loss (Table 6a) and historical context (Table 6b) still contributes positively to policy learning.

Effectiveness of World Model Post-Training. Table 4 investigates the effects of different post-training strategies on downstream policy learning across various simulation benchmarks. Our results reveal that action-only post training suffers from limited transferability, primarily due to the non-uniformity of action spaces across diverse robot datasets—where variations in embodiments, control frequencies, and normalization result in heterogeneous distributions. In contrast, most vision-based post-training approaches significantly enhance policy learning and demonstrate superior generalization without relying on explicit action labels, thereby highlighting the crucial role of visual learning in transferability. Among these, the world model post-training approach yields the most substantial gains, enhancing both generalization and long-horizon planning capabilities. A comparison with text-to-image (T2I) training emphasizes the importance of modeling temporal dynamics in video data, while contrasting with video-only training highlights the essential role of textual guidance in state transitions. Notably, this world model training requires no action annotations, enabling scalable learning from large-scale video data and providing a promising direction for future VLA research.

Data and Training Efficiency for Fine-tuning. Table 5 shows that post-training substantially enhances downstream policy learning efficiency. On the CALVIN benchmark (Table 5a), our method achieves higher success rates using only 10% of the fine-tuning data, outperforming prior approaches such as GR-1 Wu et al. (2024) and RoboVLMs Li et al. (2024b). In addition, Table 5b highlights improved training efficiency, as the model rapidly converges with fewer fine-tuning iterations. The Simpler-Env results further demonstrate the effectiveness of world-model-based post-training for efficient policy adaptation across diverse robotic setups. While similar effects are observed in latent-action methods Ye et al. (2025); Chen et al. (2024b); Gao et al. (2025), our world model offers a simpler paradigm without latent actions, achieving better transferability.

Effectiveness of Visual Prediction. While post-training proves effective, it is also crucial that the model demonstrates strong performance without relying on it. As shown in Table 6a, our findings indicate that, even without post-training, fine-tuning with visual loss supervision—leveraging the autoregressive nature of the model—naturally integrates world model learning into the policy learning process. This approach leads to a significant improvement in the model's performance.

Table 5: **Post-training enables data-efficient and training-efficient downstream policy learning.**

(a) **Data efficiency comparison.**

Method	Data	CALVIN
RT-1 Brohan et al. (2022)	10%	0.34
MT-R3M Nair et al. (2022)	10%	0.61
HULC Mees et al. (2022a)	10%	1.11
GR-1 Wu et al. (2024)	10%	2.00
RoboVLMS Li et al. (2024b)	10%	2.52
UniVLA (w/o post-train)	10%	0.15
UniVLA	10%	**3.19**

(b) **Training efficiency comparison.**

Fast convergence (CALVIN)			
Training Iters	2k	4k	8k
w/o post-train	0.37	0.82	1.46
w/ post-train	4.21	4.56	4.61

Fast adaptation (SimplerEnv-Bridge)			
Method	Batch size	Iters	Success
RoboVLMs Li et al. (2024b)	128	50k	37.5
UniVLA	128	12k	64.6

Table 6: **Ablation study on the visual prediction and historical context in policy learning.**

(a) **Effectiveness of visual prediction.**

Post-train	Visual prediction	CALVIN	LIBERO
✓		4.61	94.2
	✓	4.42	88.7
		1.46	48.5

(b) **Effectiveness of history context.**

Observations		**Avg. Len ↑**
History Window	**Current + History**	
0	1 + 0	4.26
10	1 + 1	4.61
10	1 + 2	4.43
20	1 + 2	4.47

Effectiveness of History Context. History context—comprising past observations and actions—provides valuable guidance for robot planning. In this section, we investigate the appropriate length of the history window during the fine-tuning stage. As shown in Table 6b, our ablation study on the CALVIN benchmark examines the impact of varying history window lengths. Incorporating a history window significantly improves performance (from 4.26 to 4.61). However, extending the window beyond a certain length yields diminishing returns, suggesting that recent observations carry the most predictive value, consistent with the Markov property in sequential planning.

