

See Once, Then Act: Vision-Language-Action Model with Task Learning from One-Shot Video Demonstrations

Guangyan Chen^{1†}, Meiling Wang¹, Qi Shao¹, Zichen Zhou¹, Weixin Mao², Te Cui¹, Minzhao Zhu², Yinan Deng¹, Luojie Yang¹, Zhanqi Zhang², Yi Yang¹, Hua Chen², Yufeng Yue^{1*}

¹ Beijing Institute of Technology ² LimX Dynamics

Abstract—Developing robust and general-purpose manipulation policies represents a fundamental objective in robotics research. While Vision-Language-Action (VLA) models have demonstrated promising capabilities for end-to-end robot control, existing approaches still exhibit limited generalization to tasks beyond their training distributions. In contrast, humans possess remarkable proficiency in acquiring novel skills by simply observing others performing them once. Inspired by this capability, we propose ViVLA, a generalist robotic manipulation policy that achieves efficient task learning from a single expert demonstration video at test time. Our approach jointly processes an expert demonstration video alongside the robot’s visual observations to predict both the demonstrated action sequences and subsequent robot actions, effectively distilling fine-grained manipulation knowledge from expert behavior and transferring it seamlessly to the agent. To enhance the performance of ViVLA, we develop a scalable expert-agent pair data generation pipeline capable of synthesizing paired trajectories from easily accessible human videos, further augmented by curated pairs from publicly available datasets. This pipeline produces a total of 892,911 expert-agent samples for training ViVLA. Experimental results demonstrate that our ViVLA is able to acquire novel manipulation skills from only a single expert demonstration video at test time. Our approach achieves over 30% improvement on unseen LIBERO tasks and maintains above 35% gains with cross-embodiment videos. Real-world experiments demonstrate effective learning from human videos, yielding more than 38% improvement on unseen tasks.

Index Terms—One-Shot Visual Imitation Learning, Vision Language Action Models, Unseen Task Generalization, Cross-embodiment Transfer, Robot Policy Learning.

I. INTRODUCTION

A fundamental objective in robotics research is the development of versatile, general-purpose robotic systems capable of performing diverse tasks across multiple domains, approaching human-level adaptability and versatility. Inspired by the remarkable success of Large Language Models (LLMs) [1–4] and Vision Language Models (VLMs) [5–9], which are capable of generalizing across a wide range of tasks and domains, the robotics community is actively developing robotic foundation models with similar capabilities. These works [10–15] typically build upon pretrained vision-language models and perform subsequent training on large-scale robot datasets spanning diverse objects, environments, and skills, thereby endowing robotic systems with extensive manipulation knowledge and enhanced generalization capabilities. Despite these advances, generalizing

to tasks beyond their training distributions remains a significant challenge. In stark contrast, humans exhibit remarkably efficient learning through visual imitation, extracting task-relevant knowledge from expert demonstrations and reproducing similar behaviors to accomplish comparable objectives. Inspired by this desirable capability, a natural question arises: *Can robotic agents similarly learn novel manipulation tasks directly by observing a single expert demonstration video?*

To answer this question, we investigate the development of a generalist policy for robotics that enables the robot to learn novel tasks beyond its training distribution by observing a single expert demonstration video without additional training. We develop our model based on VLMs to leverage their extensive prior knowledge and video comprehension capabilities. Upon this foundation, we train the model to predict subsequent robotic actions conditioned on a single expert demonstration video, endowing the model with the capacity to learn novel tasks from a single expert demonstration video at test time. While this approach is conceptually promising and benefits from powerful VLMs, achieving such capabilities remains non-trivial due to the following reasons:

(I) **Model capability: Lack of fine-grained action recognition capability.** A critical capability for enabling such generalist robotic policies is the acquisition of fine-grained manipulation knowledge from expert videos. However, existing VLMs predominantly focus on semantic-level video comprehension and exhibit limited proficiency in discerning fine-grained manipulation actions within video sequences. To address these limitations, we introduce a fine-grained action reasoning objective during training, wherein the model is trained to explicitly articulate the manipulation actions observed in the demonstration video prior to generating corresponding robotic control outputs, thereby enhancing its capacity for fine-grained action recognition and comprehension. Furthermore, we incorporate a temporal localization task that inserts the agent’s observation images into the demonstration video sequence, training the model to identify the temporal position of these observations within the video, thereby facilitating cross-modal information exchange between video and image representations.

(II) **Action representation: Lack of action labels for video data and the discrepancy in action spaces between the agent and the expert.** Video data typically lacks action annotations, particularly in human videos, impeding the training of VLMs on fine-grained actions. Furthermore, experts in demonstration videos and target agents typically involve different embodiments, and the resulting disparity in action

*Corresponding author.

†Guangyan Chen completed this work during an internship at LimX Dynamics.

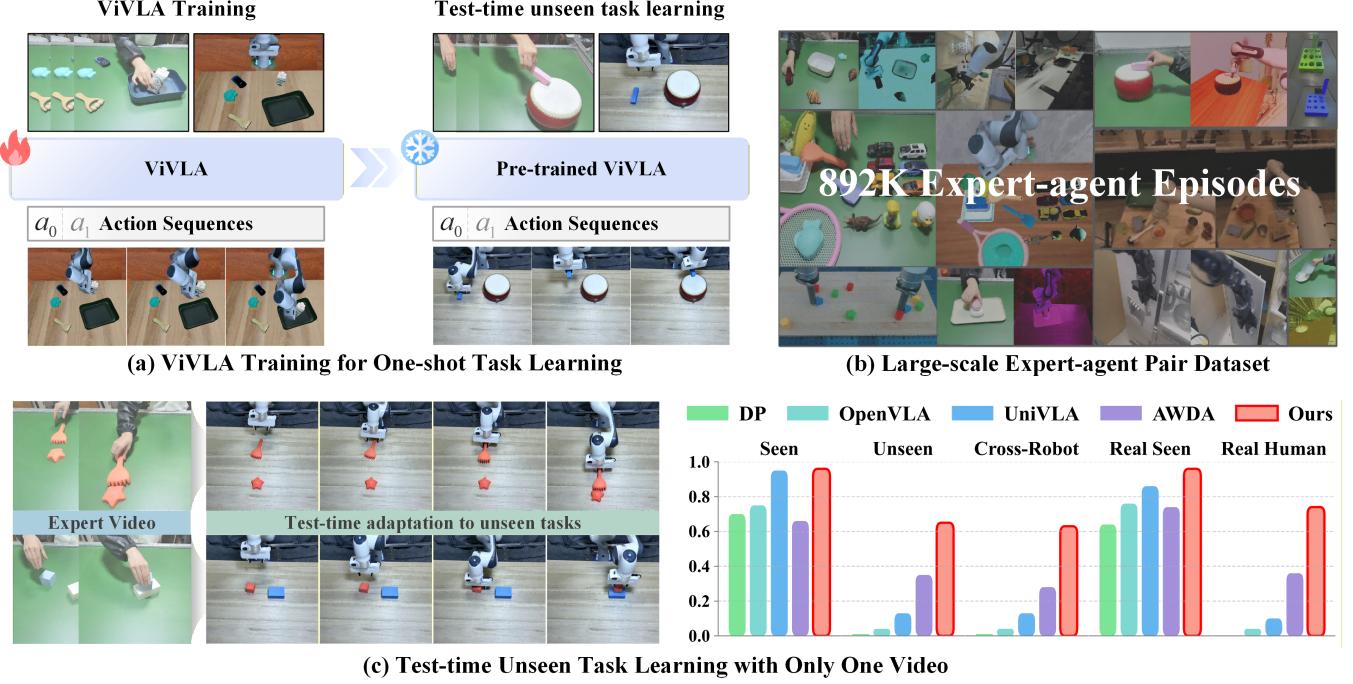


Fig. 1: Illustration of our ViVLA. (a) Our ViVLA is trained to predict subsequent robotic actions conditioned on a single expert demonstration video, endowing the model with the capacity to learn novel tasks from a single expert demonstration video at test time. (b) To push the performance limit of our proposed ViVLA, we develop a scalable expert-agent pair data generation pipeline and compile a large-scale expert-agent pair dataset. (c) Extensive experiments demonstrate that our proposed ViVLA efficiently learns unseen tasks and achieves state-of-the-art performance.

spaces complicates effective knowledge transfer from expert demonstration videos to the agent robot. To address the absence of action annotations, we propose training a latent action tokenizer that derives action representations directly from visual observations. To further bridge the embodiment gap, we train this tokenizer jointly on both expert demonstration videos and agent trajectory data, and introduce an action-centric cycle consistency objective to regularize the learned latent action space, establishing a unified action representation across the agent and the expert. Following the training of the latent action tokenizer, VLMs are trained to predict latent actions from both expert videos and agent observations within this unified latent action space, enabling effective knowledge transfer from expert demonstrations to robotic agents.

(III) Action modeling strategy: Autoregressive action modeling strategy results in shortcut learning and increased inference latency. In autoregressive next-action prediction training, the model has access to all preceding ground-truth action tokens within the sequence. This accessibility allows the model to exploit preceding ground-truth action tokens to predict subsequent actions, thereby inducing shortcut learning and hindering the development of a genuine understanding of expert demonstration videos and agent observations. Moreover, the autoregressive modeling paradigm necessitates sequential token-by-token generation, which introduces substantial inference latency. To mitigate these limitations, we adopt a parallel decoding strategy where the model receives empty action embeddings as input and generates all action tokens concurrently in a single forward pass. This modification prevents information

leakage from preceding action sequences, thereby compelling the model to ground its predictions in a comprehensive analysis of video content and agent observations. Furthermore, the parallel generation of action sequences substantially improves inference efficiency compared to sequential autoregressive decoding. To further enhance the model’s comprehension of expert demonstration videos, we introduce a temporal-spatial masking strategy that stochastically masks video tokens across both temporal and spatial dimensions. This approach reduces computational complexity during training while concurrently establishing a more challenging learning objective that necessitates action prediction from partially observed expert demonstrations, thereby fostering holistic video understanding.

(IV) Dataset: Scarcity of expert-agent pair data. Training the generalizable ViVLA model necessitates rich and diverse expert-agent pair data, which remain scarce in the robot learning domain, typically human-robot pair data. To address this data scarcity challenge, we develop a video-driven generation pipeline that synthesizes human-robot pairs from readily accessible human demonstration videos. The pipeline first grounds interaction information within human videos, and employs 3D Gaussian splatting to render realistic 4D scenes depicting an agent robot executing the demonstrated tasks, then produces observation-action data. Human-robot training pairs are constructed by pairing human videos with the generated robot demonstrations for the same task. We collect 7,421 human videos covering over 100 distinct manipulation tasks, and construct the Human2Robot dataset containing 89,736 human-robot paired training samples through this pipeline.

Additionally, we utilize the open source datasets and construct pairs between data with similar tasks, yielding 803,175 paired samples. In total, we culminate 892,911 expert-agent samples for training ViVLA.

Building upon the aforementioned insights, we introduce a novel VLA paradigm for robotic learning, termed ViVLA, which empowers robots to learn novel tasks from a single expert demonstration video at test-time. The ViVLA framework comprises two key stages: (I) Latent Action Learning with Action-Centric Cycle-Consistency (A3C). We develop a latent action tokenizer that extracts latent action representations from observation sequences, incorporating action-centric cycle-consistency constraints to establish a unified latent action space spanning both expert videos and agent demonstrations. (II) ViVLA Training for One-Shot Task Learning. The ViVLA model is trained to predict latent action sequences for both expert videos and agent observations through parallel decoding, conditioned on expert demonstration videos processed via a temporal-spatial masking strategy, in conjunction with language instructions and agent observations. To enhance the performance of ViVLA, we develop a video-driven expert-agent paired data generation pipeline capable of generating paired samples from easily accessible human videos. Additionally, we leverage the publicly available datasets to construct pairs between data with similar tasks. In total, we culminate 892,911 expert-agent paired samples for training our ViVLA. Experimental results demonstrate that our approach achieves more than 30% improvement on unseen tasks in the LIBERO benchmark and exhibits over 35% improvement using videos from different embodiments. Our method effectively distills knowledge from human videos, consistently yielding improvements exceeding 38% on real-world unseen tasks.

