Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

Mimic Intent, Not Just Trajectories

Renming Huang[1,2], Chendong Zeng[1,2], Wenjing Tang[1], Jintian Cai[1,2], Cewu Lu[1,2], Panpan Cai[1,2,†]
[1]Shanghai Jiao Tong University [2]Shanghai Innovation Institute
[†]Corresponding author
https://renming-huang.github.io/MINT

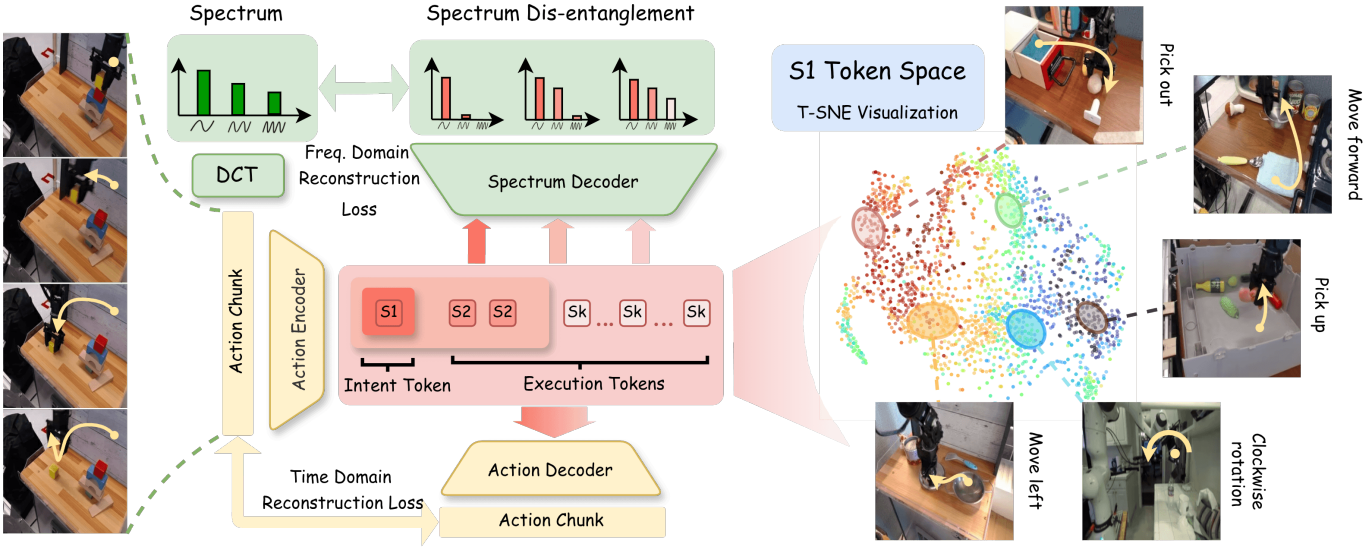

Fig. 1: **Left:** We propose Spectrally Disentangled Action Tokenizer, which encodes action chunks into multi-scale tokens via scale-wise frequency domain reconstruction constraints, where the coarsest scale captures global intent and finer scales encode execution residuals. **Right:** The T-SNE visualization of the S_1 token space demonstrates that the learned S_1 tokens form distinct clusters corresponding to semantically consistent behaviors (e.g., "Pick up","Move forward" and "Clockwise Rotation")

***Abstract*—While imitation learning (IL) has achieved impressive success in dexterous manipulation through generative modeling and pretraining, state-of-the-art approaches like Vision-Language-Action (VLA) models still struggle with adaptation to environmental changes and skill transfer. We argue this stems from mimicking raw trajectories without understanding the underlying intent. To address this, we propose explicitly disentangling behavior intent from execution details in end-2-end IL: *"Mimic Intent, Not just Trajectories" (MINT)*. We achieve this via *multi-scale frequency-space tokenization*, which enforces a spectral decomposition of action chunk representation. We learn action tokens with a multi-scale coarse-to-fine structure, and force the coarsest token to capture low-frequency global structure and finer tokens to encode high-frequency details. This yields an abstract *Intent token* that facilitates planning and transfer, and multi-scale *Execution tokens* that enable precise adaptation to environmental dynamics. Building on this hierarchy, our policy generates trajectories through *next-scale autoregression*, performing progressive *intent-to-execution reasoning*, thus boosting learning efficiency and generalization. Crucially, this disentanglement enables *one-shot transfer* of skills, by simply injecting the Intent token from a demonstration into the autoregressive generation process. Experiments on several manipulation benchmarks and on a real robot demonstrate state-of-the-art success rates, superior inference efficiency, robust generalization against disturbances, and effective one-shot transfer.**

I. Introduction

Imitation learning from demonstrations has become a dominant paradigm for learning robot manipulation policies. Recent advances are largely driven by vision–language–action (VLA) models [4, 5, 21], which map visual observations and language instructions directly to continuous control commands, achieving impressive performance on dexterous tasks such as folding laundry, pouring coffee, and object rearrangement. However, despite their success in closed settings, these models often generalize poorly to environmental variations and new task instances [15]. We argue that a key limitation is that most existing approaches learn to mimic trajectories as raw signals, without modeling why a particular sequence of actions is executed. As a result, learned policies tend to overfit to surface-level correlations in demonstrations, rather than capturing the underlying behavioral intent that governs task execution.

To address this limitation, recent work has explored action tokenization, which maps continuous trajectories into discrete latent representations [9, 46, 49] . Discrete tokens align with the intuition that action semantics are structured and compositional, and token-based policies predict abstract action sequences

before decoding them into executable trajectories. However, existing tokenization methods largely function as compression mechanisms [37, 46] rather than semantic abstractions. Their learning objectives are typically agnostic to action meaning, providing no explicit constraint that aligns the token space with interpretable behavioral concepts such as intent. Even when multi-scale or hierarchical tokenization is adopted [17], the semantics of coarse representations remain unconstrained.

To fill the gap, we introduce MINT — Mimic Intent, Not just Trajectories, an imitation learning framework based on multi-scale frequency-space action tokenization. MINT explicitly disentangles behavioral intent from execution details through *spectral decomposition*. The key insight is that a trajectory can be viewed as a superposition of signals at different frequencies: low-frequency components characterize the global shape and long-horizon structure of the behavior, while high-frequency components encode fine-grained execution details and reactive adjustments.

Concretely, we transform action chunks from the time domain into the frequency domain using the Discrete Cosine Transform (DCT) [2]. We train a multi-scale variational autoencoder (VAE) [42] with a frequency-domain reconstruction objective, enforcing consistency between the spectral representations of the original and reconstructed trajectories. The latent space is organized into multiple token scales $(S_1, S_2, \ldots, S_k)$, with progressively increasing capacity. The coarsest scale contains a single token, while finer scales introduce additional tokens to capture residual information.

