LatentVLA: Taming Latent Space for Generalizable and Long-Horizon Bimanual Manipulation

Junming Wang[1]
[1]University of Hong Kong
jmwang@cs.hku.hk

Abstract

Current paradigms for robotic imitation learning face a stark trade-off between the motion fidelity of diffusion models and the data scalability of inverse dynamics models. The latter, while scalable, often learns a latent action space disconnected from physical reality. This flaw leads to critical failures: *temporal entanglement*, where the model cannot distinguish between visually similar states requiring distinct actions, e.g., a gripper approaching versus receding from an object. This ambiguity, compounded by *discretization artifacts* and sensitivity to *task-irrelevant dynamics*, renders robust planning infeasible. We introduce **LatentVLA**, a vision-language-action framework designed to overcome these limitations by learning a continuous and spatiotemporally grounded latent action representation. Its progressive three-stage architecture first employs a **Temporal-Attentive Latent Action Model (TA-LAM)** to resolve ambiguities using language-guided attention and explicit temporal encoding. Subsequently, a **Latent Action Diffusion Transformer (LADT)** performs planning via diffusion directly within this continuous latent space, preserving motion fidelity without tokenization. Finally, an expert policy head translates these latent plans into precise robot actions. Experiments show LatentVLA sets a new state-of-the-art across a suite of real-world bimanual tasks, outperforming prior methods and demonstrating superior zero-shot generalization and few-shot efficiency.

Introduction

The ambition to create general-purpose robots capable of executing complex, long-horizon, dual-arm tasks has driven significant progress in imitation learning. This pursuit has largely bifurcated into two dominant paradigms. On one hand, diffusion-based policies (Chi et al. 2023; Liu et al. 2024, 2025) demonstrate remarkable motion fidelity, yet their reliance on vast, meticulously annotated expert datasets poses a significant barrier to scalability. On the other hand, inverse dynamics models (Ye et al. 2024; Bu et al. 2025a; Chen et al. 2024) offer a path to greater data efficiency by leveraging large quantities of unlabeled video.

However, this scalability comes at the cost of a foundational flaw: ***the learned latent action is often unmoored from the robot's physical and temporal dynamics***, establishing an unreliable bridge between perception and control.

Method	Venue	Data	Discretization Artifacts	Temporal Entanglement	Task-irrelevant Dynamics
DP	RSS'23	🤖	☑	✗	✗
IGOR	arXiv'24	🌐	✗	✗	✗
LAPA	ICLR'25	🌐	✗	✗	✗
Moto	ICCV'25	🌐	✗	✗	✗
LDP	ICML'25	🌐	☑	✗	✗
RDT-1B	ICLR'25	🌐	☑	✗	✗
LatentVLA	—	🤖🌐	☑	☑	☑

🌐 Internet Videos 🤖 Robot Traj. ☑ Solved ✗ Unsolve

Figure 1: LatentVLA's architecture is uniquely engineered to resolve three critical failure modes that plague prior methods. Its continuous latent space circumvents discretization artifacts; its explicit temporal encoding resolves temporal entanglement; and its language-guided attention filters task-irrelevant dynamics. This holistic design enables robust and generalizable performance.

This core deficiency manifests in several systematic failure modes that undermine policy performance, as illustrated in Fig. 1 and Fig. 2. First, the reliance on action tokenization (e.g., VQ-VAE) to enable learning from unlabeled data introduces *discretization artifacts*, which break the representation's connection to continuous physical motion and result in jerky, physically implausible trajectories. This issue is exacerbated by a high sensitivity to *task-irrelevant dynamics*, where spurious visual changes, such as shifting light or camera motion, contaminate the latent space and mislead the policy. Most critically, these models suffer from *temporal entanglement* (or perceptual aliasing), a condition where they collapse visually similar yet contextually distinct states into a single, ambiguous representation. For instance, a gripper approaching an object is observationally similar to it retracting, but these situations demand opposite actions. This ambiguity makes the state-action mapping ill-posed and fundamentally compromises the reliability of long-horizon planning.

Overcoming these limitations requires a conceptual shift manifested in two design principles. First, to resolve temporal entanglement and sensitivity to irrelevant dynamics, the model must achieve a deep **spatiotemporal contextualization** of actions. A latent action cannot merely describe vi-

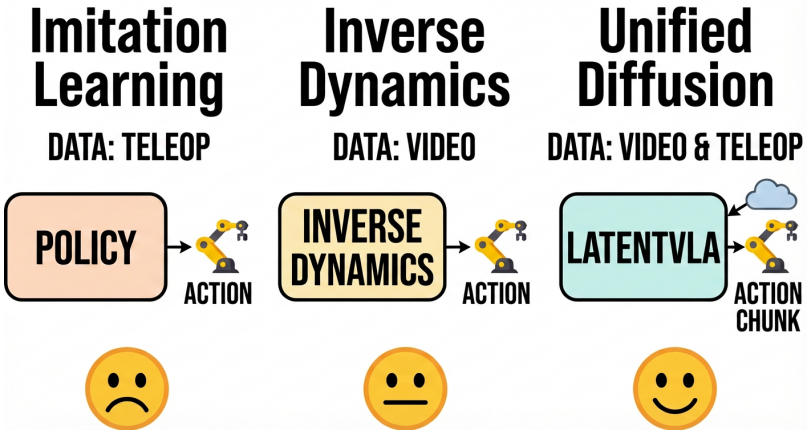

Figure 2: The LatentVLA framework. This paradigm learns a unified latent action space from both expert demonstrations and unlabeled videos, overcoming the generalization limits of prior approaches.

sual change between frames; it must encode physical intent anchored both spatially to task-relevant objects via mechanisms like language-guided attention, and temporally within the broader sequence of a task. Second, to address the trade-off between representational fidelity and data scalability that causes discretization artifacts, we must **decouple representation learning from planning**. Instead of forcing a choice between a high-fidelity continuous space and a scalable discrete one, we can first learn a robust, *continuous* latent action space from heterogeneous data (i.e., both labeled and unlabeled) and subsequently perform planning within this high-quality space.