Our contributions can be summarized as follows: (I) We propose a novel VLA paradigm, ViVLA, which is able to effectively distill fine-grained manipulation knowledge from expert behavior and transfer it seamlessly to the agent. This paradigm enables policy models to acquire novel manipulation skills at test time from a single demonstration without necessitating further training or fine-tuning. (II) We introduce a latent action learning framework incorporating cycle-consistency constraints to establish a unified latent action space that encompasses both expert demonstration videos and robot demonstrations. Furthermore, we employ a parallel decoding mechanism to mitigate the shortcut learning issue and enhance computational efficiency during inference. (III) We present a scalable expert-agent pair data generation pipeline that synthesizes trajectory pairs from easily available video sources and integrates curated examples from publicly available datasets. Leveraging this pipeline, we construct a large-scale dataset containing 892,911 expert-agent paired trajectories across a diverse range of manipulation tasks. (IV) Experiments demonstrate that ViVLA achieves over 30% improvement on unseen tasks in the LIBERO benchmark and 35% improvement when leveraging videos from different embodiments. Furthermore, our approach effectively distills knowledge from human videos, demonstrating improvements exceeding 38% on real-world unseen tasks.

II. RELATED WORK

A. Vision Language Action Models

Building upon the success of pretrained vision foundation models, large language models (LLMs), and vision-language models (VLMs), Vision-Language-Action models (VLAs) [13, 15–20] have emerged as a promising approach for processing multimodal inputs to generate robotic actions. These methods leverage pretrained vision-language models and adapt them for robotic manipulation, effectively transferring semantic knowledge acquired from web-scale pretraining to robotics applications. A pioneering work, RT-2 [16], introduced a discretization strategy that uniformly quantizes continuous action values into 256 bins per dimension. This approach enables the co-training of web-scale language models on robot trajectory data, facilitating the transfer of semantic understanding to manipulation tasks. Building on this foundation, OpenVLA [17] adopts a similar action discretization approach while training vision-language models on the large-scale Open X-Embodiment (OXE) dataset [21], which aggregates robotic data from 22 distinct embodiments spanning 21 institutions. An alternative direction [12, 18, 22] employs action experts to generate continuous signals. π_0 [18] adapts the PaliGemma VLM by integrating a specialized action expert module that generates continuous actions via flow matching, enabling precise and fluent manipulation skills. While Vision-Language-Action (VLA) models have demonstrated promising capabilities, existing approaches still struggle to generalize to tasks beyond their training distributions. This constraint stands in contrast to human capabilities, where individuals learn novel manipulation knowledge by seamlessly visually imitating others. Motivated by this observation, we take a step forward by endowing VLAs with the ability to learn skills from a single expert demonstration video at test time.

B. One-Shot Imitation Learning

To enable agents to rapidly adapt to novel tasks from minimal demonstrations while incorporating human instructions at test time, Duan et al. [23] first formulated the one-shot imitation learning (OSIL) problem. They proposed a soft-attention-based learning framework for block stacking tasks, where an agent processes both a single successful demonstration and the current observation to predict actions. Building upon this foundation, subsequent research has explored diverse demonstration formats, including cross-embodiment demonstrations [24, 25], demonstrations from varying viewpoints [26], and demonstrations paired with natural language instructions [27, 28]. To enhance task adaptation from these demonstrations, researchers have pursued multiple solution paradigms. These include task embedding conditioning [29, 30], meta-learning frameworks [24, 25], sub-goal prediction mechanisms [26, 31], expressive transformer architectures [32, 33], and contrastive learning of visual representations [33]. Finn et al. [34] extended model-agnostic meta-learning (MAML) to visual imitation learning, enabling robots to rapidly adapt to new manipulation tasks from single demonstrations. T-OSVI [32] combined transformer-based attention mechanisms with self-supervised inverse dynamics losses to learn manipulation

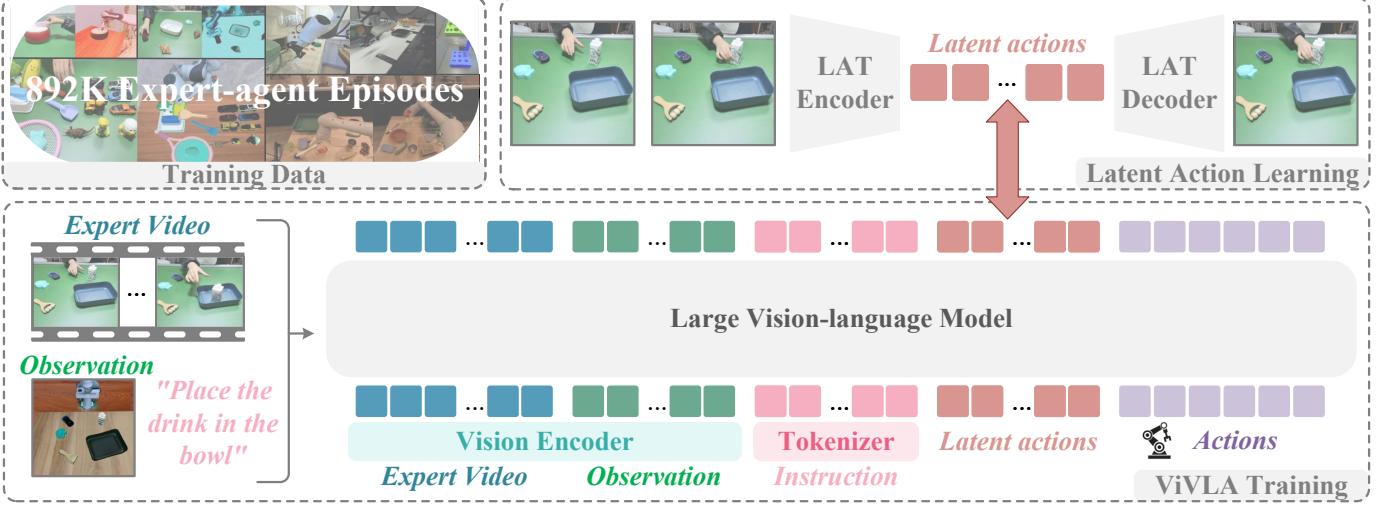


Fig. 2: Overview of our ViVLA. (I) The latent action tokenizer (LAT) learns quantized latent actions from observation sequences, obtaining latent actions for both expert videos and agent demonstrations. (II) The ViVLA model is trained to predict the learned latent action sequences and subsequent robot actions, enabling the robot to acquire novel manipulation skills from only a single expert demonstration video at test time.

skills from single demonstration videos. AWDA [35] improved generalization through a combination of attributed waypoint generation, demonstration augmentation, and image mixup techniques. More recently, OSVI-WM [36] proposed learning world models from expert demonstration videos by encoding visual observations into a shared latent space and predicting future latent trajectories, which are subsequently decoded into physical waypoints for robot execution.

C. Learning from Cross-embodiment Data

A promising goal of one-shot visual imitation learning is to enable robots to acquire skills directly from cross-embodiment demonstration videos, particularly human demonstrations, which offer the advantage of being readily accessible and providing an intuitive means for users to guide robotic behavior. To address the challenges of cross-embodiment learning, including substantial variations in camera perspectives, proprioceptive inputs, and action spaces across different embodiments, numerous approaches have been proposed. Early efforts [37] attempt to bridge these gaps through manual alignment of action spaces. Recent transformer-based methods [38, 39] have been developed to accommodate variable observations and actions more effectively. CrossFormer [39] demonstrates the capability to co-train across four distinct action spaces without imposing constraints on observation spaces or requiring explicit action-space alignment. Flow representations, which capture future trajectories of query points in images or point clouds, have been explored for cross-embodiment learning [40–43]. ATM [40] leverages video demonstrations for pre-training trajectory generation models, utilizing annotations from tracking models [44–46]. More recently, latent action-based methods [15, 47–50] are proposed to learn a discrete codebook, which exhibits greater suitability for the autoregressive training paradigm inherent to VLAs. LAPA [47] introduces an unsupervised learning framework for quantized latent actions between successive

video frames, training models to reconstruct subsequent frames using latent actions and current frames. UniVLA [15] proposes task-centric latent actions that incorporate instructions to decompose transition dynamics into task-irrelevant and task-relevant components. Although these approaches have demonstrated promising performance, they remain limited in learning semantically consistent latent action representations. Moreover, the latent action spaces of different embodiments are typically fragmented, constraining manipulation knowledge transfer across embodiments.

D. Data Augmentation for Policy Learning

In scenarios with limited training data, data augmentation has emerged as a promising approach to generate diverse training samples and enhance policy robustness. Prior research has explored diverse augmentation techniques to improve the resilience of visuomotor policies [51–57]. A representative method is MimicGen [58], which presents an automated system for synthesizing large-scale robotic demonstration datasets from a limited number of human demonstrations. This is achieved by decomposing demonstrations into object-centric segments, spatially transforming these segments and subsequently stitching them together to generate novel trajectories. However, these approaches are predominantly evaluated in simulated environments, requiring task-specific environment construction and facing challenges related to the sim-to-real gap. To facilitate the deployment of learned policies in real-world settings, recent work has explored augmenting scene appearance through image inpainting models [59–62]. For instance, Mirage [63] masks the target robot in images and inpaints the source robot at corresponding poses using URDFs and rendering techniques, enabling the transfer across different robot arms and grippers without requiring target robot training data. Similarly, VISTA [64] generates augmented task

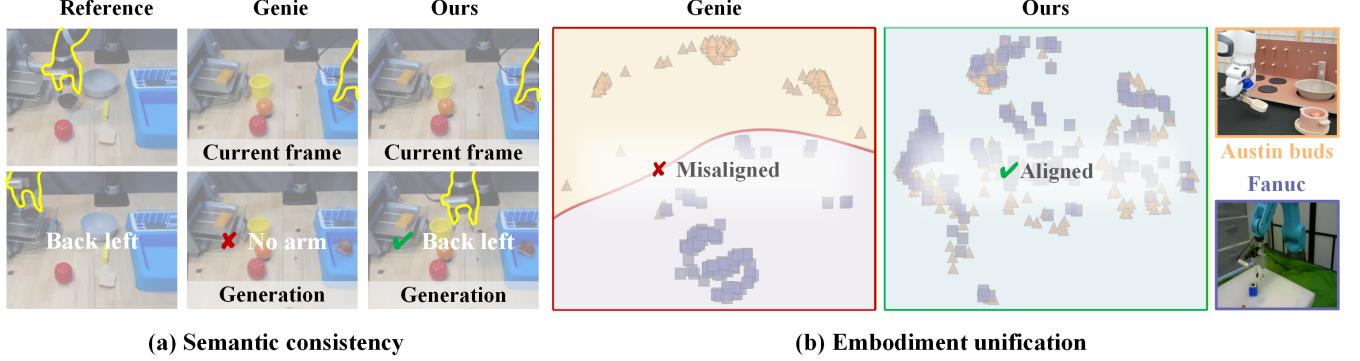


Fig. 3: Motivation for our action-centric cycle consistency. (a) We apply latent actions encoded from reference video frames to the current frame. Existing methods, such as Genie [49], generate frames with divergent motion, revealing limited semantic consistency. (b) We visualize latent action spaces across embodiments, revealing limited cross-embodiment alignment in existing methods. Our method addresses these limitations and constructs a unified latent action space.

demonstrations from diverse camera viewpoints through single-image novel view synthesis models, aiming to learn view-invariant policies [65]. Furthermore, Rovi-aug [66] develops a cross-embodiment pipeline by inpainting robot embodiments into image observations. Nevertheless, these studies primarily perform augmentation on 2D images, which inherently lack spatial information. To address this limitation, RoboSplat [67] reconstructs scenes using 3D Gaussian Splatting and edits the 3D representation for data augmentation. In this work, we propose a video-driven expert-agent pair data generation pipeline. Our pipeline takes human videos as input and employs Gaussian splatting to render robot execution processes. Expert-agent pairs are constructed by pairing human videos with their generated robot demonstrations for the same task.

III. ViVLA

Let $\mathcal{M} = \{\mathcal{M}_1, \dots, \mathcal{M}_{|\mathcal{M}|}\}$ denote a set of tasks, partitioned into disjoint training and testing sets (\mathcal{M}_{train} and \mathcal{M}_{test}). Each training task comprises a set of expert trajectories $\mathcal{T}^e = \{(v_i, \ell_i)\}_{i=1}^{N_e}$, alongside an agent demonstration corpus annotated with actions $\mathcal{T}^a = \{(\tau_i^a, \ell_i)\}_{i=1}^{N_a}$, each episode i is paired with its corresponding language instruction ℓ_i . The expert demonstration trajectory comprises expert video frames $\{v_{i,t}\}_{t=1}^{T_e}$, whereas agent demonstration $\tau_i^a = \{o_{i,t}, a_{i,t}\}_{t=1}^{T_a}$ encompasses both observation data o and their associated action commands a . Within each task, all expert and agent trajectories correspond to variations (e.g., different object configurations) of the same high-level task. The model, trained on \mathcal{M}_{train} , is evaluated on \mathcal{M}_{test} , which contains only expert videos.