To enforce disentanglement, we design a progressive reconstruction scheme: the model is trained to reconstruct the frequency-domain trajectory using (i) $S1$ alone, (ii) $S_1 + S_2$, (iii) $S_1 + S_2 + S_3, ...$, etc.. This structure induces a clear learning behavior—different levels of abstraction naturally attend to different regions of the frequency spectrum: the S_1 token is forced to capture the dominant, low-frequency components to minimize reconstruction error, while finer tokens specialize in modeling high-frequency residuals. This spectral separation induces a principled disentanglement between intent and execution, rather than relying on heuristic or post-hoc interpretation of latent variables. We therefore interpret S_1 as an "Intent token", and $S_2 \sim S_K$ as "Execution tokens".

This representation enables several key benefits. First, progressive prediction of $S_1 \sim S_K$ naturally induces an intent-to-execution reasoning process in latent space, improving sample efficiency and stabilizing long-horizon generation. Second, the Intent token provides a more compact, reusable task specification than language instructions. Given a single demonstration of a novel task, we can extract its Intent token and inject it into the policy's autoregressive generation process, enabling one-shot skill transfer to new layouts, new tasks, and extended horizons.

We evaluate MINT on four manipulation benchmarks, LIBERO [31], MetaWorld [50], CALVIN [33], and the more challenging LIBERO-Plus [15], as well as on a real robotic system. MINT achieves state-of-the-art performance on standard benchmarks, outperforming strong pretrained VLA models ($\pi_{0.5}$ [6]), action-tokenization-based methods (UniVLA [9]), and classic imitation learning approaches (ACT [53], Diffusion Policy [12]). When trained on LIBERO and evaluated on LIBERO-Plus under stronger disturbances, MINT demonstrates substantially improved robustness, achieving 15% higher success rates than the strongest baseline, OpenVLA-OFT [22]. Leveraging intent-level representations, MINT further enables one-shot skill transfer, achieving 60% higher transfer performance on novel tasks and environments from a single demonstration. Real-robot experiments confirm that MINT transfers effectively to physical systems, requiring only around 20 demonstrations per task while outperforming the strongest baseline ($\pi_{0.5}$) by 29%.

II. Related Work

A. Vision Language Action Models

The integration of Large Language Models (LLMs) [1, 43] and Vision-Language Models (VLMs) [20, 26] has evolved the prevailing Behavior Cloning paradigm [12, 18, 19, 25, 51, 53], into powerful Vision-Language-Action (VLA) models [4, 5, 6, 14, 16, 21, 41, 55]. However, despite leveraging internet scale pre-training, current VLAs have yet to exhibit the emergent generalization and learning efficiency characteristic of their LLM and VLM counterparts. We argue that this disparity stems from the fundamental limitation of mimicking raw trajectories without explicitly comprehending the underlying intent. Consequently, a framework that can disentangle high-level intent from low-level motion details, while ensuring the learned representations physically executable, is highly desired.

B. Action Tokenization

Action tokenization [8, 9, 11, 30, 39, 49] has emerged as a promising avenue for structuring continuous motor control. Mathematical approaches, including direct binning [7, 53] as well as FAST and BEAST [37, 54], discretize actions in a structured way and guarantee reconstruction, but they do not enforce explicit constraints to capture behavioral intent. Learning based methods, such as VQ-VAE variants [24, 34, 46], learn tokens automatically and achieve strong compression, but without internal constraints the learned tokens often preserve low-level kinematics rather than intent. We address this by constraining tokenization to disentangle intent from execution while keeping actions executable, producing tokens suitable for intent-to-action reasoning.

C. Coarse-to-Fine Tokenization

Standard Residual VQ methods in VLA [32, 46] employ a flat hierarchy with uniform capacity across scales, failing to accommodate the inherent asymmetry between sparse, abstract intent and dense, high-frequency execution details. The potential of Multi-Scale VQ is demonstrated by VAR [42] in image generation. CARP [17] mimic this architecture for robotic action chunks. By relying exclusively on time-domain reconstruction over aggregated multi-scale tokens, this design lacks explicit scale wise supervision, leading the hierarchy to prioritize local fidelity over the intent-to-execution structure

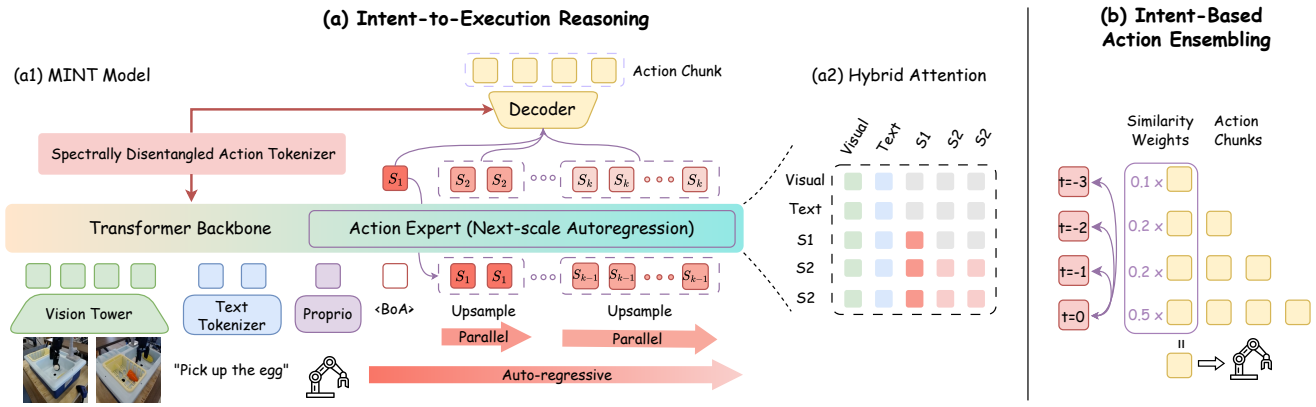

Fig. 2: **MINT Policy Overview.** (a) MINT autoregressively predicts action tokens across K temporal scales—moving from a global intent token to high-frequency execution tokens—which are subsequently mapped to continuous trajectories via the decoder. (b) Intent-based action ensemble ensures temporal consistency and smooth behavioral transitions, enhancing stability in long-horizon tasks.

essential for manipulation. In contrast, we diverge by imposing scale-wise reconstruction constraints explicitly in the frequency domain. This spectral decomposition forces the coarsest scale to exclusively capture global, low-frequency dynamics, ensuring a structural disentanglement of high-level intent from low-level execution details. The intent token opens up two possibilities: intent-based ensembling, and more crucially, task specification using the intent token, thus one-shot transfer.