4.5 Broader Applications

Real-world ALOHA Tasks. We further validate the effectiveness of our approach on the real-world ALOHA platform through two tasks: *pouring water* and *folding clothes*. Our model relies exclusively on visual observations, without access to any state information.

As shown in Table 7, we decompose the pouring task into three sequential subtasks: *right-hand grasping the bottle*, *left-hand grasping the cup*, and *pouring*. UniVLA demonstrates competitive performance against π_0 Black et al. (2024) in real-world ALOHA tasks. While π_0 achieves a marginally higher success rate in the grasping stage, both methods exhibit comparable performance in the pouring stage. All results are averaged over 8 trials. Notably, we observe that precise pouring remains a challenge for both methods. Interestingly, we observe that world model post-training yields a substantial performance improvement even in real-world experiments. Additional details and demonstration videos are provided in the appendix.

Table 7: **Comparison across real-world ALOHA manipulation tasks.**

Method	Right-hand Grasp Bottle	Left-hand Grasp Cup	Pour Water	Overall
π_0 Black et al. (2024)	87.5	75.0	37.5	37.5
UniVLA (w/o world model)	12.5	0.0	0.0	0.0
UniVLA (w/ world model)	87.5	62.5	37.5	37.5

As shown in Figure 3, despite the precision loss from discrete tokens, our method is still able to perform fine-grained tasks such as folding clothes.

End-to-end Learning for Autonomous Driving. To further explore the potential of our method, we perform a preliminary transfer to the autonomous driving domain by finetuning the model on the NAVSIM benchmark. Notably, our method is a pure autoregressive, token-based framework, modeling

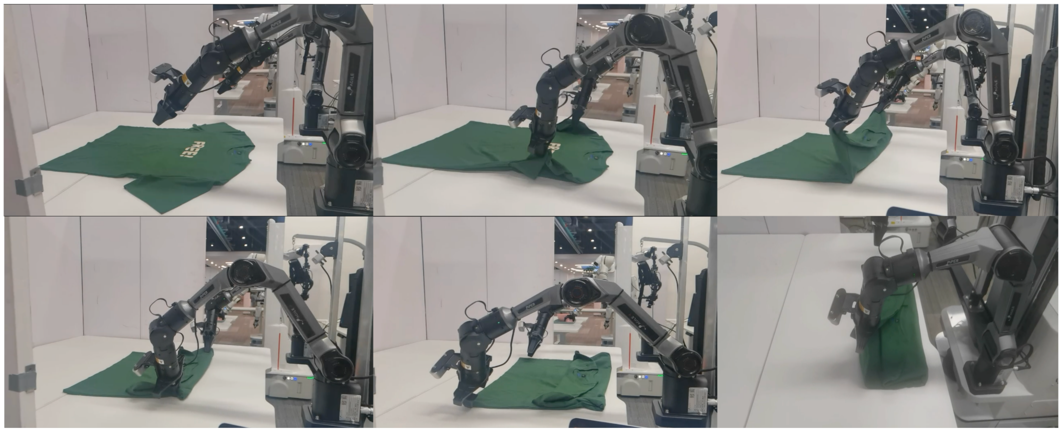


Figure 3: **Visualizations of the *folding clothes* task performed by our method.**

the driving task as causal sequence prediction over discretized multimodal tokens. Despite using only front-view camera inputs—without relying on BEV representations or multi-sensor fusion—our model achieves powerful performance on the NAVSIM test set. Notably, world model post-training is also effective in the driving domain and can lead to significant performance improvements. These results highlight the strong potential of our method for broader real-world applications.

Table 8: **Broader applications of UniVLA for end-to-end autonomous driving on the NAVSIM.** MC: Multi Camera. L: LiDAR. FC: Front Camera.