Fig. 2 presents the overall architecture of our ViVLA. Our recipe for ViVLA consists of two key stages: (I) Latent action learning with cycle-consistency. Our approach learns latent action representations from observation sequences, while introducing action-centric cycle-consistency constraints to establish a unified latent action space. (II) ViVLA training for one-shot task learning. The ViVLA model is trained to predict action sequences for both expert videos and agent observations via parallel decoding, conditioned on expert demonstration videos processed with a temporal-spatial masking strategy, along with language instructions and agent observations.

A. Latent Action Learning with Cycle Consistency

We learn latent actions from both agent demonstrations and expert videos, thereby providing latent action annotations for these two data sources. Existing approaches learn latent action representations with a temporal window size H , achieved by modeling transition dynamics between successive observation frames I_t and I_{t+H} , training the latent action tokenizer to reconstruct subsequent frames I_{t+H} conditioned on both the learned latent actions and current frames I_t . However, the correspondence between current and future frames is strictly one-to-one, exclusively pairing I_t with I_{t+H} . This permits future frame reconstruction with little understanding of transition dynamics, impeding the extraction of semantically consistent latent actions, as demonstrated in Fig. 3(a). Moreover, we visualize latent action spaces on distinct embodiments in Fig. 3(b) and observe fragmentation within these representational spaces, where distinct regions are assigned to individual embodiments, indicating limited cross-embodiment unification.

To overcome such limitations and establish a unified latent action space for ViVLA training, we propose a latent action learning framework with action-centric cycle consistency and learns latent actions jointly on both expert videos and agent demonstration data. For brevity, we use I to denote visual frames from both expert videos v and agent observations o . As illustrated in Fig. 4, we extract latent actions z_t^q from observation frames $\{I_t, I_{t+H}\}$ and maintain them in a latent action buffer \mathcal{Z} . We concurrently enforce action-centric cycle consistency, where latent actions z_s^q sampled from this buffer are decoded with observation frame I_c to generate future frames \hat{I}_g , and the tokenizer is trained to predict these sampled latent actions from both the observation frames I_c and generated frames \hat{I}_g . As demonstrated in Fig. 3, our approach facilitates the construction of a unified latent action space that achieves both semantic consistency and cross-embodiment unification.

Latent action tokenization. Our method begins by learning latent actions from observation frames through an encoder-decoder architecture. The encoder \mathcal{E} extracts image embeddings f_t and f_{t+H} from both current frames I_t and future frames I_{t+H} using DINOv2 [68]. These extracted embeddings are

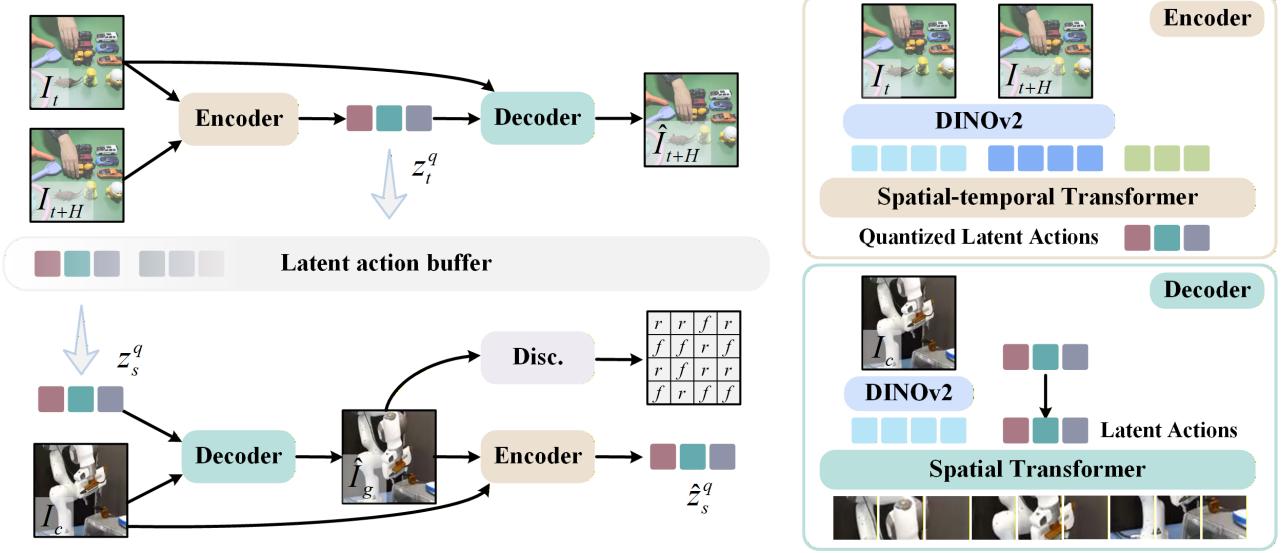


Fig. 4: Illustration of our latent action framework with action-centric cycle consistency. Our approach learns latent action representations from observation frames, while simultaneously introducing action-centric cycle-consistency constraints to establish a unified latent action space.

concatenated with learnable latent action tokens. The combined representations are then processed by a spatial-temporal (ST) transformer, which consists of L spatiotemporal blocks, each incorporating interleaved spatial and temporal self-attention layers. The ST-transformer models temporal transition dynamics between frames and aggregates the learned information into the latent action tokens. The encoded latent action tokens z_t^e are quantized into discrete representations z_t^q , where each latent action is represented using l_z tokens selected from a codebook vocabulary of size K . These quantized action tokens are optimized using the VQ-VAE [69] objective. The encoding procedure for the latent action z_t^q is illustrated below:

$$\begin{aligned} z_t^e &= \text{ST-Transformer}([f_t, f_{t+H}, z]), \quad z_t^e \in \mathbb{R}^{l_z \times c_z} \\ z_t^q &= \text{VQ}(z_t^e), \quad z_t^q \in \mathbb{R}^{l_z \times c_z}. \end{aligned} \quad (1)$$

The decoder \mathcal{D} , implemented as a spatial transformer containing spatial blocks with spatial attention layers, reconstructs the future frames I_{t+H} by processing the learned latent actions z_t^q alongside the current frames I_t . In summary, the latent action tokenization procedure is formulated as follows:

$$\begin{aligned} z_t^q &= \mathcal{E}(I_t, I_{t+H}), \quad z_t^q \in \mathbb{R}^{l_z \times c_z}, \\ \hat{I}_{t+H} &= \mathcal{D}(I_t, z_t^q), \quad \hat{I}_{t+H} \in \mathbb{R}^{w \times h \times c_o}. \end{aligned} \quad (2)$$

Action-centric cycle consistency. We introduce action-centric cycle consistency to regularize the learned latent actions, establishing a unified latent action space that achieves both semantic consistency and cross-embodiment unification. Our method takes observation frames I_c from the dataset and samples a latent action z_s^q from the latent action buffer \mathcal{Z} , where the latent action buffer \mathcal{Z} is constructed by collecting encoded latent actions z_t^q over the previous B batches. We then decode the observation frames I_c with the sampled latent action z_s^q to generate the corresponding subsequent frame \hat{I}_g . The cycle consistency is enforced by providing both the observation frames I_c and generated frames \hat{I}_g into the encoder \mathcal{E} , which

is trained to recover the originally sampled latent action z_s^q :

$$\begin{aligned} \hat{I}_g &= \mathcal{D}(I_c, z_s^q), \quad z_s^q \sim \mathcal{Z}, \\ \hat{z}_s^q &= \mathcal{E}(I_c, \hat{I}_g), \quad \hat{z}_s^q \in \mathbb{R}^{l_z \times c_z}. \end{aligned} \quad (3)$$

To enable gradient propagation, we utilize the latent action embeddings \hat{z}_s^e before quantization and compute their distances to codebook vectors e in the VQ-VAE as similarity measures. We then optimize consistency with the sampled latent actions z_s^q through cross-entropy loss, using the codebook indices of sampled latent actions as supervision signals. The objective is formulated as follows:

$$\mathcal{L}_C = -\sum_{k=1}^K y_k \log\left(\frac{\exp(-d(\hat{z}_s^e, e_k)/\tau)}{\sum_{j=1}^K \exp(-d(\hat{z}_s^e, e_j)/\tau)}\right). \quad (4)$$

where y_k is the one-hot target vector corresponding to the indices of sampled latent actions z_s^q , τ is the temperature parameter, \hat{z}_s^e is the latent action embedding prior to quantization, $\{e_k\}_{k=1}^K$ are the codebook vectors, and $d(\cdot, \cdot)$ denotes the distance metric. Unlike previous methods that are constrained by fixed pairing patterns between current and future frames, limiting the effectiveness of latent action acquisition, our method establishes a challenging self-supervised learning objective. This formulation compels the tokenizer to extract semantically consistent latent actions. Furthermore, our method enables cross-embodiment learning to facilitate the unification of the latent action space across embodiments. Specifically, we sample latent actions z_s^q encoded from embodiment E_i and apply them to observation frames I_c from embodiment E_j , generating frames \hat{I}_g and predicting latent actions \hat{z}_s^q for E_j . Cycle consistency enforces $\hat{z}_s^q \approx z_s^q$ to compel the tokenizer to learn the unified latent actions across embodiments.

Discriminator. The distributional discrepancy between decoder-generated images and dataset images degrades encoder performance. Furthermore, the decoder may leak latent action information directly to the encoder through the generated frame,

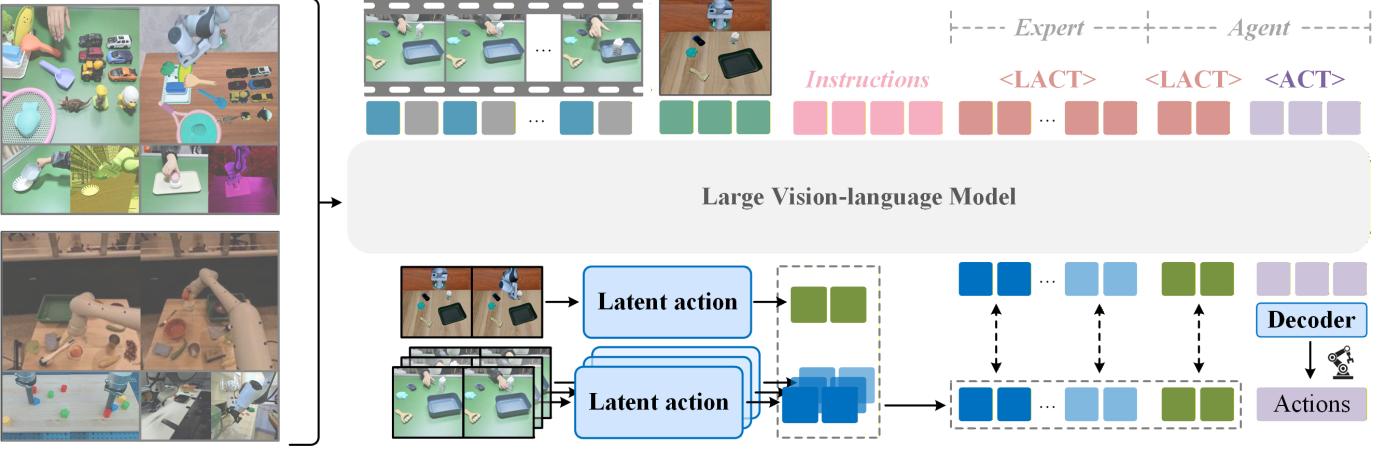


Fig. 5: Illustration of our ViVLA. Our approach jointly processes an expert demonstration video alongside the robot’s visual observations to predict both the demonstrated action sequences and subsequent robot actions, facilitating the ViVLA to distill fine-grained manipulation knowledge from expert behavior and transfer it seamlessly to the agent.

resulting in correspondences that appear cycle-consistent yet remain fundamentally erroneous. To address this issue, we introduce a local-global discriminator that aligns the distribution of generated frames with dataset images across both local details and global style. This regularization simultaneously prevents information leakage since embedding latent actions into generated frames induces distributional deviations that the discriminator penalizes. The discriminator Ψ processes input image frames, including both observation frames from datasets and generated frames, through the spatial transformer to extract corresponding patch features h_l , which are subsequently processed by an MLP to produce patch logits σ_l . Additionally, we apply 2D convolution and global pooling operations to the patch features to derive global features h_g , from which global logits σ_g are obtained, formally expressed as:

$$\begin{aligned} \mathcal{F}_l &= \text{Spatial-Transformer}(I), \quad \sigma_l = \text{MLP}(\mathcal{F}_l), \\ \mathcal{F}_g &= \text{GlobalPool}(\text{Conv2D}(\mathcal{F}_l)), \quad \sigma_g = \text{MLP}(\mathcal{F}_g). \end{aligned} \quad (5)$$

Based on these local and global logits, we define the adversarial losses \mathcal{L}_{GAN} for both the decoder \mathcal{D} and discriminator Ψ , which are formulated as follows:

$$\begin{aligned} \mathcal{L}_{GAN}^{\Psi} &= -\log(\Psi(o)) - (1 - \log(\Psi(\mathcal{D}(o, z)))), \\ \mathcal{L}_{GAN}^{\mathcal{D}} &= 1 - \log(\Psi(\mathcal{D}(o, z))). \end{aligned} \quad (6)$$

These loss formulations are applied across both local and global levels. The objective of the discriminator is to maximize the probability of correctly distinguishing between dataset samples and decoder-generated samples. The decoder is trained to maximize the probability that the discriminator classifies its generated samples as originating from the dataset distribution.