III. Overview

MINT is a two-stage imitation learning framework that explicitly disentangles behavioral intent from execution details. It consists of (1) a Spectrally Disentangled Action Tokenizer (SDAT) that learns structured discrete representations from demonstration trajectories (Fig. 1), and (2) MINT policy that generates actions through progressive intent-to-execution reasoning in the learned token space (Fig. 2). The SDAT tokenizer provides a shared action codebook and a decoder, while the MINT policy learns to predict action tokens in a coarse-to-fine manner and decode them into executable trajectories. Training of MINT contains two phases:

In the first phase (Section IV), we train SDAT on demonstration trajectories to obtain multi-scale action representations. Each trajectory is segmented into overlapping action chunks using a sliding window, and each chunk is transformed from the time domain to the frequency domain using the DCT. SDAT adopts a VQ-VAE [44] architecture to learn a discrete action codebook and a quantizer that maps action chunks to tokens.

To induce disentanglement, SDAT decomposes actions into K temporal scales with progressively increasing capacity (Fig. 1 Left). The coarsest scale (the *Intent token*) contains a single token intended to capture global, low-frequency structure, while finer scales (the *Execution tokens*) introduce additional tokens that model residual information not explained by coarser ones. All scales share a single codebook. Crucially, the SDAT tokenizer is trained using progressive reconstruction in the frequency domain: the model is required to reconstruct the frequency-domain trajectory using only the coarsest representation first, then using progressively finer representations, up to all K scales. This design constrains the functionality of tokens at different scales, forcing coarse tokens to explain dominant low-frequency components (Fig. 1 Right) and finer tokens to specialize in high-frequency residuals. Finally, a full reconstruction in the time domain using the union of all scales is applied as an auxiliary objective to ensure faithful recovery of execution details.

In the second phase, we train the MINT policy that predicts and executes action tokens produced by SDAT (Section V). The policy takes as input the current visual observation, language instruction, and robot proprioceptive state, and outputs an action trajectory. It consists of a vision-language backbone and an action expert. The backbone encodes visual and language inputs using either a standard transformer or a pretrained vision–language model. Conditioned on these features and the robot state, the action expert autoregressively predicts action tokens from coarse to fine scales, generating all tokens within a scale in parallel while maintaining autoregression across scales (Fig. 2 (a)). The predicted tokens are then decoded into continuous trajectories using the decoder inherited from SDAT.

We train two variants of MINT: a language-conditioned version and a language-free version (MINT-Zero). The former is used to evaluate task performance and robustness on standard manipulation benchmarks, while the latter is designed for one-shot skill transfer. In the transfer setting, an intent token is extracted from a single demonstration and injected into the policy by fixing the coarsest-scale token, while the policy generates execution tokens conditioned on it. We further consider two model scales: a 30M-parameter model adapted from a standard transformer architecture trained from scratch (MINT-30M), and a 4B-parameter model that combines a pretrained vision–language backbone with a randomly initialized action head (MINT-4B). Both variants are trained end-to-end in their respective settings.

IV. Spectrally Disentangled Action Tokenizer

We propose the **S**pectrally **D**isentangled **A**ction **T**okenizer (SDAT), a multi-scale framework that explicitly disentangles behavioral intent from low-level execution details. SDAT introduces a spectral decoder together with a scale-wise spectral reconstruction objective that supervises the frequency composition of actions at different scales, as shown in Algorithm 1.

A. Action Encoder and Spectrum Decoder

Let $\mathbf{A} \in \mathbb{R}^{H\times D}$ denote a continuous action sequence of horizon H with action dimension D. An action encoder $\mathcal{E}$ maps the input sequence into a compressed latent embedding: $f = \mathcal{E}(\mathbf{A}), \quad f \in \mathbb{R}^{L\times C}$, where L denotes the compressed temporal length and C is the latent feature dimension.

Given the latent embedding f, a spectrum decoder $\mathcal{D}_{\text{spec}}$ reconstructs the action sequence $\hat{\mathbf{A}} \in \mathbb{R}^{H\times D}$ via an action decoder $\mathcal{D}$ and converts it into frequency-domain representation via the DCT applied along the temporal dimension. For each action dimension $d \in \{1,\dots,D\}$, the DCT coefficients are computed as:

$$\mathbf{F}_{k,d} = \sum_{h=0}^{H-1} \hat{\mathbf{A}}_{h,d} \cos\Big[\frac{\pi}{H}\left(h+\tfrac{1}{2}\right)k\Big], k = 0,\dots,H-1, \quad (1)$$

where $\mathbf{F} \in \mathbb{R}^{H\times D}$ denotes the resulting frequency-domain representation.

B. Multi-Scale Residual Quantization

SDAT utilizes a Multi-Scale Residual Quantization scheme [23, 42] to decompose the continuous latent embedding $f^{(0)}$ into a multi-scale discrete representation $\mathbf{S} = \{\mathbf{s}_1,\dots,\mathbf{s}_K\}$, where each $\mathbf{s}_k \in \{1,\dots,V\}^{l_k}$ is a discrete token map at resolution l_k, representing the quantized features at scale k. Let $\mathcal{Z} \in \mathbb{R}^{V\times C}$ denote a shared codebook containing V code vectors, and let $\{l_1,\dots,l_K\}$ be a set of increasing resolutions with $l_K = L$. Quantization is performed recursively on residual features. Let $f^{(k)}$ denote the residual feature at scale k. At each scale, the feature is first interpolated to resolution l_k and quantized via $\mathcal{Q}$, producing $\mathbf{s}_k = \mathcal{Q}(\text{Interpolate}(f^{(k)}, l_k))$. The discrete indices are then mapped to embeddings $\mathbf{z}_k = \text{Lookup}(\mathcal{Z}, \mathbf{s}_k)$. The quantized embeddings $\mathbf{z}_k$ are projected back to the original latent resolution L through an interpolator and a scale-specific projector ϕ_k, and the residual feature is updated as: $f^{(k+1)} = f^{(k)} - \phi_k(\mathbf{z}_k)$. This forms a coarse-to-fine structure across multiple scales.