Method	Model	Input	NC↑	DAC↑	EP↑	TTC↑	C↑	PDMS↑
Human	–	–	100.0	100.0	87.5	100.0	99.9	94.8
Ego Status MLP	–	Ego State	93.0	77.3	62.8	83.6	100.0	65.6
VADv2 Chen et al. (2024a)	BEV-based	MC	97.9	91.7	77.6	92.9	100.0	83.0
UniAD Hu et al. (2023b)	BEV-based	MC	97.8	91.9	78.8	92.9	100.0	83.4
Transfuser Chitta et al. (2022)	BEV-based	MC&L	97.7	92.8	79.2	92.8	100.0	84.0
UniVLA	Auto-regressive	FC	96.9	91.1	76.8	91.7	96.7	81.7
UniVLA (w/ world model post-train)	Auto-regressive	FC	98.3	93.8	80.0	94.2	100.0	85.6

5 Conclusion

In this paper, we present UniVLA, a unified framework for vision–language–action modeling that bridges heterogeneous modalities through a shared token space and models them autoregressively. The proposed unified design facilitates deeper cross-modal integration and inherently supports flexible multimodal tasks. By leveraging a world model trained to capture dynamics and causality from videos, we observe significant improvements in downstream policy learning, both in terms of performance and efficiency. Extensive simulation experiments further demonstrate the model's strong generalization ability, efficient policy learning, and broad applicability across diverse domains. These findings highlight the great potential of our method as a new paradigm for vision–language–action modeling.

Limitations and Future Work Our exploration of post-training scalability is currently limited by compute availability, yet early indicators show promise for scaling to extensive video datasets. Although our unified framework excels in cross-modal learning, further research is needed to seamlessly integrate reinforcement learning for robust policy adaptation. Regarding real-world deployment, we have validated the feasibility of our approach on tasks such as dual-arm pouring and cloth folding; however, generalizing to more complex, open-ended manipulation tasks remains an open challenge.

References

Leonardo Barcellona, Andrii Zadaianchuk, Davide Allegro, Samuele Papa, Stefano Ghidoni, and Efstratios Gavves. Dream to manipulate: Compositional world models empowering robot imitation learning with imagination. *arXiv preprint arXiv:2412.14957*, 2024. 4

Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language. *arXiv preprint arXiv:2403.01823*, 2024. 3

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 2

Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation. In *European Conference on Computer Vision*, 2024. 3

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. 2

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023. 3, 7

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 2, 3, 4, 9

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 7, 9, 17

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 2, 3, 8

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 4

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. *arXiv preprint arXiv:2411.00769*, 2024. 4

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024. 3

Lawrence Yunliang Chen, Simeon Adebola, and Ken Goldberg. Berkeley UR5 demonstration dataset. https://sites.google.com/view/berkeley-ur5/home. 17

Lili Chen, Shikhar Bahl, and Deepak Pathak. Playfusion: Skill acquisition via diffusion from language-annotated play. In *Conference on Robot Learning*, pp. 2012–2029. PMLR, 2023. 17

Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024a. 10

Yi Chen, Yuying Ge, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for robot manipulation. *arXiv preprint arXiv:2412.04445*, 2024b. 8

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, pp. 02783649241273668, 2023. 7

Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022. 10

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024. 19

Pengxiang Ding, Jianfei Ma, Xinyang Tong, Binghong Zou, Xinxin Luo, Yiguo Fan, Ting Wang, Hongchao Lu, Panzhong Mo, Jinxin Liu, et al. Humanoid-vla: Towards universal humanoid control with visual integration. *arXiv preprint arXiv:2502.14795*, 2025. 2

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, pp. 8469–8488. PMLR, 2023. 3

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023. 3, 4

Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. *ICLR*, 2024. 4

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 2786–2793. IEEE, 2017. 4

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024. 4

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions. *arXiv preprint arXiv:2503.18938*, 2025. 4, 8