Grounded in these principles, we introduce **LatentVLA**, a novel framework whose architecture is not an assemblage of components but a deeply integrated, three-stage pipeline designed to systematically resolve these challenges (Fig. 3). The process begins with the **Temporal-Attentive Latent Action Model (TA-LAM)**, which forges a semantically rich and temporally coherent latent space from massive, heterogeneous data. By leveraging language-guided visual attention to filter distractions and explicit temporal encoding to resolve perceptual aliasing, it builds the foundational representation for robust planning. Building upon this foundation, the **Latent Action Diffusion Transformer (LADT)** performs long-horizon planning directly within this continuous latent space, circumventing the discretization artifacts of prior methods. Finally, this high-level latent plan is operationalized by an efficiently fine-tuned expert policy head, which translates the abstract sequence into precise, physically consistent robot actions, bridging the gap between representation and execution. This strategy ensures efficient adaptation while preserving the general knowledge acquired during large-scale pre-training.

To our knowledge, LatentVLA is the first architecture to address these challenges in such a synergistic manner, and comprehensive evaluations across 8 real-world tasks and 3 simulation benchmarks reveal its superior capabilities.

- **LatentVLA is High-Performance.** It achieved state-of-the-art results on 3 public benchmarks (i.e., RoboTwin 1.0, CALVIN, and SIMPLER), and outperformed RDT-1B and LAPA by 7.2% and 20.9% respectively in real-world long-horizon tasks.
- **LatentVLA has Zero-shot Generalization Capabilities.** Benefiting from its temporal-aware representations and data scaling, it maintains a 56% success rate under multiple interference settings in the real world (e.g., lighting variations).
- **LatentVLA is Few-shot Efficient.** Due to its enhanced representation learning, it transfers to novel objects with only 20 expert demonstrations, achieving 27% higher success rates than RDT-1B on out-of-domain tasks (e.g., folding a long towel).

Related Work

Vision-Language-Action (VLA) Models. Recent Vision-Language-Action (VLA) models (Zhen et al. 2025; Xie et al. 2025; Bu et al. 2025b) have achieved significant success in translating multimodal observations and language instructions into robotic actions (Chen et al. 2024; Bu et al. 2025a; Wang et al. 2024b; Zhang et al. 2024; Guan et al. 2023). Seminal works like RT-1 (Brohan et al. 2022) established a paradigm of pre-trained encoders followed by action decoders. This was advanced by RT-2 (Brohan et al. 2023) and OpenVLA (Kim et al. 2024), which unified vision, language, and action into a shared token space, leveraging the capabilities of Large Language Models (LLMs). To enhance generalization, subsequent frameworks such as RoboFlamingo (Li et al. 2023), Octo (Ghosh et al. 2023), and RT-X (Collaboration et al. 2023) aggregated vast, diverse datasets across multiple embodiments. A primary limitation of these models, however, is their reliance on large-scale, action-annotated data, which poses a significant bottleneck to scalability.

Diffusion-Based VLA Models. Diffusion-based (Gu et al. 2024; Wang et al. 2024a; Song et al. 2025) robotic control has evolved through successive architectural refinements. Diffusion Policy (Chi et al. 2023) pioneered denoising processes for action generation, while RDT

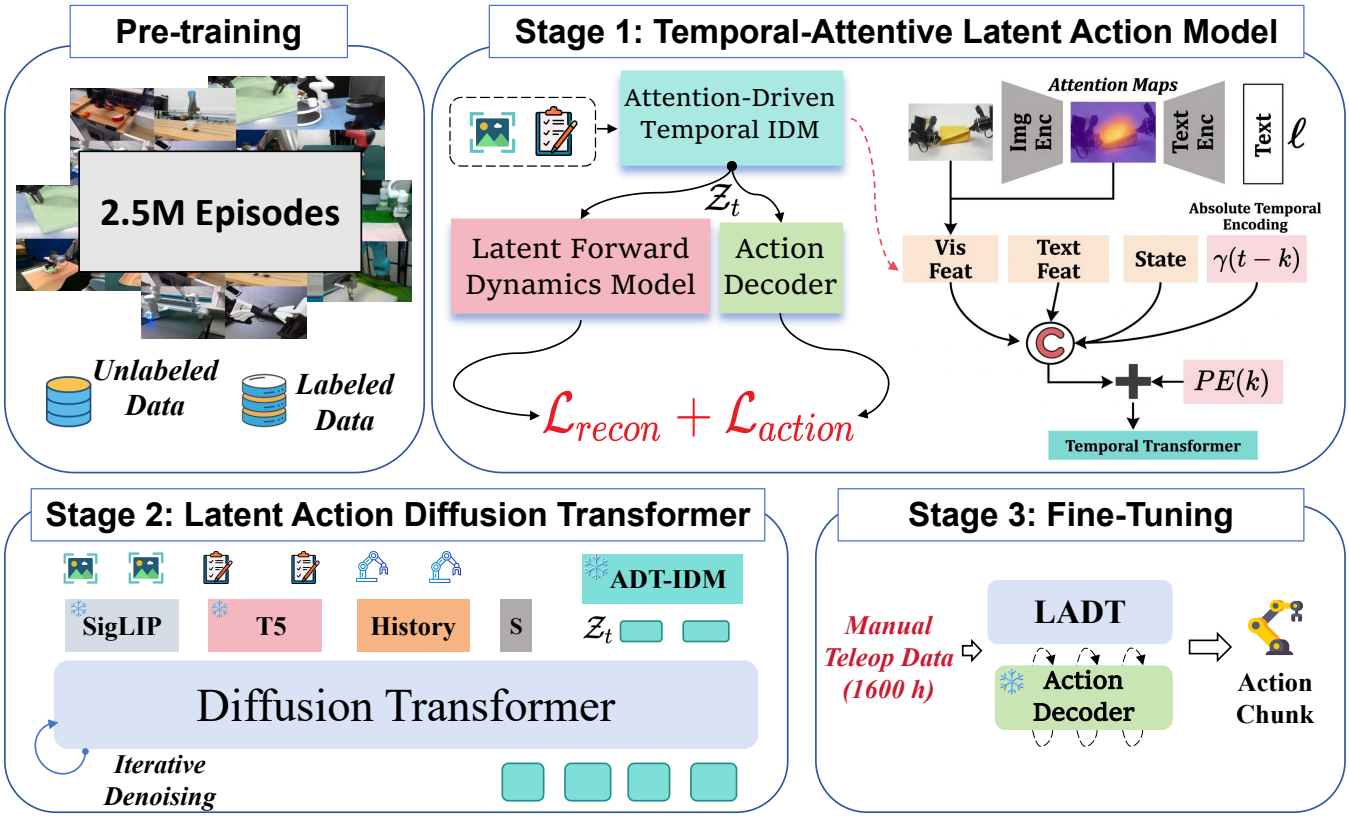

Figure 3: LatentVLA Overview. The framework comprises three stages: (1) a TA-LAM extracts semantically rich latent representations via Language-guided attention; (2) an LADT generates coherent latent plans from visual-language inputs through diffusion; and (3) a low-level Expert Policy maps these plans to precise bimanual actions.