B. ViVLA Training for One-Shot Task Learning

The overview of our ViVLA are illustrated in Fig. 5. Our framework builds upon the Qwen2.5-VL [5] vision-language model, which has demonstrated promising performance in visual recognition and semantic-level video understanding. The architecture consists of a Vision Transformer (ViT) with window attention for efficient processing at native resolutions,

an MLP-based vision-language merger that compresses visual features, and the Qwen2.5 large language model that excels at multi-modal understanding. We initialize the weights with pre-trained parameters and train the model on the expert-agent pair data to enhance the action understanding and skill learning capabilities. The policy model receives expert demonstration videos with temporal-spatial masking, along with robot agent observations, and language instructions, aiming to predict action sequences demonstrated in expert video sequences as well as the subsequent actions performed by agents.

Temporal spatial masking strategy. Video data constitute natural signals characterized by substantial temporal and spatial redundancy. The incorporation of video data generates extensive token sequences that impose considerable computational burdens during training. To mitigate these challenges, we propose a temporal-spatial masking strategy that masks video data across both temporal and spatial dimensions. We apply temporal masking using the same temporal window size as Qwen2.5VL for input video sequences, while preserving the absolute time encoding corresponding to original timesteps to maintain the temporal information. The retained video frames are subsequently processed through the vision encoder, generating a set of patch-wise token representations. We then apply spatial masking to the resulting token representations, forwarding only unmasked tokens to the language model components. This temporal-spatial masking approach substantially reduces video information redundancy while compelling VLMs to perform action prediction on partially observed video sequences, thereby enhancing their capacity for comprehensive video understanding. Following this procedure, the processed video tokens, together with agent observation tokens and language instruction tokens, are encoded into latent representations $\{h_v, h_o, h_{\ell}\}$ for subsequent action prediction.

Parallel decoding. The encoded tokens $\{h_v, h_o, h_{\ell}\}$ are fed into the language model components LM, which are trained to predict action sequences. To mitigate the shortcut learning problem and reduce inference latency, the language model employs parallel decoding to predict action sequences,

encompassing both latent actions and robot actions. Specifically, for latent action prediction, we extend the vocabulary with latent action query tokens for latent action prediction, denoted as LACT. The model receives these latent action query tokens as inputs and simultaneously decodes them in a single forward pass. To enable the model to adaptively determine the initiation of parallel decoding and the number of tokens to be decoded concurrently, we introduce START tokens to the vocabulary: $\{\text{START_LACT_n} \mid n \in \{1, 2, \dots, \text{MAX_N}\}\}$, where MAX_N indicates the maximum number of latent action query tokens. When the model generates START tokens, the corresponding latent action query tokens are appended to the subsequent input according to their designated parallel decoding quantity, generating all action tokens concurrently in a single forward pass. The parallel decoding process is represented as:

$$\begin{aligned} s_t &= \text{LM}(h_{\leq t}), h_{\leq t} = [h_v, h_o, h_\ell, h_p] \\ \{\hat{z}_{t+i}\}_{i=0}^{n-1} &= \text{LM}([h_{\leq t}, s_t, \text{LACT}^n]). \end{aligned} \quad (7)$$

where h_p represents previously generated tokens, $s_t \in \{\text{START_LACT_n} \mid n \in \{1, 2, \dots, \text{MAX_N}\}\}$, LACT^n indicates n LACT tokens, \hat{z} denotes the predicted latent action tokens. Similarly, robot action prediction follows the same parallel decoding mechanism, with the introduction of corresponding START tokens $\{\text{START_ACT_n} \mid n \in \{1, 2, \dots, \text{MAX_N}\}\}$ and action query tokens ACT to enable parallel prediction of action sequences. In contrast to the autoregressive modeling strategy, which has access to all preceding ground truth action tokens during training and necessitates sequential token-by-token processing, inducing the shortcut learning issue and increasing inference latency. The parallel decoding approach effectively prevents information leakage and encourages action prediction based on the understanding of the expert videos and agent observations. Furthermore, concurrent generation of all action sequences substantially enhances inference efficiency.

ViVLA training. The training prediction objectives comprise the latent action sequences $\{z_{v,nH}^q\}_{k=0}^{N_v-1}$ from expert video sequences, along with the subsequent latent actions $z_{o,t}^q$ and robot actions of agent robots, where $N_v = \lfloor T/H \rfloor$, T denotes the video length, and H indicates the temporal window size of the latent action encoding. The latent actions $z_{v,t}^q$ and $z_{o,t}^q$ corresponding to expert video frame v_t and agent observation frame o_t are encoded using the pre-trained latent action tokenizer, following the procedure outlined below:

$$\begin{aligned} z_{v,t}^q &= \mathcal{E}(v_t, v_{t+H}), \\ z_{o,t}^q &= \mathcal{E}(o_t, o_{t+H}). \end{aligned} \quad (8)$$

These latent actions are represented using l_z tokens selected from a codebook vocabulary of size K , which naturally aligns with the discrete prediction paradigm employed by VLMs. We extend the vocabulary by incorporating K specialized tokens: $\{\text{LACT_1}, \text{LACT_2}, \text{LACT_3}, \dots, \text{LACT_K}\}$. Each latent action is mapped to this extended vocabulary based on its corresponding index within the latent action codebook. The optimization objective centers on minimizing the sum of negative log-probabilities for subsequent latent actions:

$$\mathcal{L}_z = \mathbb{E}_z \left[-\sum_{i=1}^{N_z} \log P(\hat{z}_i^q = z_i^q) \right], \quad (9)$$

where N_z represents the total length of latent action tokens. The action query tokens ACT are appended following the latent action query tokens for the robot action prediction, and an action decoder is integrated to transform the predicted action tokens ACT into continuous robot action values. The action decoder receives action embeddings from the final layer of the LM as input. It then aggregates this information through an attention mechanism and pools the representations into a unified embedding, which is subsequently mapped to robot actions via an MLP. The complete architecture is trained end-to-end by jointly optimizing the latent action prediction loss and the L1 loss between ground truth and predicted robot actions.

To facilitate cross-modal information exchange between video and image representations, we incorporate a temporal localization task. Specifically, the agent's observation images are inserted into the expert demonstration video sequence, and the model is trained to identify their temporal positions within the video, thereby enhancing feature exchange across modalities. Additionally, we randomly exclude the expert demonstration video during training, requiring the model to predict latent actions and robot actions based solely on robot observations and language instructions. This training strategy enables our policy to remain robust for seen tasks without access to expert demonstrations.

ViVLA post-training. To effectively transfer the knowledge acquired during pre-training to the target robotic platform, we conduct post-training of our ViVLA model on the target robot. During this post-training phase, the action decoder undergoes full parameter fine-tuning to adapt to the specific action space and control requirements of the target robot, while the pre-trained VLM backbone is fine-tuned using Low-Rank Adaptation (LoRA) [70]. Such a design preserves the rich semantic representations and generalization capabilities learned during pre-training, while simultaneously enabling efficient adaptation to the target robotic domain with minimal computational overhead. The overall procedure follows the same framework established in the pre-training stage.

C. Video-driven Expert-agent Data Generation

One of the keys to training generalizable robotic models lies in diverse and high-quality training data. To generate diverse and high-quality expert-agent pair data, we construct a video-driven expert-agent pair data generation pipeline, as shown in Fig. 6. This pipeline takes human videos as input, utilizes existing vision foundation models to obtain hand poses and object poses, and employs Gaussian splatting to reconstruct 4D scenes of the robot performing the tasks. Expert-agent pairs are constructed by pairing human videos with the generated robot demonstrations for the same task.

Interaction grounding. We utilize vision foundation models to estimate hand and object poses from the provided human video v . For hand tracking, we first apply HaMeR [71] to predict hand shape and pose parameters, while generating a hand mesh model. The Iterative Closest Point (ICP) [72, 73] is further implemented to align the hand mesh with the segmented hand point cloud, yielding precise hand pose trajectories $\xi_H = \{x_C^{H^0}, \dots, x_C^{H^T}\}$ in the camera frame. Subsequently,

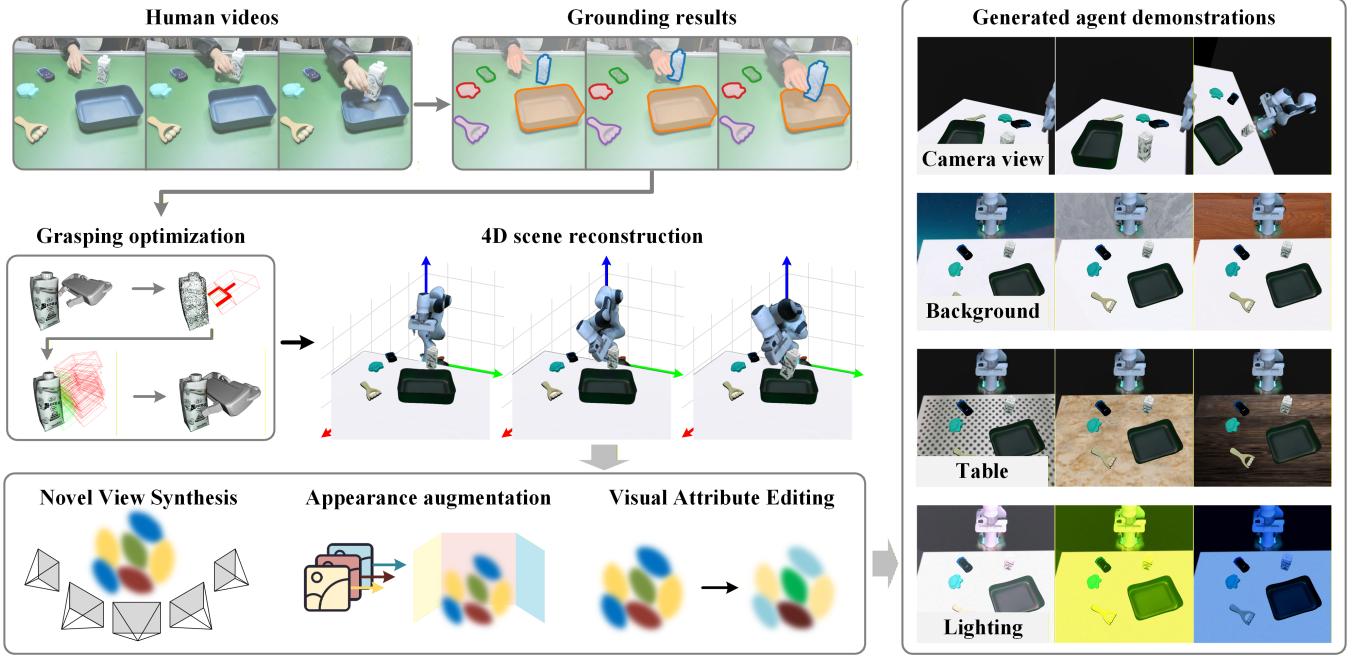


Fig. 6: Illustration of our video-driven expert-agent data generation pipeline. This pipeline takes human videos as input, utilizes existing vision foundation models to obtain hand poses and object poses, and employs Gaussian splatting to reconstruct 4D scenes of different robots performing the tasks. Expert-agent training pairs are constructed by pairing human videos with their corresponding generated robot demonstrations for the same task.