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017. 17

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. Maniskill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations*, 2023. 17

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2

Yanjiang Guo, Yucheng Hu, Jianke Zhang, Yen-Jen Wang, Xiaoyu Chen, Chaochao Lu, and Jianyu Chen. Prediction with action: Visual policy learning via joint denoising process. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 4

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a. 4

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019b. 4

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023a. 4

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17853–17862, 2023b. 10

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. \pi_ {0.5}: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. 3

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 2

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on robot learning*, pp. 651–673. PMLR, 2018. 17

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024. 17

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 2, 3, 4, 7, 8, 17

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025. 3

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022. 4

Peiyan Li, Hongtao Wu, Yan Huang, Chilam Cheang, Liang Wang, and Tao Kong. Gr-mg: Leveraging partially-annotated data via multi-modal goal-conditioned policy. *IEEE Robotics and Automation Letters*, 2025a. 3

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024a. 8

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025b. 4

Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. Towards generalist robot policies: What matters in building vision-language-action models. *arXiv preprint arXiv:2412.14058*, 2024b. 7, 8, 9, 17

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. In *ICLR*, 2024c. 7

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024d. 2, 6

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023. 2, 6, 17

Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025. 3

Jianlan Luo, Charles Xu, Fangchen Liu, Liam Tan, Zipeng Lin, Jeffrey Wu, Pieter Abbeel, and Sergey Levine. Fmb: a functional manipulation benchmark for generalizable robotic learning. *The International Journal of Robotics Research*, pp. 02783649241276017, 2023. 17

Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data. *arXiv preprint arXiv:2005.07648*, 2020. 7

Oier Mees, Lukas Hermann, and Wolfram Burgard. What matters in language conditioned robotic imitation learning over unstructured data. *IEEE Robotics and Automation Letters*, 7(4):11205–11212, 2022a. 9

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022b. 2, 6, 17

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022. 9

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024. 2, 8

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025. 3, 5, 6, 7

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025. 3, 7, 8

Moritz Reuss, Jyothish Pari, Pulkit Agrawal, and Rudolf Lioutikov. Efficient diffusion transformer policies with mixture of expert denoisers for multitask learning. *arXiv preprint arXiv:2412.12953*, 2024. 7

Erick Rosete-Beas, Oier Mees, Gabriel Kalweit, Joschka Boedecker, and Wolfram Burgard. Latent plans for task-agnostic offline reinforcement learning. In *Conference on Robot Learning*, pp. 1838–1849. PMLR, 2023. 17

Rutav Shah, Roberto Martín-Martín, and Yuke Zhu. Mutex: Learning unified policies from multimodal task specifications. *arXiv preprint arXiv:2309.14320*, 2023. 17

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. 7

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. *arXiv preprint arXiv:2412.15109*, 2024. 7

Quan Vuong, Sergey Levine, Homer Rich Walke, Karl Pertsch, Anikait Singh, Ria Doshi, Charles Xu, Jianlan Luo, Liam Tan, Dhruv Shah, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023. 3

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023. 17

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a. 2

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pp. 55–72. Springer, 2024b. 4

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024c. 5, 6

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14749–14759, 2024d. 4

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 7, 8, 9

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pp. 2226–2240. PMLR, 2023. 4

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023. 4

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *ICLR*, 2025. 3, 8

Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiang Zhu, and Jianyu Chen. Up-vla: A unified understanding and prediction model for embodied agent. *arXiv preprint arXiv:2501.18867*, 2025. 7

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. *arXiv preprint arXiv:2503.22020*, 2025. 7

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. In *International Conference on Machine Learning*, pp. 61229–61245. PMLR, 2024. 3

Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425, 2022. 5

Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024. 3

Gaoyue Zhou, Victoria Dean, Mohan Kumar Srirama, Aravind Rajeswaran, Jyothish Pari, Kyle Hatch, Aryan Jain, Tianhe Yu, Pieter Abbeel, Lerrel Pinto, et al. Train offline, test online: A real robot learning benchmark. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9197–9203. IEEE, 2023. 17

