Vision-Language-Action Pretraining from Large-Scale Human Videos

Hao Luo [*1,2] **Yicheng Feng** [*1,2] **Wanpeng Zhang** [*1,2] **Sipeng Zheng** [*2] **Ye Wang** [3,2] **Haoqi Yuan** [1,2]
Jiazheng Liu [1] **Chaoyi Xu** [2] **Haiweng Xu** [4,2] **Qin Jin** [3] **Zongqing Lu** [†1,2]

Abstract

Existing Vision-Language-Action (VLA) models struggle with complex manipulation tasks requiring high dexterity and generalization, primarily due to their reliance on synthetic data with significant sim-to-real gaps or limited teleoperated demonstrations. To address this bottleneck, we propose leveraging human hands as a *manipulator template*, capitalizing on the rich dexterity and scalability present in web data of human manipulation. Our approach introduces *physical instruction tuning*, a novel training paradigm that combines large-scale VLA pretraining from human videos, perspective spatial alignment for reasoning in a unified physical space, and post-training adaptation in physical environments. Additionally, we introduce a part-level motion tokenization method that achieves millimeter-level reconstruction accuracy to model precise hand trajectories serving as scalable motion primitives. To support our paradigm, we develop a comprehensive data curation pipeline that integrates heterogeneous sources into a large-scale dataset with millions of motion-based instructional instances. Empirically, our model demonstrates superior performance in hand motion generation and instruction following, adhering to favorable scaling laws with respect to model and data sizes. Importantly, we demonstrate promising capabilities to robotic dexterous manipulation, validating the effectiveness of bridging the human-robot embodiment gap. Project page is available at https://research.beingbeyond.com/being-h0.

[*]Equal contribution [1]School of Computer Science, Peking University [2]BeingBeyond [3]School of Information, Remin University of China [4]School of Electronics Engineering and Computer Science, Peking University. Correspondence to: [†] Zongqing Lu <zongqing.lu@pku.edu.cn>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

The advance of ChatGPT and its successors have endowed large multimodal models (LMMs) with versatile capabilities across various domains, yet their application in robotics lags behind. Recent efforts aim to bridge this gap by adapting LMMs into Vision-Language-Action models (VLAs) (Kim et al., 2024; Black et al., 2024), harnessing their multimodal reasoning for robotic tasks. However, these models' generalization is severely limited by a reliance on small-scale, lab-collected teleoperated demonstrations (O'Neill et al., 2024; Khazatsky et al., 2024), which are orders of magnitude smaller than the internet-scale data used to train LMMs. Consequently, VLAs lack robustness across diverse objects and environments. This data scarcity is especially acute for dexterous manipulators, where operational complexity and high hardware costs have largely restricted VLAs to simple grippers (Cutler et al., 2024; An et al., 2025). Although simulators offer a path to low-cost synthetic data (Wan et al., 2023; Zhong et al., 2025), their limited diversity and persistent sim-to-real gap have thus hindered the successful deployment of dexterous hands.

Human videos offer a promising alternative for VLA training, providing abundant real-world data with a minimal reality gap. However, prior work relying on implicit learning techniques (e.g., contrastive learning (Nair et al., 2022), masked autoencoder (Radosavovic et al., 2023), or latent action (Bjorck et al., 2025)) has yielded unclear learning mechanisms and limited transfer effectiveness. These methods fall short of the dramatic performance gains achieved in LMMs, where techniques like visual instruction tuning (Liu et al., 2024a) have proven remarkable success. We argue this disparity stems from a fundamental difference in data structure. LMMs benefit from isomorphic data where pretraining and downstream tasks are aligned, as visual-text understanding directly translates to multimodal reasoning tasks. In contrast, VLAs face a heterogeneous alignment challenge, grappling with significant gaps between textual/2D visual inputs and the 3D physical action space, which includes critical proprioceptive requirements. Although recent explorations into explicit human-centric representation (Qiu et al., 2025) show promise, their limited scale contradicts the core vision of leveraging web-scale data — the very resource that enables LMMs' success through massive pretraining.