(Liu et al. 2024) extended this approach with vision-conditioned transformers. HybridVLA (Liu et al. 2025) introduced autoregressive-diffusion integration through co-training mechanisms, while the π-series advanced the paradigm through flow-matched experts (π_0 (Black et al. 2024)) and heterogeneous cross-embodiment training ($\pi_{0.5}$ (Intelligence et al. 2025)), achieving impressive generalization across unseen environments.

Robotic Learning with Latent Video Representations. A parallel line of research investigates learning latent action spaces from unlabeled video corpora (Hu et al. 2024). Models like Genie (Bruce et al. 2024), LAPO (Schmidt and Jiang 2023), and DynaMo (Cui et al. 2024) pioneered unsupervised action discovery through frame prediction or direct latent acquisition from visual sequences. Similarly, LAPA (Ye et al. 2024), Moto (Chen et al. 2024), and IGOR enabled VLA systems to leverage discrete latent representations (Jang et al. 2025; Bai et al. 2025; Wang and Shi 2024) from demonstration videos. A common pitfall in these methods is the indiscriminate encoding of all visual changes, inadvertently capturing task-irrelevant environmental dynamics. *In contrast, LatentVLA is designed to explicitly disentangle task-centric dynamics from such extraneous distractors. By separately modeling temporal context, we establish a structured and robust latent space suitable for effective policy learning and transfer.*

Method

We introduce LatentVLA, a hierarchical framework for bimanual manipulation that decouples representation from planning. It employs a three-stage pipeline: (1) learning a continuous latent space via TA-LAM from heterogeneous data, (2) long-horizon planning with LADT, and (3) fine-tuning for execution. Leveraging 2.5M+ multi-modal examples, LatentVLA achieves superior generalization.

Problem Formulation

We formulate the problem as learning a generalizable bimanual manipulation policy from two distinct data sources: a large-scale pre-training dataset, $\mathcal{D}_{\text{pretrain}}$, and a separate fine-tuning dataset, $\mathcal{D}_{\text{finetune}}$. Our pre-training dataset, $\mathcal{D}_{\text{pretrain}}$, unites over 2.5 million sequences from 52 sources, systematically combining four data categories: unlabeled human videos to learn rich visual dynamics; labeled multi-embodiment robot trajectories from the RDT-1B corpus for broad action supervision; simulation data for enhanced robustness; and our large-scale, real-world demonstrations, comprising the AgiBot World Beta and LatentVLA-Dexterous sets ($\sim$0.5M sequences). In contrast, $\mathcal{D}_{\text{finetune}}$ is a held-out collection of 1600 hours of new demonstrations

across 8 specific bimanual tasks, reserved exclusively for target-platform adaptation. Each trajectory consists of multi-view observations $\mathbf{O}_t \in \mathbb{R}^{V\times H\times W\times 3}$, proprioceptive states $\mathbf{s}_t \in \mathbb{R}^{14}$, a language instruction ℓ, and, when available, bimanual actions $\mathbf{a}_t \in \mathbb{R}^{14}$.

Stage 1: Temporal-Attentive Latent Action Model

The foundational goal of Stage 1 is to build a unified latent action space from the entirety of our heterogeneous pre-training dataset, $\mathcal{D}_{\text{pretrain}}$. To achieve this, our Temporal-Attentive Latent Action Model (TA-LAM) learns to infer a single-step latent action Z_t that is both semantically meaningful and temporally disentangled. TA-LAM comprises three core modules—an Inverse Dynamics Model, a Forward Dynamics Model, and an Action Decoder—which are optimized jointly to leverage all available data signals.

Attention-Driven Inverse Dynamics Model (IDM) The IDM is the cornerstone of our representation learning. Its goal is to infer a single-step latent action $Z_t \in \mathbb{R}^{d_z}$ that explains the observed state transition from time t to $t+1$. To achieve this, it processes a history of multi-modal inputs, $(O_{t-h:t}, s_{t-h:t})$, to produce Z_t, where $d_z = 512$ and the history length is $h = 4$. To ensure this latent action is robust against environmental distractors and temporal ambiguity, the IDM integrates two synergistic mechanisms.

Language-Guided Spatial Attention. This mechanism leverages a language instruction ℓ to focus on task-relevant visual elements, filtering out environmental distractors. Our approach employs a unified, pre-trained vision-language model, SigLIP (Zhai et al. 2023), as the perceptual backbone. Its text encoder, E_{text}, embeds the instruction, while its image encoder plays a dual role. First, as E_{img}, it generates image features to compute an attention mask $A_t^{(v)}$ for each multi-view image $I_t^{(v)}$ via dot-product similarity with the text embedding. Second, as the primary feature extractor E_{vis}, it processes the attention-masked images to produce the final, task-focused visual features f_t^{visual}.

$$A_t^{(v)} = \text{softmax}\left(\frac{E_{\text{img}}(I_t^{(v)}) \cdot E_{\text{text}}(\ell)^T}{\sqrt{d_{\text{feat}}}}\right) \tag{1}$$

$$f_t^{\text{visual}} = \text{Concat}_{v\in\{\text{views}\}}\Big(E_{\text{vis}}\big(I_t^{(v)} \odot \text{Upsample}(A_t^{(v)})\big)\Big) \tag{2}$$

Here, E_{img}, E_{text}, and E_{vis} are components of the same frozen SigLIP model, ensuring representational consistency. d_{feat} denotes the dimensionality of the SigLIP feature space. This unified architecture allows the model to ground visual perception in language commands effectively.

Temporal Disentanglement. To resolve temporal entanglement—where visually similar states from different task phases are confused—the IDM explicitly models both short-term dynamics and long-term task progression. Given a history of spatially-focused features $f_{t-h:t}^{\text{visual}}$ and proprioceptive states $s_{t-h:t}$, the model first constructs a context sequence. Each element C_{t-k} in this sequence is formed by concatenating the visual, language, and proprioceptive features for that timestep.