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024. 4

Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. In *Conference on Robot Learning*, pp. 1199–1210. PMLR, 2023. 17

APPENDIX

A IMPLEMENTATION DETAILS

Post-training Stage We began by selecting several high-quality robotics datasets for post-training, as summarized in Table 9. To account for differences in data collection frequencies across datasets, we applied dataset-specific frame sampling intervals to ensure that the temporal gap between keyframes is approximately one second. We further filtered out video sequences containing fewer than six frames, as well as those lacking corresponding text instructions. Due to the large number of videos from the Kuka Kalashnikov et al. (2018) dataset, we randomly retained 100k videos to prevent it from dominating the overall training data.

Table 9: **Post-training datasets**.

Dataset	Source	Data Type	Raw Videos	Used Videos	Interval
RT-1 Brohan et al. (2022)	Real	Text, Video, Action	87212	84084	3
BridgeV2 Walke et al. (2023)	Real	Text, Video, Action	60064	28083	5
DROID Khazatsky et al. (2024)	Real	Text, Video, Action	275997	145641	15
Kuka Kalashnikov et al. (2018)	Real	Text, Video, Action	580392	100000	3
TOTO Zhou et al. (2023)	Real	Text, Video, Action	902	899	20
Taco Play Rosete-Beas et al. (2023)	Real	Text, Video, Action	3242	3242	5
FMB Luo et al. (2023)	Real	Text, Video, Action	8611	7876	5
Berkeley autolab ur5 Chen et al.	Real	Text, Video, Action	896	896	5
VIOLA Zhu et al. (2023)	Real	Text, Video, Action	135	135	15
Cmu Play Fusion Chen et al. (2023)	Real	Text, Video, Action	576	576	10
Utaustin Mutex Shah et al. (2023)	Real	Text, Video, Action	1500	1500	10
CALVIN Mees et al. (2022b)	Sim	Text, Video, Action	22966	22966	5
LIBERO Liu et al. (2023)	Sim	Text, Video, Action	3386	3386	10
ManiSkill2 Gu et al. (2023)	Sim	Text, Video, Action	30213	193273	10
SSV2 Goyal et al. (2017)	Real	Text, Video	220847	220847	1

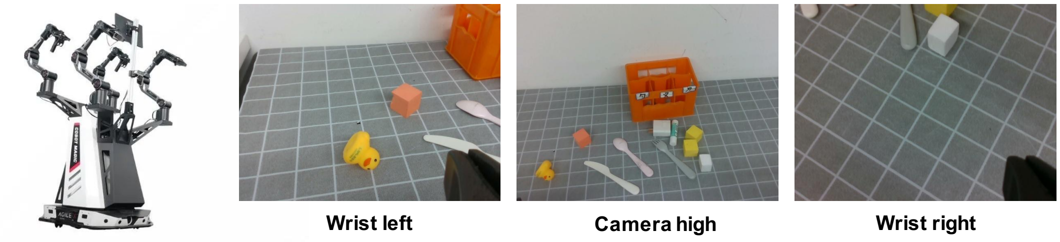
For the experiments in Table 4, to ensure a fair comparison of different post-training strategies, all models are trained on the same dataset (excluding SSV2 Goyal et al. (2017), which does not contain action annotations), with only the post-training strategy varied. For the *action prediction* task, we organize the input as (T, I, A), where T denotes the text instruction, I the image observations, and A the action sequence. During training, only the action tokens A are supervised in the loss computation. For the *text-to-image* task, the input is organized as (T, I), where T denotes the input text and I denotes the target image. During training, the loss is only computed on the vision tokens corresponding to I. For the *video prediction* task, the input is organized as $(I_1, ..., I_t)$, where I denotes the video frame. During training, the loss is computed on the vision tokens. For the *world model* task, the input is organized as $(T, I_1, ..., I_t)$, where T denotes the input text, I denotes the video frame. During training, the loss is computed on the vision tokens.