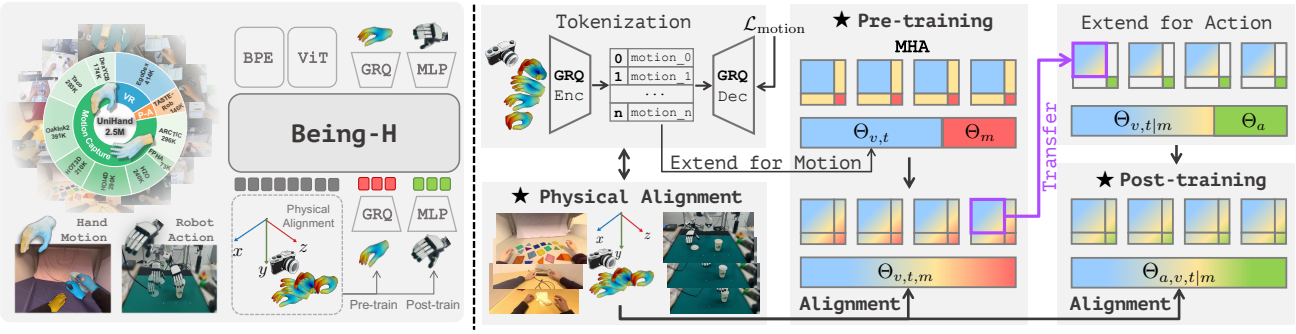

Figure 1. **Overview of Being-H pipeline (Left) and Physical Instruction Tuning (Right): (1)** Tokenize continuous hand motions into discrete representations. Physical alignment unifies heterogeneous hand motion and robot data via coordinate alignment. **(2)** Extend a vision-text model $\Theta_{v,t}$ with motion parameters Θ_m, enabling attention for vision and text, motion, and cross modality within a unified sequence. **(3)** The model leverages pretrained cross-modal dependencies ($\Theta_{v,t|m}$), then incorporates action parameters Θ_a via post-training to produce the final VLA $\Theta_{a,v,t|m}$ for downstream tasks[1]. The green part represents action attention.

These inspire us to ask: *Can we pretrain a dexterous VLA from large-scale human videos to explicitly imitate human actions and adapt to robot manipulators via post-training?*

Our motivation is straightforward: the human hand is the template for dexterous manipulation (Kim et al., 2021), exhibiting unparalleled versatility in natural settings. Learning from human motion offers a powerful pathway to bridge the pretraining-downstream data heterogeneity in robotics, but scaling to massive video data presents new challenges: (1) Data Heterogeneity. Human videos span varying camera systems, coordinate frames, and recording conditions, complicating model learning, which require unifying disparate sources and embedding essential 3D spatial reasoning capabilities. (2) Hand Quantization. To preserve granular control, continuous, fine-grained movements must be discretized into language-compatible tokens without losing millimeter-level precision. (3) Robot Control Transfer. Considering morphological differences, human hand motions require careful skill transfer to ensure learned strategies adapt effectively to end-effectors.

To address these challenges, we propose an advanced yet sample-efficient dexterous VLA trained on large-scale human videos. To train this VLA, we introduce a training framework named ***Physical Instruction Tuning*** (Figure 1) by extending visual instruction tuning to the physical domain via: (1) VLA pretraining on human videos, (2) perspective spatial alignment to unify heterogeneous data from diverse camera systems and recording conditions for reasoning in the physical space, and (3) post-training adaptation to ground pretrained priors in physical environments. Unlike implicit methods (Bjorck et al., 2025; Nair et al., 2022), we use explicit hand motion prior to guide robot learning. Our VLA employs a unified autoregressive architecture with shared attention for seamless cross-modal reasoning. For precise motion tokenization, we introduce a part-level tokenizer with grouped residual quantization (Yang et al., 2023) that achieves millimeter-level accuracy. To support large-scale learning, we curate ***UniHand***, a comprehensive 150M dataset integrating motion capture, VR, and RGB videos across diverse manipulation tasks. To our knowledge, this is the first VLA model for dexterous hands using explicit motion modeling from large-scale human videos.