Crucially, to provide a clear signal of task progression, we also include an absolute temporal encoding $\gamma(t-k)$, which encodes the timestep's index within the entire episode. This entire bundle is then augmented with a standard sinusoidal positional encoding $\text{PE}(k)$ to embed its relative position within the context window (where $h = 4$, so the window size is 5). A Transformer encoder T_{enc} then processes this entire sequence to produce the final latent action Z_t.

$$C_{t-k} = \text{Concat}(f_{t-k}^{\text{visual}}, E_\ell(\ell), s_{t-k}, \gamma(t-k)) + \text{PE}(k), \quad k \in \{0,\dots,h\} \tag{3}$$

$$Z_t = \text{LatentHead}(T_{\text{enc}}(C_{t-h},\dots,C_t)) \tag{4}$$

The absolute temporal encoding $\gamma(t)$ and the relative positional encoding $\text{PE}(k)$ are both instances of the standard sinusoidal function (Vaswani et al. 2017), defined for a position p and dimension j as:

$$\text{PosEnc}(p,j,d) = \begin{cases}\sin(p/10000^{j/d}) & \text{if } j \text{ is even}\\ \cos(p/10000^{(j-1)/d}) & \text{if } j \text{ is odd}\end{cases} \tag{5}$$

For absolute encoding, we set $p = t$ and $d = d_\gamma$, while for relative encoding, we set $p = k$ and $d = d_{\text{pe}}$.

These two mechanisms operate in synergy: spatial attention isolates *what* is task-relevant, absolute time encoding informs the model of the current task phase, and relative positional encoding models the *evolution* of these salient features within the recent past. This design ensures the model disentangles the dynamics of relevant objects, not background noise, yielding a purified latent action Z_t that encodes a directed, task-relevant state transition.

Latent Forward Dynamics Model (FDM) The FDM enforces a key constraint: the latent action Z_t must contain sufficient information to predict the future state. This is crucial for learning from unlabeled video. It takes the latent action Z_t and the final hidden state from the IDM's Transformer to predict the subsequent visual observation $\hat{O}_{t+1}$:

$$\hat{O}_{t+1} = D_{\text{obs}}(Z_t, T_{\text{enc}}(\cdot)_{\text{last\_hidden}}) \tag{6}$$

This prediction forms the basis of our unsupervised reconstruction loss, enabling TA-LAM to learn from vast quantities of human videos.

Action Decoder The Action Decoder grounds the abstract latent space by mapping latent actions to concrete robot commands. It employs a lightweight Multi-Layer Perceptron (MLP) to translate the latent action Z_t into a continuous robot action $\hat{a}_t \in \mathbb{R}^{14}$:

$$\hat{a}_t = D_{\text{act}}(Z_t) \tag{7}$$

This decoder is supervised exclusively on labeled robot data, bridging representation and physical execution.

Joint Optimization Objective TA-LAM's effectiveness stems from its joint training on both supervised and unsupervised signals harvested from $\mathcal{D}_{\text{pretrain}}$. The overall loss is

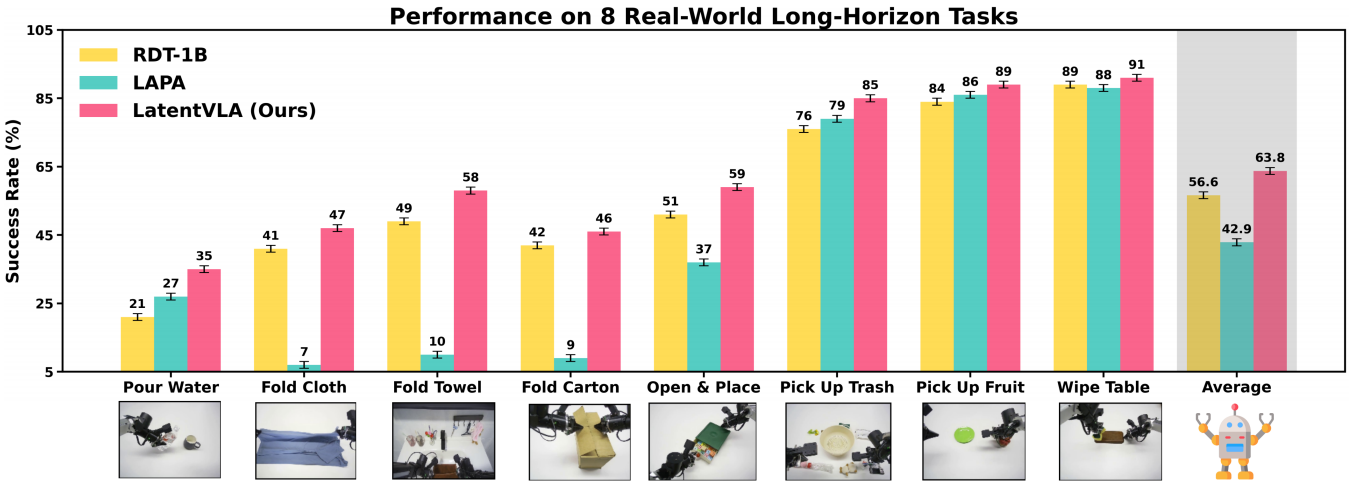

Figure 4: Real-world bimanual task performance. Success rates on 8 challenging real-world bimanual tasks demonstrate LatentVLA's superior performance under in-distribution (ID) settings.

a weighted sum of two Mean Squared Error (MSE) objectives:

$$\mathcal{L}_{\text{TA-LAM}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{act}} \mathcal{L}_{\text{action}} \tag{8}$$

The **reconstruction loss** $\mathcal{L}_{\text{recon}}$ is applied to **all sequences** in $\mathcal{D}_{\text{pretrain}}$, including unlabeled human videos. It measures the MSE between the predicted and ground-truth next observation, forcing the model to learn world dynamics:

$$\mathcal{L}_{\text{recon}} = \|\hat{O}_{t+1} - O_{t+1}\|_2^2 \tag{9}$$

The **action loss** $\mathcal{L}_{\text{action}}$ is applied **only to the labeled subsets** of $\mathcal{D}_{\text{pretrain}}$ ($\mathcal{D}_{\text{robot}}$, $\mathcal{D}_{\text{sim}}$, $\mathcal{D}_{\text{real}}$). It measures the MSE between the decoded and ground-truth action, grounding the latent space in physical embodiment:

$$\mathcal{L}_{\text{action}} = \|\hat{a}_t - a_t\|_2^2 \tag{10}$$

This dual-loss framework yields a universally applicable latent action space, enriched by diverse video dynamics and grounded by multi-embodiment robot data.