During training, the observations are resized to 256×256, using six frames as input, with the maximum sequence length set to 6400. We perform full-parameter training for 50k steps using 32 A100 GPUs (40GB), which takes approximately 4–5 days.

Simulation Finetuning The training setup is described in the main paper. We adopt full-parameter training, and for evaluation, we follow the testing protocols of OpenVLA Kim et al. (2024) and RoboVLMs Li et al. (2024b) across various benchmarks. By default, our model is trained using video-format sequences; however, it also supports fine-tuning with image-format sequences. In the ablation study evaluating the effect of visual prediction, when post-training is not applied, the visual token weight is set to 0.5 while the action token weight is set to 1.0, in order to maintain balance between the two modalities.

Real-robot Finetuning For real-world evaluation, we conduct experiments on the ALOHA platform, using images captured from three perspectives: `cam high`, `wrist left`, and `wrist right`. The real-robot is controlled using end-effector (EE) pose. All input images are resized to a resolution of 128×128. The model outputs a 14-dimensional action vector. The action chunk size is set to 20. For each task, we train for 8k steps with a batch size of 256. The learning rate is set to 5×10^{-5}, and all other settings remain consistent with the above. We also leverage world model pretraining, using video-based post-training on a collected real-aloha dataset (Table 10). Interestingly, this post-training provides substantial benefits even when transferring to real-robot execution.

B Real-Robot Experiments



Wrist left

Camera high

Wrist right

Figure 4: **Real-world setup of the AgileX Cobot Magic dual-arm robot.** The system is equipped with three Orbbec RGB cameras for visual observation: one mounted on the left wrist, one on the right wrist, and one positioned above for a high-angle view.

B.1 ALOHA Experimental Setup

The robotic platform used in this paper is **AgileX Cobot Magic V2.0**, a dual-arm robot. As shown in Figure 4, the robot is equipped with two arms and three camera views, enabling it to perform a variety of manipulation tasks. For example, Figure 5 illustrates a range of manipulation tasks collected from real-world scenarios.

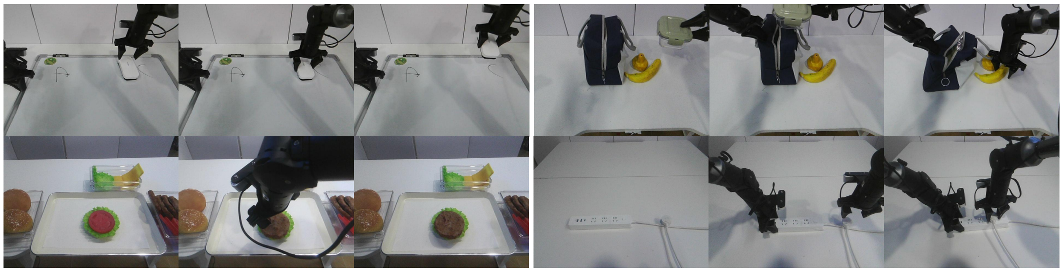


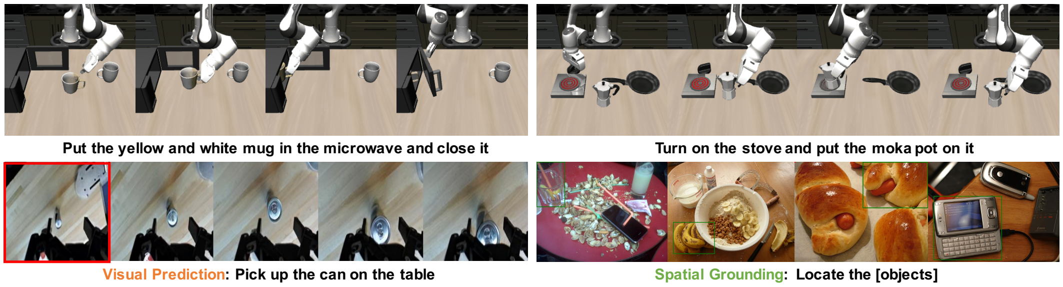
Figure 5: **Real-world task examples.** These include diverse tasks such as wiping a whiteboard, organizing tableware, making a burger, and plugging in a connector.