Our key contributions are: **(1) Physical Instruction Tuning.** A novel paradigm that establishes the human hand as a template prior for robot control, bridging human videos to embodied action. **(2) Part-Level Motion Tokenization.** A quantization method that achieves millimeter-level precision for continuous motions while ensuring compatibility with discrete, autoregressive models. **(3) The UniHand Dataset.** A large-scale dataset of over 150M instruction-following samples from diverse manipulation scenarios, unified via a scalable pipeline integrating motion capture, VR, and RGB videos. **(4) Being-H.** Integrating these innovations, we develop a VLA model trained on large-scale videos of human-**Being H**ands. Our model enables robust cross-modal reasoning across vision, language, and motion, with effective adaptation for downstream robot tasks.

Conflict of Interest Disclosure. Several authors are employed by BeingBeyond, which leads the development of the system and technologies presented and evaluated in this paper. This relationship is disclosed in accordance with the conference policy on financial conflicts of interest.

2. Related Work

Hand Motion Modeling. Existing research on human hand motion mainly focuses on hand-object interaction (HOI)(Jiang et al., 2021) and fine-grained action precision. Recent benchmarks(Banerjee et al., 2025; Zhan et al., 2024) capture these interactions, but their reliance on motion-capture systems or multi-camera setups con-

[1]The condition on m in $\Theta_{v,t|m}$ indicates that the weights are influenced by pretraining on hand motions. Similarly, we denote the post-trained weights as $\Theta_{a,v,t|m}$.

fines them to tabletop scenarios. Egocentric videos (Grauman et al., 2024) from head-mounted cameras offer richer environments but often lack precise 3D annotations. Recent monocular 3D hand modeling (Pavlakos et al., 2024; Dong et al., 2024) enables pseudo-label extraction, yet the weak-perspective assumption is incompatible with shifted-perspective egocentric data (Grauman et al., 2022). Integrated with SLAM, Dyn-HaMR (Yu et al., 2025) advances in camera tracking and occlusion-robust refinement. Beyond hands modeling, HOI-centric methods (Liu et al., 2022a) predict interaction hotspots, future trajectories, and affordances. Meanwhile, HOI has progressed from 2D recognition/detection (Gkioxari et al., 2015; 2018; Qi et al., 2018) to 3D motion generation: earlier works use multi-stage pipelines (Christen et al., 2024; Cha et al., 2024) from action labels (Ghosh et al., 2023; Brahmbhatt et al., 2019), while recent diffusion models (Ho et al., 2020) improve generation and autoregressive LLMs (Huang et al., 2025) in long-term consistency. However, most methods overlook visual inputs until MEgoHand (Zhou et al., 2025). Inspired by these advances, we pretrain our model on large-scale human videos with generation objectives to learn strong hand-motion priors for downstream manipulation.

Learning VLAs from Human Videos. The emergence of LMMs has enabled VLAs to map perceptions to physical actions. Most approaches (Kim et al., 2024; Black et al., 2024; Bjorck et al., 2025) improve generalization via large-scale robot-data pretraining (O'Neill et al., 2024; Khazatsky et al., 2024). FAST (Pertsch et al., 2025) accelerates autoregressive training with discrete cosine transforms, while Octo (Team et al., 2024) and RDT-1B (Liu et al., 2024b) adopt diffusion heads for more flexible action prediction. Yet dexterous hand datasets remain small due to the high cost of data collection compared to grippers. Simulation scales better (Ye et al., 2025; Deng et al., 2025) but still suffers from the sim-to-real gap. Human videos are a promising alternative, providing transferable representations such as visual features (Nair et al., 2022), 3D priors (Xu et al., 2025), and interaction knowledge (e.g., affordance, contacts, and grasps)(Gavryushin et al., 2025; Chen et al., 2025). However, existing methods still differ substantially from our setting. One line of work reduces the visual gap by editing videos, such as masking human hands or rendering robot embodiments into human demonstrations (Lepert et al., 2025; Kareer et al., 2024). Another line aligns human and robot action spaces through retargeting, human-centric state-action representations, or physical trajectory refinement (Niu et al., 2025; Yang et al., 2025; Qiu et al., 2025; Li et al., 2025b; Chen et al., 2024d). These approaches are effective for transferring demonstrations, but limited by specific embodiments and retargeting quality. Recent hand modeling methods also study grasp prediction or physically plausible HOI generation (He et al., 2025; Zhang et al., 2026); however, they primarily target grasp or HOI sequence generation rather than scalable pretraining for downstream control. More recent VLA pretraining methods, such as VITRA (Li et al., 2026), learn compact representations from videos, but mainly rely on visual dynamics rather than explicit hand-motion supervision. Related latent-action methods, such as UniVLA (Bu et al., 2025) and IGOR (Chen et al., 2024c), compress visual transitions into task-centric latent codes; in contrast, our motion tokens are not used as task-specific latent actions during post-training, but as structured physical supervision for learning transferable manipulation priors. Being-H therefore treats human hands as a universal template for downstream manipulators: it explicitly tokenizes part-level hand motion from large-scale human videos, aligns heterogeneous viewpoints through perspective spatial alignment, and maps the resulting motion-informed VLA representation to robot actions during post-training. To our knowledge, this is the first work to pretrain a scalable, generalizable VLA by explicitly modeling motion from large-scale human videos, enabling robots to learn diverse skills.