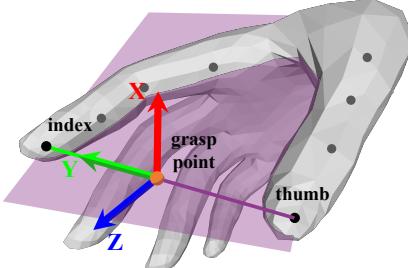


Fig. 7: Calculation of the 6D gripper pose from the estimated hand pose. The grasp point is computed as the midpoint between the thumb and index finger tips. The coordinate frame is defined with the X -axis normal to the plane spanned by all tracked points on both digits, the Y -axis pointing from the grasp point to the index finger tip, and the Z -axis obtained through $\vec{z} = \vec{x} \times \vec{y}$.

these sequences are converted to robot end-effector trajectories $\xi_E = \{\boldsymbol{x}_C^{E^0}, \dots, \boldsymbol{x}_C^{E^T}\}$ [74], as illustrated in Fig. 7. For object tracking, FoundationPose [75] estimates temporal object poses $\xi_\Omega = \{\boldsymbol{x}_C^{\Omega^0}, \dots, \boldsymbol{x}_C^{\Omega^T}\}$ given the object mesh Ω , which is reconstructed from multi-view images via TRELLIS [76].

Video parsing We segment videos into individual clips $\{\tau_i\}_{i=1}^V$, where each clip encapsulates a distinct subtask. This segmentation process is predicated on the identification and utilization of interaction markers, which denote the onset of contact and the termination of contact. Specifically, the point clouds \mathcal{P} of objects and hands are obtained through their respective pose estimation and mesh models. We then compute

inter-object distances and identify contact transitions as follows:

$$\begin{aligned} d &= \text{dist}(\mathcal{P}), \quad t_b = \{t | d^{t-1} > \epsilon \wedge d^t < \epsilon\}, \\ t_e &= \{t | d^{t-1} < \epsilon \wedge d^t > \epsilon\}, \end{aligned} \quad (10)$$

where function dist calculates the distance between any two point clouds. t_b and t_e denote contact initiation and termination, respectively. We classify the clips into the grasping phase and the manipulation phase. In the grasping phase, the objects remain stationary, and the agent executes a reach-and-grasp maneuver targeting the object. In the manipulation phase, the agent manipulates the grasped object, performs a motion, and makes contact between objects.

End-effector pose optimization. We optimize the converted end-effector pose trajectories ξ_E to further refine the contact relationship between the end-effector and the object. During the grasping phase, we replicate the robot end-effector’s trajectory using grounded trajectories and optimize the grasping pose $\boldsymbol{x}_C^{E^t g}$ at the moment of contact t_g , where t_g denotes the termination of the grasping phase. Specifically, we sample N candidate grasps $\{\boldsymbol{x}_{C,i}^{E^t g}\}_{i=1}^N$ within a 6D neighborhood around the initial grounded grasping pose $\boldsymbol{x}_C^{E^t g}$, and subsequently filter for feasible grasping configurations. A grasp is considered feasible if two conditions are satisfied: (I) the end-effector does not collide with the object, and (II) the object lies within the grasping region of the end-effector. For parallel-jaw grippers, the grasping region is typically defined as the region between the gripper jaws. Among feasible candidates, we compute a stability score for each grasp as the normalized perpendicular distance between the gripper plane and the object’s center of gravity (COG) [77]. The feasible grasping pose $\hat{\boldsymbol{x}}_C^{E^t g}$ with the highest stability score is then selected and expressed relative



Fig. 8: The set of objects used for the data generation pipeline.

to the object’s coordinate frame. In the manipulation phase, we adopt the commonly used assumption that the relative pose between the end-effector and object remains invariant throughout the manipulation. Under this constraint, the end-effector trajectories are derived by composing the object’s motion with the optimized grasping pose. The resulting optimized end-effector poses $\hat{\xi}_E$ enable our data generation pipeline to synthesize high-quality 4D robot–object manipulation scenes with improved physical consistency.

High quality 4D scene reconstruction. The grounded object trajectories and optimized end-effector pose trajectories are used to reconstruct 4D robot–object manipulation scenes via 3D Gaussian splatting. During reconstruction, each object is represented using its grounded pose and corresponding Gaussian model \mathcal{G}_Ω , while the robot is represented using joint positions and its Gaussian model \mathcal{G}_R . The robot joint positions are derived from the optimized end-effector trajectories $\hat{\xi}_E$ through motion planning, the object Gaussian models \mathcal{G}_Ω are converted from the object meshes, and the robot Gaussian models \mathcal{G}_R are obtained from the corresponding URDF file or real-world reconstruction. The reconstructed 4D scenes exhibit high visual fidelity and align well with the trajectories observed in the human video v , as demonstrated in Fig. 6.

Diverse expert-agent data generation. Leveraging the reconstructed 4D scene, we generate robot demonstrations through diverse demonstration augmentations [67], which operate along three dimensions: (I) Camera view. A key advantage of 3D Gaussian splatting lies in its capability for novel view synthesis. We exploit this property to generate demonstrations from multiple viewpoints by rendering the reconstructed 4D scene from diverse camera poses, thereby enriching the visual diversity of the training data. (II) Scene appearance. We augment scene appearance by applying varied textures to the table and background 3D Gaussian planes surrounding the manipulation workspace. This enables the generation of diverse scene compositions that enhance the model’s robustness to environmental variations. (III) Lighting condition. We diversify lighting conditions by applying random scaling factors, offsets, and noise perturbations to the diffuse color attributes of each Gaussian in the reconstructed scene. We

TABLE I: Training data mixture.

Datasets	Num of Traj.
Fractal [79]	87,212
Bridge [80]	60,064
Droid [81]	49,933
Language Table [82]	442,226
BC-Z [27]	43,264
FMB Dataset [83]	4,592
Ego4D [81]	18,368
EgoDex [84]	97,516
Human2Robot	89,736
Overall	892,911

curate 7,421 human demonstration videos spanning over 100 distinct manipulation tasks, and construct the Human2Robot dataset containing 89,736 human-robot paired training samples through this pipeline. The objects used in the data collection process are visualized in Fig. 8.

Dataset collection from publicly available dataset. In addition to our generated dataset, we incorporate publicly available datasets to further enrich the training corpus by constructing expert-agent pairs from demonstrations of similar tasks. Task similarity is quantified by computing the cosine similarity between Sentence-BERT embeddings [78] of task instructions, and we select pairs that exceed a predefined similarity threshold (set to 0.9 in our implementation). Through this process, we collect 803,175 paired samples from publicly available datasets. In total, we culminate 892,911 expert-agent paired samples for training our ViVLA.

IV. IMPLEMENTATION DETAILS

Our training corpus consists of three primary data categories: robotic manipulation data, human video data, and synthetically generated human-robot pair data (our Human2Robot dataset). For robotic manipulation, we incorporate a selected subset from the Open X-Embodiment dataset [85], specifically focusing on single-arm end-effector control tasks. The human video data comprises ego-centric recordings from the Ego4D dataset [86] and EgoDex [84], capturing first-person perspectives of diverse human activities. We leverage the demonstrations from robotic manipulation data and human video data by generating expert-agent pairs through the combination of trajectories with semantically similar task demonstrations. Our Human2Robot is synthesized through a video-driven expert-agent data generation pipeline, which we detail in Sec. III-C. A comprehensive breakdown of the dataset composition is provided in Table I.

For latent action learning with action-centric cycle consistency, our encoder leverages a pre-trained ViT-base DINov2 model to extract visual embeddings from input images. These embeddings are subsequently processed through a spatial-temporal transformer consisting of 12 layers with 768-dimensional hidden states. The resulting latent action representations are discretized via vector quantization using a codebook containing 16 entries. Both the decoder and discriminator networks employ spatial transformers with identical architectural specifications, each comprising 12 layers and maintaining 768-dimensional representations throughout. To

facilitate temporal consistency during training, we maintain a latent action buffer that accumulates encoded latent actions from the preceding 4 batches. The entire latent action tokenizer is optimized using the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-2} .

For training ViVLA, we adopt a global batch size of 256, distributed across GPUs with 8 samples per device. The model is trained for 30,000 optimization steps using a constant learning rate of 2×10^{-5} . To enhance robustness, we apply spatiotemporal masking to the input videos, with masking ratios randomly sampled from the range $[0, 0.5]$ for both temporal and spatial dimensions. To strengthen the model’s fine-grained action reasoning capabilities, we perform fine-grained action reasoning training on individual videos with probability 0.4, where the model exclusively predicts latent action sequences within the expert video demonstration without proceeding to subsequent latent action and robot action prediction for the robot agent.

V. EXPERIMENTS

We perform experiments to answer the following questions:

- Can ViVLA effectively learn unseen tasks from a single expert demonstration video, enhancing the adaptability of robotic agents?
- Does ViVLA possess robust cross-embodiment learning capabilities that enable skill acquisition from expert demonstration videos with different robotic platforms?
- Can ViVLA directly learn robotic skills from human demonstration videos and perform robustly in real-world robotic scenarios?

We first assess the capability for unseen task learning through a comparative evaluation of ViVLA against state-of-the-art VLA baselines and one-shot imitation learning (OSIL) methods (Sec. V-A). We then examine ViVLA’s capacity for cross-embodiment skill transfer in Sec. V-B. We further demonstrate that ViVLA enables effective learning from human demonstration videos and achieves robust task execution in real-world environments (Sec. V-C). Finally, we present comprehensive ablation studies in Sec. V-E to systematically validate the contribution of each component within our framework.

A. Unseen Task Learning

We first investigate the capacity of ViVLA to learn unseen tasks from videos with the same embodiment.

Baselines. ViVLA is compared with four representative methods: (1) Diffusion Policy [87], which represents robot visuomotor policies as conditional denoising diffusion processes, enabling stable training and natural handling of multimodal action distributions in high-dimensional spaces. (2) AWDA [35], which achieves one-shot visual imitation by predicting attributed waypoints from demonstration videos, executing them via hand-crafted motor primitives. (3) OpenVLA [10], a VLA based on Prismatic7B [88] and trained on the OXE [21] dataset, was post-trained on the LIBERO benchmark. (4) UniVLA [15], which learns task-centric latent actions from diverse cross-embodiment videos without requiring action labels, pre-training the VLA model on the action-less dataset.

Experiment setup. We conduct our evaluation on the LIBERO benchmark [89], a comprehensive suite comprising 130 language-conditioned manipulation tasks. Following the experimental protocol established by OpenVLA [10], we focus on four specialized suites: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-Long. LIBERO-Spatial evaluates robustness to environmental configuration changes while maintaining consistent object types. LIBERO-Object assesses generalization across object variations under fixed spatial arrangements. LIBERO-Goal tests adaptability to different task objectives while preserving object categories and spatial layouts. LIBERO-Long presents the most challenging scenarios, requiring long-horizon reasoning across diverse object categories and spatial configurations. Each suite contains 10 distinct tasks. To evaluate generalization capabilities, we designate 8 tasks per suite as seen tasks for training and reserve 2 tasks as unseen tasks for testing. The training set is constructed by aggregating demonstrations from all seen tasks into a consolidated dataset. For expert-agent pair construction, we leverage robot execution trajectories from the dataset, pairing videos of the same task performed by the robot as expert demonstrations. All baseline methods are trained on this consolidated dataset and subsequently evaluated on both seen and unseen task splits to assess in-distribution performance and generalization ability.

Experiment results. Table II presents a comprehensive performance comparison across all evaluated task suites, demonstrating that ViVLA consistently outperforms baseline methods. Diffusion Policy (DP) exhibits limited generalization capabilities when applied to unseen tasks. Despite pre-training on the large-scale OXE dataset and subsequent post-training on the integrated LIBERO dataset, OpenVLA and UniVLA exhibit significant degraded performance on unseen tasks. This observation corroborates our hypothesis that current methods encounter significant difficulties in generalizing to novel tasks, typically requiring task-specific data collection and fine-tuning procedures to acquire novel capabilities. Our approach also surpasses AWDA on both seen and unseen tasks, where AWDA similarly exploits expert demonstration videos for unseen task generalization. These results substantiate our method’s capacity to effectively extract fine-grained manipulation knowledge from expert demonstration videos and effectively transfer this knowledge to novel task configurations, enabling our approach to learn novel tasks from only a single expert video. The substantial performance gains on unseen tasks highlight the effectiveness of our approach in learning transferable representations that extend beyond the training distribution.