Real-World Task Collection Table 10 provides a summary of the real-world data collected from the physical robot, recorded at an actual frequency of 30 Hz. A total of 8 tasks were included, with each task collecting approximately 500 trajectories on average. During preprocessing, static frames at the beginning and end of each trajectory were filtered out.

Table 10: **Real-world task trajectories.**

Fold Clothes	Clear Desk	Store Glasses	Food Packing	Pour Water	Clean Blackboard	Insert Plug	Make Hamburger
528	500	500	500	496	500	500	640

Data Processing To reduce redundancy and improve training efficiency, we select keyframes based on thresholding the changes in recorded action joint values. For each selected sequence, the action chunk is normalized by subtracting the joint values of the first frame.



Put the yellow and white mug in the microwave and close it

Turn on the stove and put the moka pot on it

Visual Prediction: Pick up the can on the table

Spatial Grounding: Locate the [objects]

Figure 6: **Multimodal capabilities of UniVLA**. Top: Action outputs for executing long-horizon tasks in the LIBERO benchmark. Bottom: Visual predictions and spatial grounding demonstrating the model's spatiotemporal understanding. The red box marks the current observation; green boxes indicate predicted object detections.

B.2 ALOHA Experiments

For real-world experiments, we perform world model post-training using video data collected from the ALOHA platform, consisting of about 1,000 videos (Pouring water and fold clothes).

Real-world Latency For real-world experiments, model inference is conducted remotely via network communication on an NVIDIA A100 GPU (40GB). The dual-arm robotic platform (AgileX Cobot Magic) receives three image observations, each at a resolution of 128×128 pixels. As shown in Table 11, actions are predicted in discrete chunks of 20 steps, corresponding to approximately a 3.3-second motion window. Each inference step on the model requires roughly 2.1 seconds, and when accounting for communication and data I/O overhead, the total system latency amounts to approximately 3.2 seconds. Consequently, the system generates 20 action steps every 3.2 seconds. These predicted actions are then executed sequentially over the 20 timesteps, with each step interpolated five times to achieve smoother and more precise control.

Table 11: **Inference latency in real-world deployment.**

Setting	**Action Chunk Size**	**Inference Time**	**Total Latency**	**Action Steps/sec**
Real ALOHA	20 steps	2.1 s	3.2 s	6.3
Real ALOHA, w/ vllm	20 steps	1.5 s	2.6 s	7.7

C Autonomous Driving Experiments

NAVSIM Setup The NAVSIM dataset Dauner et al. (2024), resampled from OpenScene to emphasize challenging scenarios, is currently one of the most established end-to-end evaluation benchmarks in the autonomous driving domain. The dataset is divided into two parts: Navtrain and Navtest, comprising 1,192 scenarios for training and validation, and 136 scenarios for testing.

For model training, the input images are resized to a resolution of 512×288. We follow the standard training setup, using the current image frame and ego status to predict trajectories for the next 8 frames. Both the action and ego status are encoded using the fast tokenizer. For post-training of the world model, we leverage six consecutive vision–action pairs, jointly supervising both vision and action outputs.

D Multimodal Capability

As illustrated in Figure 6, we qualitatively showcase the model's ability to interleave multiple modalities—action, language, and vision—within a unified framework. This design enables policy learning for embodied control, spatial reasoning through language output, and future state prediction via visual output, highlighting the model's capacity for generalizable multimodal understanding.

E USE OF LLMS

Large Language Models (LLMs) are used for polishing writing in this manuscript.