C. Scale-wise Spectral Reconstruction

To enforce spectral disentanglement across quantization scales, we supervise the contribution of each residual level in the frequency domain. Let $\hat{f}^{(k)}$ be the cumulative latent approximation up to scale k, formed by summing the quantized residuals:

$$\hat{f}^{(k)} = \sum_{i=1}^{k} \phi_i(\text{Lookup}(\mathcal{Z}, \mathbf{s}_i)), \quad (2)$$

Algorithm 1 Spectrally Disentangled Action Tokenizer

Inputs: Action sequence $\mathbf{A}$
Hyperparameters: scales K, resolutions $(l_k)_{k=1}^K$
Initialize: $f^{(0)} \leftarrow \mathcal{E}(\mathbf{A})$, $\hat{f}^{(0)} \leftarrow 0$, $\mathbf{S} \leftarrow [\,]$, $\mathcal{F} \leftarrow [\,]$
for $k = 1,\dots,K$ **do**
  $\mathbf{s}_k \leftarrow \mathcal{Q}(\text{Interpolate}(f, l_k))$
  $\mathbf{S} \leftarrow \mathbf{S} \cup \{\mathbf{s}_k\}$
  $\mathbf{z}_k \leftarrow \text{Lookup}(\mathcal{Z}, \mathbf{s}_k)$
  $\mathbf{z}_k \leftarrow \text{Interpolate}(\mathbf{z}_k, l_K)$
  $f^{(k)} \leftarrow f^{(k-1)} - \phi_k(\mathbf{z}_k)$
  $\hat{f}^{(k)} \leftarrow \hat{f}^{(k-1)} + \phi_k(\mathbf{z}_k)$
  $\mathbf{F}^{(k)} \leftarrow \mathcal{D}_{\text{spec}}(\hat{f}^{(k)})$
  $\mathcal{F} \leftarrow \mathcal{F} \cup \{\mathbf{F}^{(k)}\}$
end for
$\hat{A} \leftarrow \mathcal{D}(\hat{f}^{(K)})$
Return: Multi-scale tokens $\mathbf{S}$, frequency domain spectrum $\mathcal{F}$, reconstruction sequences $\hat{\mathcal{A}}$

Each cumulative feature $\hat{f}^{(k)}$ is decoded by the shared spectral decoder $\mathcal{D}_{\text{spec}}$ into a progressively refined action sequence $\hat{\mathbf{A}}^{(k)}$ which is then transformed into the frequency domain as $\mathbf{F}^{(k)} = \text{DCT}(\hat{\mathbf{A}}^{(k)})$. Let the $\mathbf{F} = \text{DCT}(\mathbf{A})$ denote the ground truth, a scale-wise spectral loss enforces consistency between the ground-truth actions and each partial reconstruction:

$$\mathcal{L}_{\text{freq.}} = \sum_{k=1}^{K} \lambda_k \left\|\mathbf{F} - \mathbf{F}^{(k)}\right\|_2, \quad (3)$$

This encourages early scales to capture low-frequency global structures, while later scales focus on high-frequency details.

D. Training Objective

The SDAT action tokenizer is trained to capture spectral structure across scales. Given an action sequence $\mathbf{A}$, the encoder $\mathcal{E}$ produces a latent f, which is discretized via multi-scale residual quantization into $\hat{f}$. The spectral decoder $\mathcal{D}_{\text{spec}}$ outputs both frequency domain spectrum $\hat{\mathbf{F}}$ and the reconstructed actions $\hat{\mathbf{A}}$.

The training loss includes the scale-wise spectral reconstruction $\mathcal{L}_{\text{freq.}}$, followed with codebook and commitment losses [44], and a auxiliary l_1 reconstruction term. Formally, it is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{freq.}} + \underbrace{\|\text{sg}(f) - \hat{f}\|_2^2}_{\text{Codebook loss}} + \underbrace{\|f - \text{sg}(\hat{f})\|_2^2}_{\text{Commitment loss}} + \alpha \underbrace{\|\mathbf{A} - \hat{\mathbf{A}}\|_1^2}_{\text{Auxiliary loss}},$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator, and α is a weighting factor.

V. MINT Policy Learning

The MINT policy learns *intent-to-execution reasoning* by operating on the multi-scale discrete action tokens produced by SDAT. Leveraging *next-scale autoregressive prediction* (Fig. 2 (a1)), the policy performs autoregressive prediction across

scales while decoding tokens in parallel within each scale using a hybrid attention mechanism (Fig. 2 (a2)).

A. Next-Scale Autoregressive Modeling

Building on the multi scale action token maps produced by the SDAT action tokenizer, denoted as $\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_K\}$, we model the joint distribution over tokens autoregressively across scales:

$$p(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K) = \prod_{k=1}^{K} p(\mathbf{s}_k \mid \mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{k-1}). \quad (4)$$

Each autoregressive unit $\mathbf{s}_k$ is treated as a *token map* rather than a token sequence, and the sequence of coarser-scale token maps $(\mathbf{s}_1, \ldots, \mathbf{s}_{k-1})$ serves as the prefix for predicting $\mathbf{s}_k$. At the k-th autoregressive step, following [42], all distributions over l_k tokens in $\mathbf{s}_k$ will be generated *in parallel*, conditioned on the prefix token maps $\mathbf{s}_{<k}$ and a scale-specific positional embedding map.

During training, we apply a *hybrid attention mask* to enforce a scale-aware dependency structure, such that the token map at scale k can attend only to token maps from coarser or equal scales $\mathbf{s}_{\leq k}$. The policy is optimized using the standard cross-entropy loss, which measures the discrepancy between the predicted token map $\hat{\mathbf{s}_k}$ and the ground-truth token map $\mathbf{s_k}$ derived from the action sequence.

B. Intent-Based Action Ensemble

Let $\mathbf{a}_t|\mathbf{o}_t$ denote the predicted action to be executed at time step t conditioned on an observation $\mathbf{o}_t$. The final action at time t is associated with a set of overlapping predictions $\{\mathbf{a}_t|\mathbf{o}_{t-H}, \ldots, \mathbf{a}_t|\mathbf{o}_{t-1}, \mathbf{a}_t|\mathbf{o}_t\}$. During inference, let $\mathbf{s}_1^{(t)} \in \mathbb{R}^C$ denote the intent token associated with the action chunk generated at time step t, and $\mathbf{s}_1^{(t-h)}$ denote the intent token of a previous chunk. We derive the final action executed at time t via intent-based action ensemble (Fig. 2 (b)):

$$\mathbf{a}_t = \sum_{h=0}^{H} w_h^{\text{intent}} \cdot \mathbf{a}_t|\mathbf{o}_{t-h}, \quad (5)$$

where w_h^{intent} is an adaptive weight determined by the similarity between behavioral intents. The ensemble weights are computed by measuring the similarity between the current intent token and historical intent tokens:

$$w_h^{\text{intent}} = \frac{\exp\big(\beta \, \langle \mathbf{s}_1^{(t)}, \mathbf{s}_1^{(t-h)} \rangle\big)}{\sum_{j=0}^{H} \exp\big(\beta \, \langle \mathbf{s}_1^{(t)}, \mathbf{s}_1^{(t-j)} \rangle\big)}, \quad (6)$$

where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity between two intent tokens, $\beta > 0$ is a temperature scaling the effect of intent similarity on weight assignment.