B. Unseen Task Learning with Cross-robot Videos

We then investigate the ability of ViVLA to learn novel skills from expert demonstration videos of manipulation tasks performed by a different robot platform.

Experiment setup. The unseen task learning experiments with cross-robot videos are conducted on the LIBERO benchmark [89]. To construct the expert-agent pair dataset, we replay the LIBERO dataset using a UR robotic arm, where the Franka arm serves as the agent and the UR arm provides expert

TABLE II: Success rate on the LIBERO benchmark. We partition the LIBERO dataset into seen and unseen tasks and report the success rates of the compared methods on each subset.

Methods	DP	OpenVLA	UniVLA	AWDA	AWDA _R	Ours	Ours _R	
Seen	0.70	0.75	0.95	0.66	0.62	0.96	0.95	
Unseen	0.01	0.04	0.13	0.35	0.28	0.65	0.63	
LIBERO-Spatial		LIBERO-Object		LIBERO-Goal		LIBERO-Long		
Methods	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
Diffusion Policy [87]	0.76	0.01	0.90	0.01	0.68	0.00	0.45	0.00
OpenVLA [17]	0.82	0.05	0.88	0.07	0.76	0.02	0.55	0.01
UniVLA [15]	0.95	0.16	0.96	0.23	0.95	0.07	0.92	0.05
<i>Learning from videos with the same embodiment</i>								
AWDA [35]	0.71	0.40	0.78	0.50	0.63	0.28	0.51	0.20
Ours	0.98	0.70	0.98	0.74	0.96	0.62	0.92	0.54
<i>Learning from videos with the different embodiment</i>								
AWDA _R [35]	0.69	0.32	0.74	0.41	0.57	0.21	0.49	0.17
Ours_R	0.95	0.71	0.98	0.73	0.95	0.58	0.92	0.51

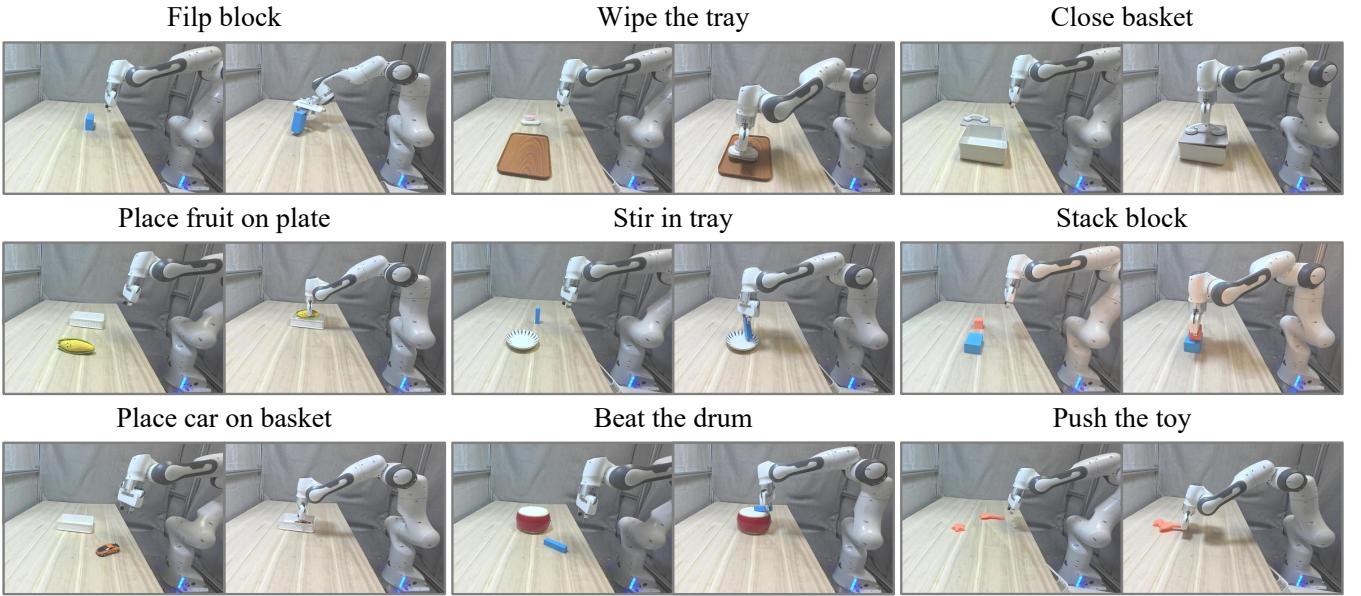


Fig. 9: Example qualitative results for real-world manipulation task.

demonstrations. The model is post-trained on this pair dataset. During inference, the model takes demonstration videos of the UR arm performing manipulation tasks as input and generates corresponding control actions for the Franka arm to execute these tasks. All other experimental settings remain consistent with those employed in Sec. V-A.

Results. The empirical results for baseline approaches and the proposed method are reported in Table II. Despite utilizing expert demonstration videos from heterogeneous robotic platforms, our method achieves superior performance on both seen and unseen tasks across all experimental suites. Notably, on unseen tasks, our approach exhibits strong skill acquisition capabilities by learning new skills from merely a single demonstration video at test time and executing them robustly.

The use of cross-embodiment videos incurs only marginal performance degradation compared to demonstrations from the same robotic platform. This robustness can be attributed to our latent action learning framework with cycle consistency, which effectively learns a unified latent action space. Such a design enables our method to extract generalizable manipulation knowledge from demonstrations across diverse embodiments, thereby facilitating effective manipulation knowledge extraction from cross-robot videos.

C. Unseen Task Learning with Human Videos

Experiment setup. To validate our model's capability to acquire novel skills from human demonstration videos, we train the model on our Human2Robot dataset, where humans provide

TABLE III: The real-world task learning experiment results with human videos. We report the success rates of comparison methods on seen tasks and unseen tasks. Seen tasks are indicated in green, and unseen tasks are indicated in red.

Methods	DP	OpenVLA	UniVLA	AWDA	Ours
Seen	0.64	0.76	0.86	0.74	0.96
Unseen	0.00	0.04	0.10	0.36	0.74
Methods	Flip block	Wipe the tray	Close basket	Place fruit on plate	Stir in tray
Diffusion Policy [87]	0.7	0.6	0.5	0.8	0.6
OpenVLA [17]	0.6	0.8	0.8	0.9	0.7
UniVLA [15]	0.9	0.9	0.8	0.9	0.8
AWDA [35]	0.8	0.6	0.7	0.8	0.8
Ours	1.0	0.9	1.0	1.0	0.9
Methods	Stack block	Place car on basket	Beat the drum	Push the toy	Pour from bowl to plate
Diffusion Policy [87]	0.0	0.0	0.0	0.0	0.0
OpenVLA [17]	0.1	0.1	0.0	0.0	0.0
UniVLA [15]	0.2	0.3	0.0	0.0	0.0
AWDA [35]	0.3	0.4	0.3	0.5	0.3
Ours	0.8	0.8	0.6	0.7	0.8



Fig. 10: Real-world experiment setup.

expert demonstrations and Franka robots serve as learning agents to construct the expert-agent pair data. We evaluate ViVLA on 12 real-world manipulation tasks, consisting of 6 seen tasks and 6 unseen tasks. During evaluation, the model receives human demonstration videos corresponding to each task as input, and generates control actions for the Franka arm to execute these tasks. The real-world experimental setup is illustrated in Fig. 10, featuring a seven-degree-of-freedom Franka Emika robot arm [90]. Task success is determined through human evaluation, with success rates computed over 10 trials featuring randomized object positions and orientations.

Results. The qualitative results are presented in Fig. 9, with quantitative analysis detailed in Table III. The experimental results demonstrate that ViVLA achieves markedly superior performance compared to baseline methods. Our model achieves high success rates on seen tasks, validating the high fidelity

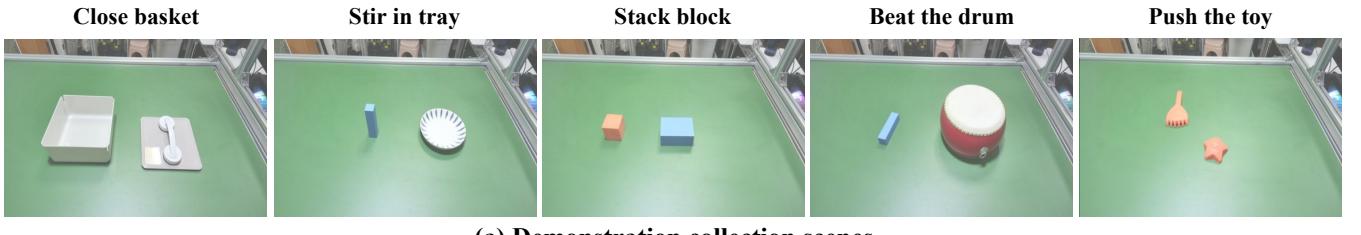
of our generated robot data from the expert-agent pair data generation pipeline. On unseen tasks, our method demonstrates promising performance, indicating its capability to effectively extract manipulation knowledge from human demonstration videos and transfer it to the robotic agent in real-world scenarios. These results substantiate that ViVLA effectively bridges the embodiment gap between human demonstrations and robotic systems through its latent action learning framework. The learned unified latent action space enables effective knowledge transfer from human videos to the robot agent, facilitating efficient acquisition of new manipulation skills. Collectively, these findings demonstrate that ViVLA can learn new skills from a single human demonstration video, providing a practical approach for end-users to teach robots novel tasks.

D. Robustness Analysis

Object count and spatial arrangement. We explore the model’s robustness to variations in object quantities and spatial distributions between expert demonstration videos and agent robot manipulation scenarios. Experiments are conducted on five representative tasks: close basket, stir in tray, stack block, beat the drum, push the toy, where the first two constitute seen tasks and the latter three represent unseen tasks. Human demonstration videos are collected in the scenarios illustrated in Fig. 11 (a), while evaluation is performed in the scenarios depicted in Fig. 11 (b). The experimental results presented in Table IV demonstrate that ViVLA maintains high success rates and exhibits robust performance despite variations in object quantities and spatial distributions between training and evaluation scenarios. This robustness validates the effectiveness of our approach in generalizing across diverse environmental configurations.

TABLE IV: Success rates on environments with different object counts and spatial distribution. vo denotes the variants evaluated in environments with different object quantities and spatial distribution.

Methods	Close basket	Stir in tray	Stack block	Beat the drum	Push the toy	Overall
Ours _{vo}	0.8	0.7	0.7	0.6	0.7	0.70
Ours	1.0	0.9	0.8	0.6	0.7	0.80



(a) Demonstration collection scenes



(b) Testing scenes

Fig. 11: Robustness analysis on object count and spatial distribution variations. Seen tasks are indicated in green, and unseen tasks are indicated in red.

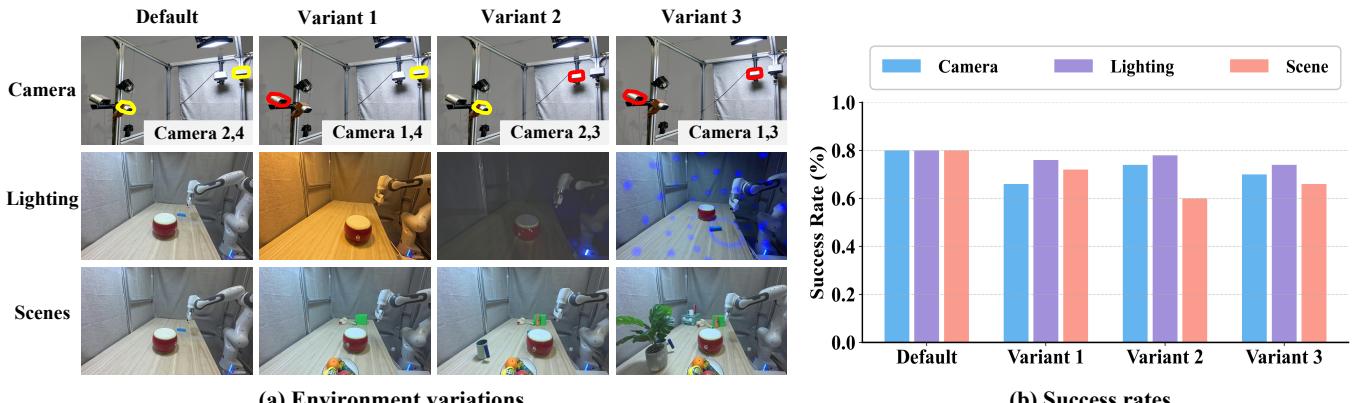


Fig. 12: Environmental variations in terms of camera viewpoints, lighting, and scenes, the varying camera viewpoints are highlighted in red. Our ViVLA exhibits strong robustness against these variations.