Stage 2: Latent Action Diffusion Transformer

The second stage transitions from single-step representation learning to long-horizon planning. Our approach, the Latent Action Diffusion Transformer (LADT), performs generative planning directly within the continuous latent space learned by the frozen TA-LAM. A core design choice is to operate in this continuous domain, which circumvents the information loss and "discretization artifacts" inherent in VQ-based methods. This allows LADT to model the subtle and fine-grained variations essential for precise manipulation, resulting in smoother and more physically plausible action sequences.

To train LADT, we leverage **only the labeled subsets of the pre-training data**, denoted as $\mathcal{D}_{\text{labeled}} = \mathcal{D}_{\text{robot}} \cup \mathcal{D}_{\text{sim}} \cup \mathcal{D}_{\text{real}}$. We first use the frozen IDM from Stage 1 to encode all expert trajectories in $\mathcal{D}_{\text{labeled}}$ into sequences of continuous latent actions. LADT, a diffusion model, is then conditioned on past observations $O_{t-h:t}$, prior latent actions $Z_{t-h':t-1}$, proprioception s_t, and the language instruction ℓ. It is trained to predict an entire future sequence of $H = 16$ latent actions, $Z_{\text{future}} = \{Z_t, \ldots, Z_{t+H-1}\}$, by reversing a Gaussian noising process. The training objective is a standard denoising score-matching loss:

$$\mathcal{L}_{\text{LADT}} = \mathbb{E}_{k,\epsilon,c_t}\left[\|\epsilon - \epsilon_\theta(Z^{\text{noisy}}_{\text{future},k}, k, c_t)\|_2^2\right] \tag{11}$$

where $Z^{\text{noisy}}_{\text{future},k}$ is the latent sequence corrupted with noise ϵ at diffusion step k, and c_t is the multimodal conditioning context. The inclusion of latent action history $Z_{t-h':t-1}$ provides a strong sequential prior, enhancing the temporal coherence of the generated plan. To ensure the language instruction robustly guides the planning process, we employ specialized conditioning mechanisms within the Transformer architecture, such as alternating the injection of visual and language features across layers to prevent the dense visual information from overshadowing the instructional signal.

Stage 3: Real-World Fine-tuning

The final stage adapts the pretrained generalist model to the target platform and a specific set of tasks. For this, we fine-tune the LADT planner from Stage 2 using our **held-out fine-tuning dataset,** $\mathcal{D}_{\text{finetune}}$, which consists of 1600 hours of new demonstrations across 8 bimanual tasks. A key aspect of our strategy is **freezing the entire TA-LAM model**—including both the IDM for latent encoding and the Action Decoder for command generation. This strategy isolates adaptation strictly to the high-level planning module (LADT), preserving the rich, generalizable representations from pre-training. By doing so, we ensure stable and highly sample-efficient specialization. The fine-tuning objective directly optimizes the policy by minimizing the MSE between the decoded action sequence and the ground-truth expert ac-

tions $A^{\text{gt}} = \{a_t, \ldots, a_{t+H-1}\}$ from $\mathcal{D}_{\text{finetune}}$:

$$\mathcal{L}_{\text{finetune}} = \mathbb{E}_{\substack{(O_{t-h:t}, s_t, A^{\text{gt}}, \ell) \\ \sim \mathcal{D}_{\text{finetune}}}} \left[\| D_{\text{act}}^{\text{frozen}}(\text{LADT}_\theta(c_t)) - A^{\text{gt}} \|_2^2 \right] \quad (12)$$

where c_t is the conditioning context. This targeted approach ensures efficient knowledge transfer for high-performance real-world execution.

Experiments

We conduct a comprehensive suite of experiments to validate LatentVLA, structured around four central questions: dissecting the efficacy of our architectural components (**Q1**), benchmarking generalization against state-of-the-art methods (**Q2**), assessing the ability to leverage large-scale data (**Q3**), and examining few-shot adaptation to novel tasks (**Q4**).

Experimental Setup. Our evaluation protocol is designed for rigor and transparency, encompassing large-scale pre-training, fine-tuning, and real-time deployment. We benchmark LatentVLA against leading models, including RDT-1B (Liu et al. 2024) and LAPA (Ye et al. 2024), re-training all baselines from scratch on our full dataset to ensure a fair comparison.

Training Protocol and Datasets. Our model is trained via a three-stage protocol designed to leverage our defined datasets, $\mathcal{D}_{\text{pretrain}}$ and $\mathcal{D}_{\text{finetune}}$. The first two stages constitute an extensive pre-training phase, which was conducted on a cluster of 112 NVIDIA A100-80G GPUs and took approximately one month.

- **Stage 1 & 2 (Pre-training):** First, a multi-modal encoder is trained on the entire $\mathcal{D}_{\text{pretrain}}$ corpus to learn a unified representation. Following this, a latent diffusion planner is trained on a labeled subset of $\mathcal{D}_{\text{pretrain}}$ to master long-horizon action generation.
- **Stage 3 (Fine-tuning):** The pre-trained model is then fine-tuned for 100,000 iterations. This adaptation stage utilizes a mixture of public simulation benchmarks and our held-out demonstration dataset, $\mathcal{D}_{\text{finetune}}$, to specialize the model for the target bimanual tasks.

Real-World Task Suite and OOD Probes. Our real-world evaluation is centered on a suite of **8 challenging bimanual tasks**. To test zero-shot generalization, we designed targeted Out-of-Distribution (OOD) probes, including systematic variations in lighting and background, the introduction of novel objects (e.g., different cups, towels, clothes), and dynamic disturbances during task execution.

Simulation Benchmarks. We also employ three standard simulation benchmarks to ensure comprehensive comparison: **RoboTwin 1.0** (Mu et al. 2024) for complex bimanual coordination, **SIMPLER** (Li et al. 2024) for long-horizon manipulation, and **CALVIN (ABC→D)** (Mees et al. 2022) for compositional generalization.