Intent based ensemble enabling smooth execution and rapid switching between behaviors. Empirical studies reveal it improves action stability and long-horizon task success.

C. Model Architectures

We instantiate our framework with two variant architectures:

a) MINT-30M: This variant is a lightweight, decoder-only Transformer model trained from scratch, with approximately 30M trainable parameters. Visual inputs are encoded by frozen SigLIP [52] and DINOv2 [35] backbones, while language instructions are processed by a frozen BERT [13] encoder, and injected into the network via Feature-wise Linear Modulation (FiLM) [36] layers.

b) MINT-4B: This is a large-scale policy model built upon an existing vision–language architecture used in π_0 and $\pi_{0.5}$. It employs a PaliGemma-2.6B [3] vision language model together with a SigLIP based visual encoder, both pretrained on large-scale robotic datasets. MINT-4B uses a transformer-based action expert with approximately 300M parameters, which is trained from scratch. The action expert performs next-scale autoregressive prediction over multi-scale action tokens.

VI. Experiments

We evaluate our framework on standard benchmarks and the LIBERO-Plus suite to show that explicitly disentangling behavioral intent from execution improves performance and generalization under severe disturbances. We also perform experiments in real-world environments, and we test the one-shot transfer capability of the framework in simulation.

A. Performance Comparison

Benchmark. We conduct experiments on three widely adopted robotic manipulation benchmarks: LIBERO, CALVIN, and MetaWorld , which together cover multi-task manipulation, long-horizon compositional reasoning, and task generalization across varying difficulty levels. LIBERO is a simulated benchmark suite composed of five task families: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, LIBERO-Long and LIBERO-90. CALVIN features 34 tabletop manipulation tasks across four scene configurations. we evaluate on CALVIN ABCD→D benchmark, require policies to follow free-form language instructions and complete 5 tasks in sequence across 500 different instruction chains. MetaWorld is a large-scale manipulation benchmark consisting of 50 tasks with varying levels of difficulty. Tasks are categorized into Easy, Medium, Hard, and Very Hard groups, reflecting increasing requirements on precision, coordination, and long-horizon control.

Baselines. We compare against a comprehensive set of baselines across three benchmarks. Specifically, on the LIBERO suite, we compare our method against Diffusion Policy, WorldVLA [10], SmolVLA [40], both train from scratch, OpenVLA [21], OpenVLA-OFT [22], LAPA [49], UniVLA [9], and the π family (π_0 [5], π_0-FAST [37], and $\pi_{0.5}$), both pretrain with large robot dataset. For the CALVIN benchmark, we evaluate against RT-1 [7], RoboVLMs [29], UniVLA, and $\pi_{0.5}$. On Meta-World, we compare against Diffusion Policy, TinyVLA [48], and π_0.

Results. The results for these experiments are summarized in Table IX. MINT consistently matches or surpasses all current state-of-the-art approaches across all reported benchmarks. Notably, our MINT-30M variant surpasses OpenVLA and π_0 on LIBERO by wide margins despite its significantly smaller size

TABLE I: Performance comparison across LIBERO, CALVIN, and MetaWorld benchmarks

LIBERO						
Method	SPATIAL	OBJECT	GOAL	LONG	**Avg.**	L90
Without Pre-training						
Diffusion Policy [12]	78.3	92.5	68.3	50.5	72.4	–
MDT [38]	78.5	87.5	73.5	64.8	76.1	–
WorldVLA [10]	87.6	96.2	83.4	60.0	81.8	–
SmolVLA [40]	93.0	94.0	91.0	77.0	88.8	–
MINT-30M	**98.6**	**99.2**	**97.4**	**93.2**	**97.1**	**97.4**
With Pre-training						
LAPA [49]	73.8	74.6	58.8	55.4	65.7	–
OpenVLA [21]	84.7	88.4	79.2	53.7	76.5	–
π_0-FAST [37]	96.4	96.8	88.6	60.2	85.5	–
π_0 [5]	90.0	86.0	95.0	73.0	86.0	–
UniVLA [9]	96.5	96.8	95.6	92.0	95.2	–
OpenVLA-OFT [22]	96.9	98.1	95.6	91.1	95.4	–
$\pi_{0.5}$ [6]	**98.8**	98.2	98.0	92.4	96.9	96.0
MINT-4B	97.4	**99.6**	**98.2**	**97.8**	**98.3**	**98.7**

CALVIN (ABCD→D)						
Method	Success @ k Tasks					**Avg.**
	1	2	3	4	5	Len
RT-1 [7]	84.4	61.7	43.8	32.3	22.7	2.45
Robo-Flamingo [28]	96.4	89.6	82.4	74.0	66.0	4.09
$\pi_{0.5}$ [6]	94.2	89.3	82.7	78.5	70.3	4.15
UnifiedVLA [47]	**97.9**	**94.8**	89.2	82.8	75.1	4.34
RoboVLMs [29]	96.7	93.0	89.9	86.5	82.6	4.49
MINT-4B	97.4	94.2	**91.7**	**88.2**	**86.1**	**4.57**

MetaWorld						
Method	Easy	Medium	Hard	Very Hard	**Avg.**	–
Diffusion Policy [12]	23.1	10.7	1.9	6.1	10.5	–
TinyVLA [48]	77.6	21.5	11.4	15.8	31.6	–
π_0 [5]	77.9	51.8	53.3	20.0	50.8	–
MINT-4B	**82.1**	**72.4**	**58.3**	**56.0**	**67.2**	–

and training from scratch. Our pre-trained MINT-4B variant also outperforms $\pi_{0.5}$ on LIBERO and the π_0 baseline on Meta-World, where it nearly triples the success rate in the most challenging "Very Hard" tasks. On CALVIN, MINT-4B demonstrate superior stability in long-horizon composition. These results demonstrate the versatility of MINT to adapt to diverse task complexities and manipulation settings, showing that MINT not only provides strong performance across scales but also demonstrates robust generalization to difficult, high-precision environments.