Environment generalization. We explore the model’s robustness to environmental variations between expert demonstration videos and agent robot scenarios, including changes in camera perspectives, lighting conditions, and scene settings. For each factor, we design three experimental variants. Validation is conducted across five representative manipulation tasks: close basket, stir in tray, stack block, beat the drum, push the toy, where the first two constitute seen tasks and the latter three represent unseen tasks. As illustrated in Fig. 12, our method exhibits remarkable stability under varying lighting conditions, with minimal performance degradation. While camera perspective and scene variations exert a more pronounced impact on performance, our approach nevertheless maintains considerable robustness across different environmental settings.

These results collectively demonstrate that our method achieves consistent performance across diverse environmental conditions, underscoring its adaptability and generalization capabilities.

E. Ablation Analysis

To investigate the fundamental designs of our ViVLA approach, we perform comprehensive ablation experiments. These design decisions are evaluated in unseen task learning experiments with cross-robot videos, with performance quantified via success rate metrics.

Latent action prediction. We investigate the role of latent action prediction in learning from video demonstrations. Removing latent action prediction for both expert videos and agent robots results in significant performance degradation. This

TABLE V: Ablation studies with ViVLA on task learning with cross-robot videos. Default settings are marked in gray. A3C denotes our action-centric cycle consistency.

Methods	LIBERO-Spatial		LIBERO-Object		LIBERO-Goal		LIBERO-Long	
	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
<i>(a) Latent action prediction</i>								
W/o prediction	0.91	0.48	0.89	0.55	0.81	0.47	0.76	0.33
Genie	0.93	0.65	0.91	0.62	0.87	0.53	0.82	0.41
A3C	0.95	0.71	0.98	0.73	0.95	0.58	0.90	0.51
<i>(b) Latent action learning framework designs</i>								
W/o discriminator	0.89	0.54	0.91	0.59	0.86	0.45	0.81	0.38
Local discriminator	0.93	0.68	0.96	0.69	0.94	0.55	0.87	0.46
W/o Latent action buffer	0.92	0.66	0.97	0.70	0.91	0.50	0.86	0.46
A3C	0.95	0.71	0.98	0.73	0.95	0.58	0.90	0.51
<i>(c) Temporal-spatial masking strategy</i>								
W/o masking	0.94	0.64	0.98	0.67	0.94	0.54	0.91	0.46
Spatial masking	0.95	0.69	0.97	0.72	0.95	0.56	0.90	0.48
Temporal-spatial masking	0.95	0.71	0.98	0.73	0.95	0.58	0.90	0.51
<i>(d) Parallel modeling</i>								
Auto-regressive	0.92	0.63	0.96	0.67	0.93	0.52	0.87	0.35
Parallel modeling	0.95	0.71	0.98	0.73	0.95	0.58	0.90	0.51

TABLE VI: Ablation analysis on language instructions and expert video demonstrations.

Methods	Close basket	Stir in tray	Stack block	Beat the drum	Push the toy	Overall
Ours _{w/o lang}	1.0	0.9	0.7	0.6	0.6	0.76
Ours _{w/o video}	0.9	0.9	0.3	0.1	0.0	0.44
Ours	1.0	0.9	0.8	0.6	0.7	0.80

observation suggests that the latent action prediction pretraining task enhances the model’s capability to recognize fine-grained actions in videos while simultaneously constructing a unified latent action space that bridges the embodiment gap between experts and agents, thereby improving the efficiency of novel skill learning from the expert video. Compared to baseline methods that employ Genie [49] for latent action learning, a widely adopted approach, our method achieves substantial performance improvements. This comparison validates the critical role of cycle consistency in learning effective latent action representations. The experimental results confirm that our latent action learning framework with cycle consistency captures semantically meaningful latent actions, leading to enhanced performance.

Latent action learning framework designs. We further investigate the contributions of key components in our latent action learning framework with cycle consistency, specifically the latent action buffer and the discriminator. The experimental results are presented in Table V(b). Removing the discriminator leads to substantial performance degradation, indicating that without discriminator supervision, the decoder can leak information about the sampled latent action to the encoder, thereby undermining the effectiveness of our

latent action learning approach. Additionally, employing a commonly used local discriminator that supervises individual patches still results in a performance decline. This validates the effectiveness of our local-global discriminator design, which effectively prevents information leakage and mitigates distribution mismatch between generated and dataset video frames, thus enhancing the efficiency of latent action learning. We also examine the role of the latent action buffer. The variant that excludes the buffer and instead directly uses latent actions learned from the current batch, exhibits significant performance degradation. This decline can be attributed to the collapse of the latent action space, demonstrating that maintaining a dynamic buffer of historical latent actions is essential for stable and effective latent action learning.

Temporal-spatial masking strategy. We investigate the impact of the temporal-spatial masking strategy on model performance. As shown in Table V(c), removing masking entirely preserves performance on seen tasks while causing degradation on unseen tasks. This effect can be attributed to the fact that seen tasks are present in the training data, allowing the model to leverage inherent manipulation knowledge to complete these tasks even with weakened video understanding capabilities. In contrast, unseen tasks require the model to

extract fine-grained manipulation knowledge directly from expert demonstration videos. The absence of the masking strategy reduces the effectiveness of latent action prediction pretraining in fostering robust video understanding capabilities, thereby impairing performance on novel tasks. The variant employing only spatial masking also exhibits performance degradation, though less severe than complete removal. This indicates that our temporal-spatial masking approach creates a more challenging pretraining objective that requires the model to predict actions from partially visible expert demonstrations through holistic spatiotemporal video comprehension.

Parallel modeling. We investigate different action modeling approaches in Table V(d). Adopting autoregressive modeling results in performance degradation on seen tasks and a decline on unseen tasks. This suggests that autoregressive modeling is susceptible to shortcut learning, which undermines the effectiveness of latent action prediction pretraining in developing the model’s capability to comprehend fine-grained manipulation. Furthermore, we observe that the performance degradation is particularly pronounced on unseen tasks within the long suite compared to other task suites. This observation indicates that autoregressive modeling poses greater challenges for long-horizon latent action prediction tasks, where the accumulation of prediction errors across extended sequences exacerbates performance decline.

Language exclusion. We validate the model’s semantic understanding capability for video demonstrations. We provide only expert videos and agent observations as input while removing language instructions, thereby requiring the model to infer task semantics solely from visual demonstrations. Experiments are conducted across five representative manipulation tasks: close basket, stir in tray, stack block, beat the drum, push the toy, where human videos serve as expert demonstrations. The experimental results presented in Table VI show that our method maintains robust task completion even without language instructions, exhibiting only marginal performance degradation. This demonstrates that ViVLA preserves strong video semantic understanding capabilities while developing fine-grained action recognition abilities. The model’s capacity to comprehend task objectives from visual demonstrations alone highlights the effectiveness of our approach in learning rich, semantically meaningful representations from expert videos.

Video exclusion. We further validate the model’s capability to operate using language instructions alone, without relying on expert video demonstrations. In this experiment, we provide only language instructions and agent observations as input while excluding expert demonstration videos. The experimental results presented in Table VI reveal a clear distinction between seen and unseen tasks. For seen tasks, ViVLA maintains robust performance with language instructions alone, achieving promising success rates that indicate effective internalization of learned skills during training. In contrast, performance on unseen tasks degrades significantly without video demonstrations, with success rates dropping dramatically. This performance gap underscores our method’s capacity to extract and learn novel behavioral patterns from video demonstrations, a capability that proves essential for generalizing to previously unseen manipulation tasks.

VI. LIMITATIONS

During generalization to unseen tasks, the predominant failure cases arise from errors in basic manipulation operations, particularly imprecise grasping and inaccurate placement. We attribute these failures to two primary factors: First, under specific spatial arrangements of objects, occlusions may impede the static camera’s ability to capture fine-grained robot-object interaction details. Second, since static cameras predominantly capture background regions, the proportion of task-relevant visual information input to the model remains limited. In contrast, the model can compensate for these perceptual limitations on seen tasks by leveraging its internalized manipulation knowledge acquired during training. We posit that incorporating a wrist-mounted camera represents a direct solution to this challenge, as it would enable the model to consistently observe detailed robot-object manipulations from an egocentric perspective while substantially increasing the density of task-relevant visual information in the input stream.

Additionally, we identify two methodological aspects warranting improvement: (I) Our experiments reveal that the model exhibits promising error recovery capabilities during task execution. However, these capabilities can be systematically strengthened through targeted data augmentation. Specifically, during the agent demonstration generation stage of our data pipeline, we can introduce controlled trajectory perturbations paired with corresponding recovery sequences. By exposing the model to diverse failure scenarios and their corrections during training, we can enable it to learn robust error recovery strategies directly from data, thereby improving task execution reliability in real-world deployment. (II) To maintain demonstration quality, our current pipeline leverages manually collected human videos. A promising direction for future work is to leverage internet-scale human videos to automatically generate expert-agent pair data. This would require developing a robust pipeline for video filtering, task identification, and quality assessment, but could significantly expand the diversity and volume of training data available for learning generalizable manipulation policies.

VII. CONCLUSION

This study presents ViVLA, a generalist policy learning architecture that enables efficient skill acquisition from single-demonstration observation without requiring subsequent fine-tuning. Our approach processes expert demonstration videos, robot agent observations, and language instructions to predict both the latent actions exhibited in expert demonstrations and the subsequent latent actions executed by the agent. To push the performance limit of our proposed ViVLA, we develop a scalable expert-agent pair data generation pipeline capable of synthesizing paired trajectories from easily accessible videos, further augmented by curated pairs from open source datasets. Through this pipeline, we compile a large-scale dataset comprising 892,911 expert-agent paired trajectories spanning diverse manipulation tasks. Extensive experiments demonstrate that our proposed ViVLA is able to learn unseen tasks from a single expert demonstration at inference time. Our approach achieves more than 30% improvement on unseen tasks in the LIBERO

benchmark, and attains exceeding 35% improvement when leveraging videos from different embodiments. Furthermore, our approach effectively distills knowledge from human videos, demonstrating a gain of over 38% on real-world unseen tasks.