Comparative Performance Analysis

Real-World Bimanual Manipulation In our suite of 8 challenging real-world tasks (Fig. 4, LatentVLA achieves a mean success rate of 63.8%, decisively outperforming the re-trained RDT-1B and LAPA by absolute margins of 12.8% and 48.5%, respectively. The true advantage of our method becomes clear when dissecting task-specific performance. LatentVLA shows its largest gains in tasks with deformable objects, such as *Fold Cloth* (47% vs. LAPA's 7%) and *Fold Towel* (58% vs. LAPA's 10%). These tasks require not just a sequence of actions, but a continuous understanding of state changes. We attribute this success to our progressive refinement architecture: the TA-LAM module first abstracts a robust, goal-oriented latent representation, while the LADT policy planner then generates a smooth, temporally coherent action sequence to achieve it, avoiding the rigid, often fragile plans of token-based methods.

Simulation Benchmarks. LatentVLA's architectural strengths are decisively validated across a suite of standard simulation benchmarks (Table 1). It excels in complex bimanual coordination on **RoboTwin 1.0**, surpassing DP3 by 26.0%. More importantly, it overcomes the characteristic failure modes of discrete methods in long-horizon tasks. On **SIMPLER**, its absolute temporal encoding resolves visual state ambiguities to outperform LAPA by 26.1%, while on **CALVIN**, its continuous latent planning mitigates the compounding quantization errors that plague discrete models, achieving a 13.5% higher success rate than Moto.

Zero-Shot Generalization

The advantage of *LatentVLA*'s continuous latent space is most evident under severe out-of-distribution (OOD) conditions. Baselines like *LAPA*, which use a discrete latent space, suffer from **catastrophic failures** under dynamic lighting; minor visual shifts can cause observations to cross a representational "cliff," mapping to incorrect latent tokens. *LatentVLA* remains robust by design. Its attention-guided *TA-LAM* filters out irrelevant distractors (e.g., lighting), while the continuous nature of its latent space ensures **graceful degradation**: small input perturbations lead to correspondingly small shifts in the latent representation. This robustness is critical for deployment in uncontrolled real-world environments.

Data Scaling and Few-Shot Adaptation

Data and Architectural Efficiency (Q3). We conducted two experiments to disentangle architectural merit from data advantage. First, we trained LatentVLA from scratch on the original 1.3M RDT-1B dataset. On this level playing field, our model still outperformed the original RDT-1B by 5.6% on average, proving the inherent efficiency of our architecture. Second, we fine-tuned the pre-trained RDT-1B model with our 1600 hours of real-world data. Our full LatentVLA model, trained on all 2.5M trajectories, surpassed this fine-tuned baseline by 7.2%, demonstrating its superior ability to effectively harness large-scale, diverse data.

Few-Shot Adaptation to Novel Objects (Q4). We evaluated LatentVLA's adaptability by fine-tuning it on a novel task variant: folding a long, previously unseen towel. With only **20 demonstrations** (under one hour of fine-tuning on a single A100 GPU), LatentVLA achieved a **61% success**

Task	DP	DP3	**LatentVLA**
Block Hammer Beat	-	58.3	**87.0**
Block Handover	-	85.0	**98.3**
Bottle Adjust	24.7	70.7	**76.7**
Container Place	16.3	74.0	**83.6**
Dual Bottles Pick (E)	26.7	60.3	**70.7**
Dual Bottles Pick (H)	32.3	48.0	**53.6**
Empty Cup Place	14.7	70.3	**75.7**
Pick Apple Messy	6.0	10.7	**33.7**

(a) RoboTwin 1.0 Success Rate (%)

Method	Rate (%)
Octo	16.9
OpenVLA	24.8
LAPA	52.1
RT-2	60.7
Moto	61.4
LatentVLA	**65.7**

(b) SIMPLER

Method	Avg. Len
RoboFlamingo	2.47
GR-1	3.06
Moto	3.10
LatentVLA	**3.52**

(c) CALVIN

Table 1: Experimental results on RoboTwin 1.0, SIMPLER, and CALVIN benchmarks. Our method, LatentVLA, achieves superior performance across all tasks. Note that LatentVLA demonstrates significant improvements in long-horizon tasks in CALVIN.

Model Variant	**Succ%**	**Δ**
LatentVLA (Full Model)	**58.0**	**–**
Ablating Core Pillars		
(1) w/ VQ-VAE + Autoregressive GPT	25.0	-33.0
(2) w/o Absolute Temporal Encoding	31.0	-27.0
Ablating Grounding		
(3) w/o Joint Training (Video only)	35.0	-23.0
(4) w/o Language-Guided Attention	39.0	-19.0
Ablating Refinement		
(5) w/o Latent History Conditioning	45.0	-13.0
RDT-1B (re-trained)	49.0	-9.0
LAPA (re-trained)	10.0	-48.0

Table 2: Ablation on the real-world towel folding task. Performance drops (Δ) reveal a clear hierarchy of importance.

rate. This was achieved by freezing the pre-trained TALAM representation and only updating the LADT planner, validating our framework's capacity for rapid, sample-efficient specialization to new object properties.

Ablation Study

An ablation study on the towel folding task (Table 2) quantified the architectural pillars of LatentVLA's success (Q1). Results reveal a clear hierarchy of importance. The core planning mechanism is most critical: replacing the continuous diffusion planner (variant 1) or removing absolute temporal encoding (variant 2) caused the most severe performance drops (-33.0% and -27.0%, respectively). These components are essential for fine-grained control and long-horizon coherence. Next, embodied grounding is crucial. Training on video alone (variant 3, -23.0%) or removing language-guided attention (variant 4, -19.0%) led to physically implausible actions and loss of object focus. Lastly, policy refinement remains vital, as removing latent history (variant 5) still degraded performance by 13.0%. Collectively, these results demonstrate that LatentVLA's effectiveness derives from the tight integration of its components, not any single element.

Conclusion and Limitations

We introduced LatentVLA, a framework achieving state-of-the-art performance in long-horizon bimanual manipulation. By integrating temporally-aware representations with latent diffusion planning, our model excels in performance, generalization, and sample efficiency, effectively leveraging both labeled and unlabeled data.

Limitations and Future Work. Primary limitations include sensitivity to complete object occlusions and visual ambiguities from highly reflective surfaces.

References

Bai, S.; Zhou, W.; Ding, P.; Zhao, W.; Wang, D.; and Chen, B. 2025. Rethinking Latent Redundancy in Behavior Cloning: An Information Bottleneck Approach for Robot Manipulation. In *Forty-second International Conference on Machine Learning*.