B. Generalization

Benchmark and Baselines. We evaluate the robustness of our model against distribution shifts on the LIBERO-PLUS benchmark. This suite assesses seven distinct dimensions of generalization. These dimensions include camera viewpoints involving variations in pose and field of view, robot initial states comprising manipulator pose variations, and language instructions for instruction following tasks. The benchmark also covers light conditions such as intensity and color shifts,

TABLE II: Generalization comparison on LIBERO-PLUS

Method	Camera	Robot	Lang.	Light	Back.	Noise	Layout	**Avg.**
OpenVLA	0.8	3.5	23.0	8.1	34.8	15.2	28.5	16.3
UniVLA	1.8	46.2	69.9	69.0	81.0	21.2	31.9	45.9
π_0	13.8	6.0	58.8	85.0	81.4	79.0	68.9	56.1
π_0-FAST	65.1	21.6	61.0	73.2	73.2	74.4	68.8	62.5
OpenVLA-OFT	56.4	31.9	79.5	88.7	**93.3**	75.8	74.2	71.4
$\pi_{0.5}$	53.0	**50.3**	65.7	83.1	77.3	53.2	72.7	65.0
MINT-30M	61.4	41.2	61.6	92.2	77.1	76.5	76.2	69.5
MINT-4B	**72.2**	42.4	**85.8**	**96.6**	88.9	**90.1**	**84.6**	**80.1**
Trained with LIBERO Plus								
OpenVLA-OFT+	92.8	30.3	**85.8**	94.9	93.9	89.3	77.6	80.7
$\pi_{0.5}$+	67.2	42.4	59.4	75.8	74.9	72.6	64.5	65.3
MINT-4B+	**95.6**	**44.6**	84.7	**95.1**	**94.5**	**95.2**	**78.7**	**84.1**

background textures reflecting scene appearance changes, sensor noise involving photometric degradation, and object layout encompassing displacement and confounding objects. We compare our models against a wide range of baselines including OpenVLA, UniVLA, π_0, π_0-FAST, and $\pi_{0.5}$. We compare MINT+ against $\pi_{0.5}$+, both finetuned on the LIBERO-PLUS dataset.

Result and Analyze. Table X demonstrates that MINT exhibits strong and consistent generalization on the LIBERO-PLUS benchmark. Across both model scales, MINT-30M and MINT-4B outperform prior baselines under camera viewpoint and robot initialization shifts, with especially pronounced gains for camera variations. MINT demonstrates stable performance under scene-level variations and sensor perturbations, while preserving reliable instruction following, indicating effective alignment between intent and low-level action execution. Fine-tuning on the LIBERO-PLUS dataset further amplifies these advantages. Compared with $\pi_{0.5}$+, MINT-4B+ leverages the same distributional diversity more effectively, achieving broader and more uniform improvements across perturbation types. This highlights the strength of the MINT framework in generalizing to complex and heterogeneous conditions, rather than relying solely on increased data diversity.

C. One-Shot Transfer via Intent Token Injection

We evaluate one-shot transfer on out-of-distribution tasks by comparing MINT-30M against MINT-Zero-30M. While MINT-30M is one-shot finetuned on a single demonstration for each evaluation task, MINT-Zero-30M conditions on an explicit intent token extracted from a single demonstration via the action tokenizer. The evaluation focuses on three types of distributional shifts as shown in Fig. 3: (i) *New Task*, introducing entirely unseen task semantics; (ii) *New Layout*, where the task is familiar but the layout is novel; (iii) *Extended Horizon*, requiring the execution of longer, sequential actions than observed during training.

We construct a base dataset from LIBERO-90 by excluding the "*LIVING ROOM SCENE 1-3*" subset, ensuring that evaluation tasks with novel layouts and unseen semantics remain strictly out-of-distribution. Both MINT-30M and MINT-

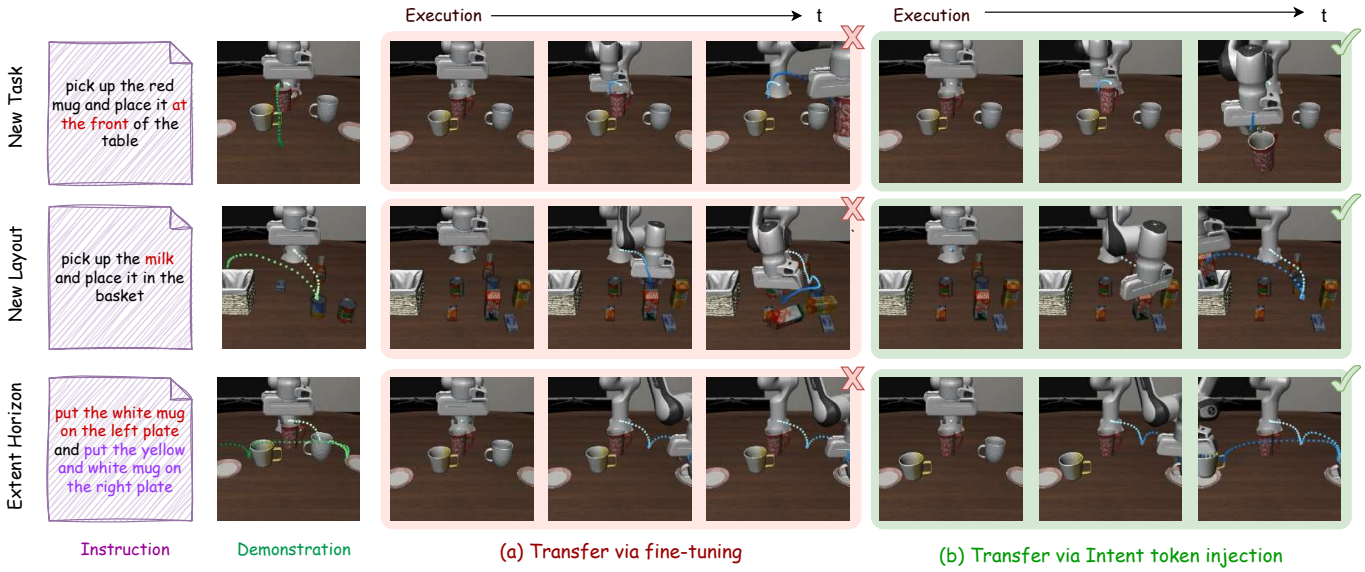

Fig. 3: **One-shot transfer evaluation on OOD tasks in simulation**. We evaluate generalization across three compositional shifts: New Layout, New Task, and Extended Horizon.

TABLE III: One-shot transfer performance comparison.