REFERENCES

- [1] A. Liu, B. Feng, *et al.*, “Deepseek-v3 technical report,” *arXiv preprint arXiv:2412.19437*, 2024.
- [2] J. Bai, S. Bai, *et al.*, “Qwen technical report,” *arXiv preprint arXiv:2309.16609*, 2023.
- [3] H. Touvron, L. Martin, *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [4] J. Achiam, S. Adler, *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [5] S. Bai, K. Chen, *et al.*, “Qwen2. 5-vl technical report,” *arXiv preprint arXiv:2502.13923*, 2025.
- [6] K. Team, A. Du, *et al.*, “Kimi-vl technical report,” *arXiv preprint arXiv:2504.07491*, 2025.
- [7] C. Team, “Chameleon: Mixed-modal early-fusion foundation models,” *arXiv preprint arXiv:2405.09818*, 2024.
- [8] H. Liu, C. Li, *et al.*, “Visual instruction tuning,” *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.
- [9] B. Xiong, B. Chen, *et al.*, “Bluelm-2.5-3b technical report,” *arXiv preprint arXiv:2507.05934*, 2025.
- [10] M. J. Kim, K. Pertsch, *et al.*, “Openvla: An open-source vision-language-action model,” *arXiv preprint arXiv:2406.09246*, 2024.
- [11] K. Black, N. Brown, *et al.*, “ π_0 : A vision-language-action flow model for general robot control,” *arXiv preprint arXiv:2410.24164*, 2024.
- [12] P. Intelligence, K. Black, *et al.*, “ $\pi_0.5$: a vision-language-action model with open-world generalization,” *arXiv preprint arXiv:2504.16054*, 2025.
- [13] D. Qu, H. Song, *et al.*, “Spatialvla: Exploring spatial representations for visual-language-action model,” *arXiv preprint arXiv:2501.15830*, 2025.
- [14] C. Fan, X. Jia, *et al.*, “Interleave-vla: Enhancing robot manipulation with interleaved image-text instructions,” *arXiv preprint arXiv:2505.02152*, 2025.
- [15] Q. Bu, Y. Yang, *et al.*, “Univla: Learning to act anywhere with task-centric latent actions,” *arXiv preprint arXiv:2505.06111*, 2025.
- [16] B. Zitkovich, T. Yu, *et al.*, “Rt-2: Vision-language-action models transfer web knowledge to robotic control,” in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.
- [17] M. J. Kim, K. Pertsch, *et al.*, “Openvla: An open-source vision-language-action model,” *arXiv preprint arXiv:2406.09246*, 2024.
- [18] K. Black, N. Brown, *et al.*, “ π_0 : A vision-language-action flow model for general robot control,” *arXiv preprint arXiv:2410.24164*, 2024.
- [19] J. Bjorck, F. Castañeda, *et al.*, “Gr0ot n1: An open foundation model for generalist humanoid robots,” *arXiv preprint arXiv:2503.14734*, 2025.
- [20] K. Pertsch, K. Stachowicz, *et al.*, “Fast: Efficient action tokenization for vision-language-action models,” *arXiv preprint arXiv:2501.09747*, 2025.
- [21] A. O’Neill, A. Rehman, *et al.*, “Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6892–6903.
- [22] Q. Bu, J. Cai, *et al.*, “Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems,” *arXiv preprint arXiv:2503.06669*, 2025.
- [23] Y. Duan, M. Andrychowicz, *et al.*, “One-shot imitation learning,” *Advances in neural information processing systems*, vol. 30, 2017.
- [24] C. Finn, P. Abbeel, *et al.*, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.
- [25] T. Yu, C. Finn, *et al.*, “One-shot imitation from observing humans via domain-adaptive meta-learning,” *arXiv preprint arXiv:1802.01557*, 2018.
- [26] P. Sharma, D. Pathak, *et al.*, “Third-person visual imitation learning via decoupled hierarchical controller,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [27] E. Jang, A. Irpan, *et al.*, “Bc-z: Zero-shot task generalization with robotic imitation learning,” in *Conference on Robot Learning*. PMLR, 2022, pp. 991–1002.
- [28] M. Ahn, A. Brohan, *et al.*, “Do as i can, not as i say: Grounding language in robotic affordances,” *arXiv preprint arXiv:2204.01691*, 2022.
- [29] S. James, M. Bloesch, *et al.*, “Task-embedded control networks for few-shot imitation learning,” in *Conference on robot learning*. PMLR, 2018, pp. 783–795.
- [30] A. Bonardi, S. James, *et al.*, “Learning one-shot imitation from humans without humans,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3533–3539, 2020.
- [31] D. Pathak, P. Mahmoudieh, *et al.*, “Zero-shot visual imitation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 2050–2053.
- [32] S. Dasari and A. Gupta, “Transformers for one-shot visual imitation,” in *Conference on Robot Learning*. PMLR, 2021, pp. 2071–2084.
- [33] Z. Mandi, F. Liu, *et al.*, “Towards more generalizable one-shot visual imitation learning,” in *2022 International conference on robotics and automation (ICRA)*. IEEE, 2022, pp. 2434–2444.
- [34] C. Finn, T. Yu, *et al.*, “One-shot visual imitation learning via meta-learning,” in *Conference on robot learning*. PMLR, 2017, pp. 357–368.
- [35] M. Chang and S. Gupta, “One-shot visual imitation via attributed waypoints and demonstration augmentation,” *arXiv preprint arXiv:2302.04856*, 2023.
- [36] R. G. Goswami, P. Krishnamurthy, *et al.*, “Osvi-wm: One-shot visual imitation for unseen tasks using world-model-guided trajectory generation,” *arXiv preprint arXiv:2505.20425*, 2025.
- [37] J. Yang, C. Glossop, *et al.*, “Pushing the limits of cross-embodiment learning for manipulation and navigation,” 2024.
- [38] D. Ghosh, H. Walke, *et al.*, “Octo: An open-source generalist robot policy,” in *RSS*, 2024.
- [39] O. M. Team, D. Ghosh, *et al.*, “Octo: An open-source generalist robot policy,” 2024.
- [40] C. Wen, X. Lin, *et al.*, “Any-point trajectory modeling for policy learning,” *arXiv preprint arXiv:2401.00025*, 2023.
- [41] C. Yuan, C. Wen, *et al.*, “General flow as foundation affordance for scalable robot learning,” *arXiv preprint arXiv:2401.11439*, 2024.
- [42] C. Gao, H. Zhang, *et al.*, “Flip: Flow-centric generative planning as general-purpose manipulation world model,” 2024.
- [43] M. Xu, Z. Xu, *et al.*, “Flow as the cross-domain manipulation interface,” *arXiv preprint arXiv:2407.15208*, 2024.
- [44] N. Karaev, I. Makarov, *et al.*, “Cotracker3: Simpler and better point tracking...” *arXiv preprint*, 2024.
- [45] Nikita, I. Rocco, *et al.*, “Cotracker: It is better to track together,” in *ECCV*.
- [46] “Tapir: Tracking any point with per-frame initialization and...” in *ICCV*, 2023.
- [47] S. Ye, J. Jang, *et al.*, “Latent action pretraining from videos,” *arXiv preprint arXiv:2410.11758*, 2024.
- [48] Y. Chen, Y. Ge, *et al.*, “Moto: Latent motion token as the bridging language for robot manipulation,” *arXiv preprint arXiv:2412.04445*, vol. 8, 2024.
- [49] J. Bruce, M. D. Dennis, *et al.*, “Genie: Generative interactive environments,” in *Forty-first International Conference on Machine Learning*, 2024.
- [50] X. Chen, J. Guo, *et al.*, “Igor: Image-goal representations are

- the atomic control units for foundation models in embodied ai,” *arXiv preprint arXiv:2411.00785*, 2024.
- [51] M. Laskin, K. Lee, *et al.*, “Reinforcement learning with augmented data,” *Advances in neural information processing systems*, vol. 33, pp. 19 884–19 895, 2020.
- [52] D. Yarats, I. Kostrikov, *et al.*, “Image augmentation is all you need: Regularizing deep reinforcement learning from pixels,” in *International conference on learning representations*, 2021.
- [53] A. Mandlekar, D. Xu, *et al.*, “What matters in learning from offline human demonstrations for robot manipulation,” *arXiv preprint arXiv:2108.03298*, 2021.
- [54] A. Mandlekar, S. Nasiriany, *et al.*, “Mimicgen: A data generation system for scalable robot learning using human demonstrations,” *arXiv preprint arXiv:2310.17596*, 2023.
- [55] L. Fan, G. Wang, *et al.*, “Secant: Self-expert cloning for zero-shot generalization of visual policies,” *arXiv preprint arXiv:2106.09678*, 2021.
- [56] N. Hansen and X. Wang, “Generalization in reinforcement learning by soft data augmentation,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 611–13 617.
- [57] N. Hansen, H. Su, *et al.*, “Stabilizing deep q-learning with convnets and vision transformers under data augmentation,” *Advances in neural information processing systems*, vol. 34, pp. 3680–3693, 2021.
- [58] A. Mandlekar, S. Nasiriany, *et al.*, “Mimicgen: A data generation system for scalable robot learning using human demonstrations,” *arXiv preprint arXiv:2310.17596*, 2023.
- [59] T. Yu, T. Xiao, *et al.*, “Scaling robot learning with semantically imagined experience,” *arXiv preprint arXiv:2302.11550*, 2023.
- [60] Z. Chen, Z. Mandi, *et al.*, “Semantically controllable augmentations for generalizable robot learning,” *The International Journal of Robotics Research*, p. 02783649241273686, 2024.
- [61] Z. Chen, S. Kiami, *et al.*, “Genaug: Retargeting behaviors to unseen situations via generative augmentation,” *arXiv preprint arXiv:2302.06671*, 2023.
- [62] Z. Mandi, H. Bharadhwaj, *et al.*, “Cacti: A framework for scalable multi-task multi-scene visual imitation learning,” *arXiv preprint arXiv:2212.05711*, 2022.
- [63] L. Y. Chen, K. Hari, *et al.*, “Mirage: Cross-embodiment zero-shot policy transfer with cross-painting,” 2024.
- [64] S. Tian, B. Wulfe, *et al.*, “View-invariant policy learning via zero-shot novel view synthesis,” *arXiv preprint arXiv:2409.03685*, 2024.
- [65] E. Ameperosa, J. A. Collins, *et al.*, “Rocoda: Counterfactual data augmentation for data-efficient robot learning from demonstrations,” in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 13 250–13 256.
- [66] L. Y. Chen, C. Xu, *et al.*, “Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning,” *arXiv preprint arXiv:2409.03403*, 2024.
- [67] S. Yang, W. Yu, *et al.*, “Novel demonstration generation with gaussian splatting enables robust one-shot manipulation,” *arXiv preprint arXiv:2504.13175*, 2025.
- [68] M. Oquab, T. Dariset, *et al.*, “Dinov2: Learning robust visual features without supervision,” *arXiv preprint arXiv:2304.07193*, 2023.
- [69] A. Van Den Oord, O. Vinyals, *et al.*, “Neural discrete representation learning,” *Advances in neural information processing systems*, vol. 30, 2017.
- [70] E. J. Hu, Y. Shen, *et al.*, “Lora: Low-rank adaptation of large language models.” *ICLR*, vol. 1, no. 2, p. 3, 2022.
- [71] G. Pavlakos, D. Shan, *et al.*, “Reconstructing hands in 3d with transformers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9826–9836.
- [72] P. J. Besl and N. D. McKay, “Method for registration of 3-d shapes,” in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.
- [73] S. Rusinkiewicz and M. Levoy, “Efficient variants of the icp algorithm,” in *Proceedings third international conference on 3-D digital imaging and modeling*. IEEE, 2001, pp. 145–152.
- [74] G. Chen, M. Wang, *et al.*, “Fmimic: Foundation models are fine-grained action learners from human videos,” *The International Journal of Robotics Research*, p. 02783649251377335, 2025.
- [75] B. Wen, W. Yang, *et al.*, “Foundationpose: Unified 6d pose estimation and tracking of novel objects,” *arXiv preprint arXiv:2312.08344*, 2023.
- [76] J. Xiang, Z. Lv, *et al.*, “Structured 3d latents for scalable and versatile 3d generation,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21 469–21 480.
- [77] H.-S. Fang, C. Wang, *et al.*, “Anygrasp: Robust and efficient grasp perception in spatial and temporal domains,” *IEEE Transactions on Robotics*, 2023.
- [78] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: <http://arxiv.org/abs/1908.10084>
- [79] A. Brohan, N. Brown, *et al.*, “Rt-1: Robotics transformer for real-world control at scale,” *arXiv preprint arXiv:2212.06817*, 2022.
- [80] H. R. Walke, K. Black, *et al.*, “Bridgedata v2: A dataset for robot learning at scale,” in *Conference on Robot Learning*. PMLR, 2023, pp. 1723–1736.
- [81] A. Khazatsky, K. Pertsch, *et al.*, “Droid: A large-scale in-the-wild robot manipulation dataset,” *arXiv preprint arXiv:2403.12945*, 2024.
- [82] C. Lynch, A. Wahid, *et al.*, “Interactive language: Talking to robots in real time,” *IEEE Robotics and Automation Letters*, 2023.
- [83] J. Luo, C. Xu, *et al.*, “Fmb: a functional manipulation benchmark for generalizable robotic learning,” *The International Journal of Robotics Research*, vol. 44, no. 4, pp. 592–606, 2025.
- [84] R. Hoque, P. Huang, *et al.*, “Egodex: Learning dexterous manipulation from large-scale egocentric video,” *arXiv preprint arXiv:2505.11709*, 2025.
- [85] A. Padalkar, A. Pooley, *et al.*, “Open x-embodiment: Robotic learning datasets and rt-x models,” *arXiv preprint arXiv:2310.08864*, 2023.
- [86] K. Grauman, A. Westbury, *et al.*, “Ego4d: Around the world in 3,000 hours of egocentric video,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 995–19 012.
- [87] C. Chi, Z. Xu, *et al.*, “Diffusion policy: Visuomotor policy learning via action diffusion,” *The International Journal of Robotics Research*, p. 02783649241273668, 2023.
- [88] S. Karamcheti, S. Nair, *et al.*, “Prismatic vlms: Investigating the design space of visually-conditioned language models,” in *Forty-first International Conference on Machine Learning*, 2024.
- [89] B. Liu, Y. Zhu, *et al.*, “Libero: Benchmarking knowledge transfer for lifelong robot learning,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [90] S. Haddadin, S. Parusel, *et al.*, “The franka emika robot: A reference platform for robotics research and education,” *IEEE Robotics & Automation Magazine*, vol. 29, no. 2, pp. 46–64, 2022.