Black, K.; Brown, N.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Groom, L.; Hausman, K.; Ichter, B.; et al. 2024. pi0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; et al. 2023. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Hsu, J.; et al. 2022. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*.

Bruce, J.; Dennis, M. D.; Edwards, A.; Parker-Holder, J.; Shi, Y.; Hughes, E.; Lai, M.; Mavalankar, A.; Steigerwald, R.; Apps, C.; et al. 2024. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*.

Bu, Q.; Cai, J.; Chen, L.; Cui, X.; Ding, Y.; Feng, S.; Gao, S.; He, X.; Huang, X.; Jiang, S.; et al. 2025a. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*.

Bu, Q.; Yang, Y.; Cai, J.; Gao, S.; Ren, G.; Yao, M.; Luo, P.; and Li, H. 2025b. Learning to Act Anywhere with Task-centric Latent Actions. *arXiv preprint arXiv:2502.14420*.

Chen, Y.; Ge, Y.; Li, Y.; Ge, Y.; Ding, M.; Shan, Y.; and Liu, X. 2024. Moto: Latent motion token as the bridging language for robot manipulation. *arXiv preprint arXiv:2412.04445*.

Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2023. Diffusion policy: Visuomotor policy learning via action diffusion. *Robotics: Science and Systems*.

Collaboration, O. X.-E.; O'Neill, A.; Rehman, A.; Maddukuri, A.; Gupta, A.; Padalkar, A.; Lee, A.; Pooley, A.; Gupta, A.; Mandlekar, A.; Jain, A.; Tung, A.; Bewley, A.; Herzog, A.; Irpan, A.; Khazatsky, A.; Rai, A.; Gupta, A.; Wang, A.; Kolobov, A.; Singh, A.; Garg, A.; Kembhavi, A.; Xie, A.; Brohan, A.; Raffin, A.; Sharma, A.; Yavary, A.; Jain, A.; Balakrishna, A.; Wahid, A.; Burgess-Limerick, B.; Kim, B.; Schölkopf, B.; Wulfe, B.; Ichter, B.; Lu, C.; Xu, C.; Le, C.; Finn, C.; Wang, C.; Xu, C.; Chi, C.; Huang, C.; Chan, C.; Agia, C.; Pan, C.; Fu, C.; Devin, C.; Xu, D.; Morton, D.; Driess, D.; Chen, D.; Pathak, D.; Shah, D.; Büchler, D.; Jayaraman, D.; Kalashnikov, D.; Sadigh, D.; Johns, E.; Foster, E.; Liu, F.; Ceola, F.; Xia, F.; Zhao, F.; Frujeri, F. V.; Stulp, F.; Zhou, G.; Sukhatme, G. S.; Salhotra, G.; Yan, G.; Feng, G.; Schiavi, G.; Berseth, G.; Kahn, G.; Yang, G.; Wang, G.; Su, H.; Fang, H.-S.; Shi, H.; Bao, H.; Amor, H. B.; Christensen, H. I.; Furuta, H.; Walke, H.; Fang, H.; Ha, H.; Mordatch, I.; Radosavovic, I.; Leal, I.; Liang, J.; Abou-Chakra, J.; Kim, J.; Drake, J.; Peters, J.; Schneider, J.; Hsu, J.; Bohg, J.; Bingham, J.; Wu, J.; Gao, J.; Hu, J.; Wu, J.; Wu, J.; Sun, J.; Luo, J.; Gu, J.; Tan, J.; Oh, J.; Wu, J.; Lu, J.; Yang, J.; Malik, J.; Silvério, J.; Hejna, J.; Booher, J.; Tompson, J.; Yang, J.; Salvador, J.; Lim, J. J.; Han, J.; Wang, K.; Rao, K.; Pertsch, K.; Hausman, K.; Go, K.; Gopalakrishnan, K.; Goldberg, K.; Byrne, K.; Oslund, K.; Kawaharazuka, K.; Black, K.; Lin, K.; Zhang, K.; Ehsani, K.; Lekkala, K.; Ellis, K.; Rana, K.; Srinivasan, K.; Fang, K.; Singh, K. P.; Zeng, K.-H.; Hatch, K.; Hsu, K.; Itti, L.; Chen, L. Y.; Pinto, L.; Fei-Fei, L.; Tan, L.; Fan, L. J.; Ott, L.; Lee, L.; Weihs, L.; Chen, M.; Lepert, M.; Memmel, M.; Tomizuka, M.; Itkina, M.; Castro, M. G.; Spero, M.; Du, M.; Ahn, M.; Yip, M. C.; Zhang, M.; Ding, M.; Heo, M.; Srirama, M. K.; Sharma, M.; Kim, M. J.; Kanazawa, N.; Hansen, N.; Heess, N.; Joshi, N. J.; Suenderhauf, N.; Liu, N.; Palo, N. D.; Shafiullah, N. M. M.; Mees, O.; Kroemer, O.; Bastani, O.; Sanketi, P. R.; Miller, P. T.; Yin, P.; Wohlhart, P.; Xu, P.; Fagan, P. D.; Mitrano, P.; Sermanet, P.; Abbeel, P.; Sundaresan, P.; Chen, Q.; Vuong, Q.; Rafailov, R.; Tian, R.; Doshi, R.; Mart'in-Mart'in, R.; Baijal, R.; Scalise, R.; Hendrix, R.; Lin, R.; Qian, R.; Zhang, R.; Mendonca, R.; Shah, R.; Hoque, R.; Julian, R.; Bustamante, S.; Kirmani, S.; Levine, S.; Lin, S.; Moore, S.; Bahl, S.; Dass, S.; Sonawani, S.; Song, S.; Xu, S.; Haldar, S.; Karamcheti, S.; Adebola, S.; Guist, S.; Nasiriany, S.; Schaal, S.; Welker, S.; Tian, S.; Ramamoorthy, S.; Dasari, S.; Belkhale, S.; Park, S.; Nair, S.; Mirchandani, S.; Osa, T.; Gupta, T.; Harada, T.; Matsushima, T.; Xiao, T.; Kollar, T.; Yu, T.; Ding, T.; Davchev, T.; Zhao, T. Z.; Armstrong, T.; Darrell, T.; Chung, T.; Jain, V.; Vanhoucke, V.; Zhan, W.; Zhou, W.; Burgard, W.; Chen, X.; Chen, X.; Wang, X.; Zhu, X.; Geng, X.; Liu, X.; Liangwei, X.; Li, X.; Pang, Y.; Lu, Y.; Ma, Y. J.; Kim, Y.; Chebotar, Y.; Zhou, Y.; Zhu, Y.; Wu, Y.; Xu, Y.; Wang, Y.; Bisk, Y.; Dou, Y.; Cho, Y.; Lee, Y.; Cui, Y.; Cao, Y.; Wu, Y.-H.; Tang, Y.; Zhu, Y.; Zhang, Y.; Jiang, Y.; Li, Y.; Li, Y.; Iwasawa, Y.; Matsuo, Y.; Ma, Z.; Xu, Z.; Cui, Z. J.; Zhang, Z.; Fu, Z.; and Lin, Z. 2023. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. https://arxiv.org/abs/2310.08864.