Method	Task Specification	New Task	New Layout	Extend Horizon	Avg.
Replay	Replay	0.28	0.12	0.04	0.11
Fine-tune (MINT-30M)	Language	0.42	0.08	0.00	0.17
Intent-injection (MINT-Zero-30M)	Intent	**0.90**	**0.68**	**0.72**	**0.77**

Zero-30M are trained on this base dataset; for MINT-30M, each evaluation task is further one-shot finetuned using a single demonstration. During inference, MINT-30M uses language-based task specifications, while MINT-Zero-30M is conditioned on the $\mathbf{s}_1$ token extracted from a single demonstration and autoregressively predicts the remaining action tokens via next-scale prediction.

Results. As shown in Table III, MINT-30M achieves limited one-shot transfer. For new layouts, one-shot transfer can cause the model to diverge, while for extended-horizon tasks, one-shot finetuning fails to capture the required behaviors. In contrast, intent-based task specification enables effective one-shot transfer, yielding high success rates across new tasks, new layouts, and extended-horizon sequences. These results highlight that explicit intent representation provides a more grounded and execution-aligned task specification than language, allowing policies to efficiently transfer to new settings to novel compositions without additional training.

D. Real-World Experiments

We evaluated MINT-4B on four real-world tasks: seen behaviors *(A) Place Banana*, *(B) Stack Blocks*, and *(C) Insert Marker*, alongside an unseen *(D) Stack Cups* task for zero-shot generalization (Fig. 4). MINT-4B significantly outperforms all baselines, successfully managing the high-precision coaxial alignment and geometric re-orientation required for structural stability and proper fit (Fig. 5).

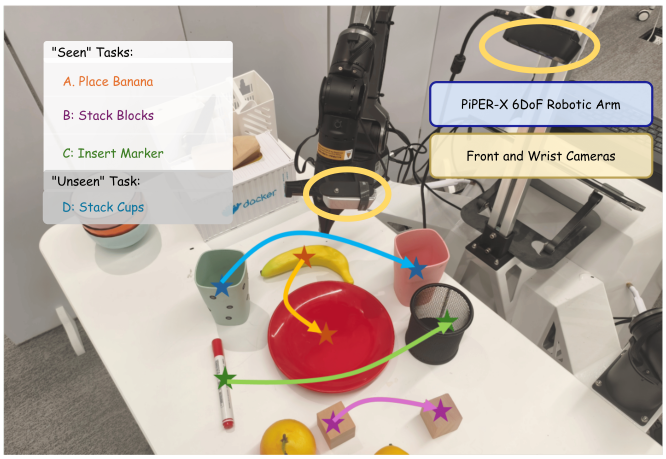

Fig. 4: **Real-world Experiment Setup.**

Training Data. The training data consists of a small task-specific dataset and a large-scale prior dataset. We use BridgeDataV2 [45], which contains over 60k manipulation trajectories across 24 environments and 13 skills, offering diverse objects, viewpoints, and workspace layouts. This dataset is used to pre-train the tokenizer and action head. For target tasks, we collect 20 demonstrations per task, totaling 2.4k frames, which are used for in-domain post-training in a multi-task learning setting.

Baselines and Setup. We evaluated our MINT-4B model against three baselines: a task-specific policy (ACT [53]), a generalist policy with pretrained weights (π_0 [5]), and a modified version of the $\pi_{0.5}$ [6] policy with a re-initialized action-expert head ($\pi^*_{0.5}$). Both MINT-4B and $\pi^*_{0.5}$ utilize the same pretrained VLM backbone and were pretrained on BridgeDataV2 before being finetuned on our collected demonstrations. In contrast, ACT was trained individually for

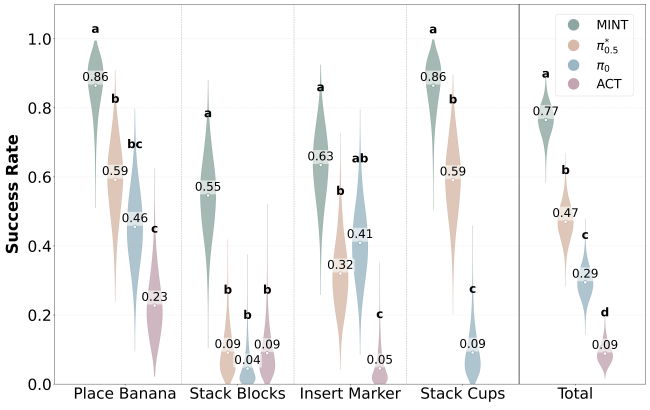

Fig. 5: **Real-world task results.** The violin plots show Bayesian posterior success rates. The distinct lettering indicates statistically distinguishable policies.

each task, and π_0 was directly finetuned from its pretrained weights. All policies were deployed on a 6-DOF Piper-X robotic arm with dual-camera RGB input, and each policy's performance was evaluated over 20 trials per task to ensure statistical significance.

Results. Using Bayesian posterior analysis, we find MINT statistically distinguishable from all baselines on seen tasks *(A)* and *(B)*. Notably, MINT significantly outperforms the runner-up $\pi^*_{0.5}$ on *(B) Stack Blocks*, demonstrating superior capability in high-precision axis alignment. Similar performance gaps are observed in the unseen task *(D) Stack Cups*. Despite novel objects, MINT effectively generalizes the shared "stacking" intent from task *(B)*, substantially outperforming baselines that overfit to specific object instances.

E. Ablation Studies

To isolate the contributions of the spectral disentanglement objective and the intent-based action ensemble, we conduct ablation studies across CALVIN and LIBERO-LONG baselines.

Efficacy of Scale-Wise Spectral Decomposition. Fig. 6 qualitatively demonstrates that our spectral objective organizes the latent space into coherent behavioral clusters, overcoming the fragmentation observed in standard time-domain reconstruction. Quantitatively (Table IV), while scale-wise time-domain constraints degrade performance (82.8%) by overfitting to high-frequency noise, our *Scale-Wise Spectral Loss* yields significant gains (93.4% on LIBERO-Long, 4.54 length on CALVIN). This confirms that enforcing spectral hierarchy is essential for disentangling global intent from execution details.

Impact of Intent-based Action Ensemble. Table IV shows that our *Intent-Based Action Ensemble* consistently outperforms both Temporal [53] (89.2%) and Action-based [27] (90.4%) baselines. By dynamically modulating aggregation weights based on intent compatibility, our method effectively resolves conflicts during behavioral transitions, achieving the highest success rate (93.2%) on LIBERO-Long and average sequence length (4.57) on CALVIN.