Cui, Z.; Pan, H.; Iyer, A.; Haldar, S.; and Pinto, L. 2024. Dynamo: In-domain dynamics pretraining for visuo-motor control. *Advances in Neural Information Processing Systems*, 37: 33933–33961.

Ghosh, D.; Walke, H.; Pertsch, K.; Black, K.; Mees, O.; Dasari, S.; Hejna, J.; Xu, C.; Luo, J.; et al. 2023. Octo: An open-source generalist robot policy.

Gu, S.; Yin, W.; Jin, B.; Guo, X.; Wang, J.; Li, H.; Zhang, Q.; and Long, X. 2024. Dome: Taming diffusion model into high-fidelity controllable occupancy world model. *arXiv preprint arXiv:2410.10429*.

Guan, X.; Wang, J.; Sun, Z.; Zhang, Z.; Duan, T.; Deng, S.; Liu, F.; and Cui, H. 2023. New Problems in Active Sampling for Mobile Robotic Online Learning. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1155–1160.

Hu, Y.; Guo, Y.; Wang, P.; Chen, X.; Wang, Y.-J.; Zhang, J.; Sreenath, K.; Lu, C.; and Chen, J. 2024. Video Prediction Policy: A Generalist Robot Policy with Predictive Visual Representations. *arXiv preprint arXiv:2412.14803*.

Intelligence, P.; Black, K.; Brown, N.; Darpinian, J.; Dhabalia, K.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; et al. 2025. pi0.5: a Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*.

Jang, J.; Ye, S.; Lin, Z.; Xiang, J.; Bjorck, J.; Fang, Y.; Hu, F.; Huang, S.; Kundalia, K.; Lin, Y.-C.; et al. 2025. DreamGen: Unlocking Generalization in Robot Learning through Video World Models. *arXiv preprint arXiv:2505.12705*.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; et al. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246*.

Li, X.; Hsu, K.; Gu, J.; Pertsch, K.; Mees, O.; Walke, H. R.; Fu, C.; Lunawat, I.; Sieh, I.; Kirmani, S.; et al. 2024. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*.

Li, X.; Liu, M.; Zhang, H.; Yu, C.; Xu, J.; Wu, H.; Cheang, C.; Jing, Y.; Zhang, W.; Liu, H.; et al. 2023. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*.

Liu, J.; Chen, H.; An, P.; Liu, Z.; Zhang, R.; Gu, C.; Li, X.; Guo, Z.; Chen, S.; Liu, M.; et al. 2025. HybridVLA: Collaborative Diffusion and Autoregression in a Unified Vision-Language-Action Model. *arXiv preprint arXiv:2503.10631*.

Liu, S.; Wu, L.; Li, B.; Tan, H.; Chen, H.; Wang, Z.; Xu, K.; Su, H.; and Zhu, J. 2024. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*.

Mees, O.; Hermann, L.; Rosete-Beas, E.; and Burgard, W. 2022. CALVIN: A Benchmark for Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks. *IEEE Robotics and Automation Letters (RA-L)*, 7(3): 7327–7334.

Mu, Y.; Chen, T.; Peng, S.; Chen, Z.; Gao, Z.; Zou, Y.; Lin, L.; Xie, Z.; and Luo, P. 2024. Robotwin: Dual-arm robot benchmark with generative digital twins (early version). *arXiv preprint arXiv:2409.02920*.

Schmidt, D.; and Jiang, M. 2023. Learning to act without actions. *arXiv preprint arXiv:2312.10812*.

Song, Z.; Jia, C.; Liu, L.; Pan, H.; Zhang, Y.; Wang, J.; Zhang, X.; Xu, S.; Yang, L.; and Luo, Y. 2025. Don't Shake the Wheel: Momentum-Aware Planning in End-to-End Autonomous Driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22432–22441.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; and Shi, Y. 2024. Neurncd: Novel class discovery via implicit neural representation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 257–265.

Wang, J.; Zhang, X.; Xing, Z.; Gu, S.; Guo, X.; Hu, Y.; Song, Z.; Zhang, Q.; Long, X.; and Yin, W. 2024a. Hedrive: Human-like end-to-end driving with vision language models. *arXiv preprint arXiv:2410.05051*.

Wang, P.; Fan, Z.; Wang, Z.; Su, H.; Ramamoorthi, R.; et al. 2024b. Lift3d: Zero-shot lifting of any 2d vision model to 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21367–21377.

Xie, A.; Rybkin, O.; Sadigh, D.; and Finn, C. 2025. Latent diffusion planning for imitation learning. *arXiv preprint arXiv:2504.16925*.

Ye, S.; Jang, J.; Jeon, B.; Joo, S.; Yang, J.; Peng, B.; Mandlekar, A.; Tan, R.; Chao, Y.-W.; Lin, B. Y.; et al. 2024. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid Loss for Language Image Pre-Training. arXiv:2303.15343.

Zhang, Z.; Duan, T.; Sun, Z.; Guan, X.; Wang, J.; Liang, H.; Cui, Y.; and Cui, H. 2024. Prediction-based Hierarchical Reinforcement Learning for Robot Soccer. In *2024 IEEE/CIC International Conference on Communications in China (ICCC)*, 1721–1726.

Zhen, H.; Sun, Q.; Zhang, H.; Li, J.; Zhou, S.; Du, Y.; and Gan, C. 2025. TesserAct: learning 4D embodied world models. *arXiv preprint arXiv:2504.20995*.