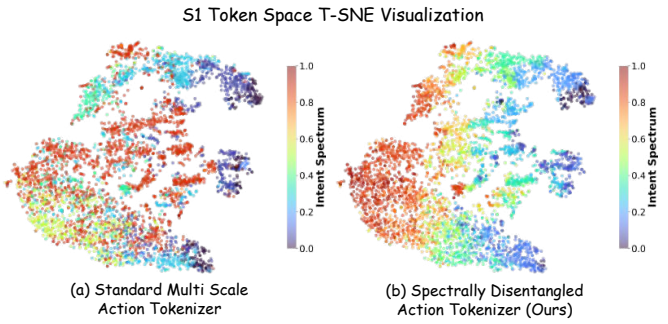

Fig. 6: **Visualization of the Intent Latent Space.** t-SNE of action chunks colored by s_1 tokens (RGB from top-3 PCs). (a) Standard time-domain reconstruction is fragmented. (b) Our SDAT produces coherent chromatic clusters aligned with action sequences structures.

TABLE IV: Ablation of training objectives and ensembling.

Ablation Setting	CALVIN	LIBERO-Long
Reconstruction Objectives		
Terminal Time-Domain Loss	4.36	87.8
+ Terminal Spectral Loss	4.41	88.2
+ Scale-Wise Time-Domain Loss	4.06	82.8
+ Scale-Wise Spectral Loss (Ours)	**4.54**	**93.4**
Action Ensemble		
No Ensemble	4.09	85.8
Temporal-based Ensemble [53]	4.32	89.2
Action-based Ensemble [27]	4.10	90.4
Intent-based Ensemble (Ours)	**4.57**	**93.2**

VII. Conclusion

We present MINT (*Mimic Intent, Not just Trajectories*), an imitation learning framework that explicitly decouples behavioral intent from low-level execution to address the generalization limitations of current VLA models caused by the entanglement of high-level planning and control dynamics. MINT leverages the Spectrally Disentangled Action Tokenizer (SDAT) to separate low-frequency global intent from high-frequency execution residuals via scale-wise spectral reconstruction, ensuring stable intent representation, improving robustness to environmental variations, achieving state-of-the-art performance across diverse manipulation benchmarks, and enabling effective one-shot skill transfer through intent injection.

Limitation and Future Work. MINT relies on trajectory demonstrations to learn intent, which limits the diversity of intents to the scope of available datasets. Exploring large-scale network data could provide a richer set of behaviors and broader coverage of tasks, while recombining discrete intent tokens offers a promising avenue for synthesizing novel, long-horizon behaviors zero-shot. Together, these directions could further enhance the generalization and flexibility of intent-driven control.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China (Grant No. 2024YFB4707600) and NSFC under grant No. 62303304.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 2006.

[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.

[4] 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.

[5] 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.

[6] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y Galliker, et al. π_0.5 : a vision-language-action model with open-world generalization. In *9th Annual Conference on Robot Learning*, 2025.

[7] 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.

[8] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[9] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

[10] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

[11] Yi Chen, Yuying Ge, Weiliang Tang, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19752–19763, 2025.

[12] 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*, 44(10-11):1684–1704, 2025.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[14] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. 2023.

[15] Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

[16] Yuxia Fu, Zhizhen Zhang, Yuqi Zhang, Zijian Wang, Zi Huang, and Yadan Luo. Mergevla: Cross-skill model merging toward a generalist vision-language-action agent. *arXiv preprint arXiv:2511.18810*, 2025.

[17] Zhefei Gong, Pengxiang Ding, Shangke Lyu, Siteng Huang, Mingyang Sun, Wei Zhao, Zhaoxin Fan, and Donglin Wang. Carp: Visuomotor policy learning via coarse-to-fine autoregressive prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13460–13470, 2025.

[18] RenMing Huang, Shaochong Liu, Yunqiang Pei, Peng Wang, Guoqing Wang, Yang Yang, and Hengtao Shen. Goal-reaching policy learning from non-expert observations via effective subgoal guidance. *arXiv preprint arXiv:2409.03996*, 2024.

[19] Renming Huang, Yunqiang Pei, Guoqing Wang, Yangming Zhang, Yang Yang, Peng Wang, and Hengtao Shen. Diffusion models as optimizers for efficient planning in offline rl. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024.

[20] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[21] 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.

[22] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed

and success. *arXiv preprint arXiv:2502.19645*, 2025.

[23] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11523–11532, 2022.

[24] Seungjae Lee, Yibin Wang, Haritheja Etukuru, H Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Behavior generation with latent actions. *arXiv preprint arXiv:2403.03181*, 2024.

[25] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[27] 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*, 2024.

[28] 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. *arXiv preprint arXiv:2311.01378*, 2023.

[29] 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*, 2024.

[30] Zuolei Li, Xingyu Gao, Xiaofan Wang, and Jianlong Fu. Latbot: Distilling universal latent actions for vision-language-action models. *arXiv preprint arXiv:2511.23034*, 2025.

[31] 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.

[32] Yicheng Liu, Shiduo Zhang, Zibin Dong, Baijun Ye, Tianyuan Yuan, Xiaopeng Yu, Linqi Yin, Chenhao Lu, Junhao Shi, Luca Jiang-Tao Yu, et al. Faster: Toward efficient autoregressive vision language action modeling via neural action tokenization. *arXiv preprint arXiv:2512.04952*, 2025.

[33] 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, 2022.

[34] Atharva Mete, Haotian Xue, Albert Wilcox, Yongxin Chen, and Animesh Garg. Quest: Self-supervised skill abstractions for learning continuous control. *Advances in Neural Information Processing Systems*, 37:4062–4089, 2024.

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[36] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[37] 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.

[38] Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. *arXiv preprint arXiv:2407.05996*, 2024.

[39] Dominik Schmidt and Minqi Jiang. Learning to act without actions. *arXiv preprint arXiv:2312.10812*, 2023.

[40] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

[41] 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.

[42] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

[43] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[44] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[45] 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*, pages 1723–1736. PMLR, 2023.

[46] Yating Wang, Haoyi Zhu, Mingyu Liu, Jiange Yang, Hao-

Shu Fang, and Tong He. Vq-vla: Improving vision-language-action models via scaling vector-quantized action tokenizers. *arXiv preprint arXiv:2507.01016*, 2025.

[47] Yuqi Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. Unified vision-language-action model. *arXiv preprint arXiv:2506.19850*, 2025.

[48] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*, 2025.

[49] 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. *arXiv preprint arXiv:2410.11758*, 2024.

[50] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

[51] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024.

[52] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[53] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[54] Hongyi Zhou, Weiran Liao, Xi Huang, Yucheng Tang, Fabian Otto, Xiaogang Jia, Xinkai Jiang, Simon Hilber, Ge Li, Qian Wang, et al. Beast: Efficient tokenization of b-splines encoded action sequences for imitation learning. *arXiv preprint arXiv:2506.06072*, 2025.

[55] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.