3. Physical Instruction Tuning

We propose ***physical instruction tuning*** for our dexterous VLA, Being-H, with three key components in Figure 2.

3.1. Pretraining on Hand Motion

We leverage human-robot anatomical similarity by pretraining on hand motion generation, treating the human hand as an ideal manipulator template and robots as simplified versions. Our pretrained VLA learns to predict MANO-parameterized motions $m = \{\theta, \mathbf{r}_{rot}, \tau, \beta\}$ (joint angles θ, wrist rotation $\mathbf{r}_{rot}$, translation τ, and shape β) from visual-text context. Similar to the multimodal instruction tuning in VLMs, each sample is treated as an instructional pair $\{\mathcal{X}_Q, \mathcal{X}_A\}$ and the objective is:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(\Theta) = -\sum_{j=1}^{L} \log P_{\Theta}(y_j \mid \mathcal{X}_Q, y_{1:j-1}), \quad (1)$$

where $\mathcal{X}_Q$ is the question inputs, $\mathcal{X}_A = \{y_{1:L}\}$ refers to the target answer tokens. Building on LMMs like InternVL3 (Chen et al., 2024e), the model incorporates motion tokens into the vocabulary, enabling autoregressive generation of hand motion sequences conditioned on visual-text inputs. Due to space limitation, we introduce multimodal integration and training details in Appendix B.1 and B.3.

Inspired by Wang et al. (2024c), we treat hand movements as a foreign language by quantizing continuous motion features into discrete embeddings using a motion tokenizer during pretraining. The tokenizer encodes T-frame hand features $\mathcal{M} = \{m_1, ..., m_T\}$ of raw motion sequence into

$\lceil T/\alpha \rceil$ d-dim token embeddings, where α denotes the temporal downsampling rate. To represent hand efficiently and effectively, we use the MANO-162 as the hand motion parameterization. Each frame is encoded as $m \in \mathbb{R}^{162}$, including joint pose $\theta \in \mathbb{R}^{15\times6}$, global rotation $\mathbf{r}_{rot} \in \mathbb{R}^6$, and translation $\tau \in \mathbb{R}^3$. Both θ and $\mathbf{r}_{rot}$ are in the axis–angle form. The LMM vocabulary is extended by integrating K discrete motion codes and special tokens <MOT> and </MOT> to mark motion block boundaries.

Hand Motion Tokenization. The precision of the motion tokenizer critically impacts both the quality of generated hand motions and the transferability of learned priors to downstream tasks. We therefore design a dedicated tokenizer based on GRQ (Yang et al., 2023) for expressive motion representation (Figure 2). Given a motion sequence $\mathcal{M} \in \mathbb{R}^{T\times D}$, an encoder converts it into a feature map $z \in \mathbb{R}^{\lceil T/\alpha \rceil \times d}$, which is discretized via a multi-stage residual quantization. First, the channel dimension d is partitioned into n groups. Each group feature $z^{(g)} \in \mathbb{R}^{\lceil T/\alpha \rceil \times d/n}$ is quantized independently using an L-layer residual vector quantizer (RVQ) with codebook $\mathcal{C}^{(g)}$. Each feature vector $z_i^{(g)} \in \mathbb{R}^{d/n}$ in group g is quantized as:

$$r_0 = z_i^{(g)}, \quad q_l = \underset{c\in\mathcal{C}^{(g)}}{\arg\min} \|r_{l-1} - c\|_2, \quad r_l = r_{l-1} - q_l, \qquad (2)$$

where q_l and r_l denote the selected code and residual at layer l. The final quantized representation is: $\hat{z}_i^{(g)} = \sum_{l=1}^{L} q_l$. We observe that reconstructing wrist parameters $\mathbf{r}_{rot}$ and τ is challenging due to broad 3D distribution, despite their importance for precision. To address this, we introduce a wrist-specific loss term: $\mathcal{L}_{\text{wrist}} = ||\mathbf{w} - \hat{\mathbf{w}}||_2^2$ where $\mathbf{w} = [\mathbf{r}_{rot}, \tau]$ and $\hat{\mathbf{w}}$ denotes the reconstructed wrist parameters. Combined with the reconstruction and quantization loss from VQ-VAE (Van Den Oord et al., 2017), the final objective is:

$$\mathcal{L}_{\text{motion}} = \mathcal{L}_{\text{recon}} + \lambda_1 \mathcal{L}_{\text{quant}} + \lambda_2 \mathcal{L}_{\text{wrist}}, \qquad (3)$$

where $\lambda_{\{1,2\}}$ denotes balancing weights. We employ an exponential moving average (EMA) update strategy for the codebook. Given the higher complexity of wrist motions, we employ separate tokenizers for wrist $\{\mathbf{r}_{rot}, \tau\}$ (global pose) and finger $\{\theta, \beta\}$ (fine-grained manipulation) parameters. This part-level strategy improves feature modeling, provides explicit token semantics, and enhances the LMM's capture of structured hand dynamics. When using separate tokenizers, $\mathcal{L}_{\text{wrist}}$ is omitted. We discuss motion feature choices and tokenization details in Appendix B.2 and B.3.

Decoding Mode. Our pretrained VLA generates both text and motion via unified next-token prediction, with motion tokens decoded to MANO parameters. We introduce three decoding modes to balance flexibility and motion validity: **(1) Free-format Mode** which allows fully flexible autoregressive sampling without any constraints and risks producing invalid motion blocks. **(2) Block-formatted Mode** which ensures structural consistency by restricting sampling to motion tokens between <MOT> and </MOT> delimiters. For evaluation, <EOS> is suppressed until the target number of motion blocks is generated. **(3) Soft-formatted Mode** which focuses on evaluating local motion quality by using a soft constraint: we blend predicted and ground-truth MANO parameters via their mean, then tokenize the hybrid for the prediction of the next block. This anchors the generation in a plausible neighborhood of real trajectories, providing a reliable measure of the model's ability to produce high-quality motion in the vicinity of real trajectories.

3.2. Perspective Spatial Alignment

Diversity of human videos introduces significant variability in camera systems, complicating effective pretraining. To alleviate this, we introduce ***perspective spatial alignment***: a unified processing toolkit that maps videos from disparate cameras into a consistent physical coordinate system. This pipeline incorporates 3D spatial reasoning and available physical attributes to enhance geometric and perceptual consistency across datasets. We adopt two alignment strategies:

Weak-Perspective Projection Alignment. Inconsistent camera systems across datasets cause divergent 3D projections, impairing model depth perception and 3D reasoning. To alleviate this, we align all data to a unified weak-perspective camera space, ensuring consistent 2D-to-3D mapping and uniform scaling for objects at similar depths. Given source camera intrinsics $\{f_x, f_y, c_x, c_y\}$ and target $\{f'_x, f'_y, c'_x, c'_y\}$, scale factors and translation offsets are:

$$s_x = \frac{f'_x}{f_x}, \; s_y = \frac{f'_y}{f_y}, \; \Delta x = c'_x - s_x \cdot c_x, \; \Delta y = c'_y - s_y \cdot c_y. \qquad (4)$$

Each pixel (u, v) in the source image is transformed as $u' = s_x \cdot u + \Delta x$ and $v' = s_y \cdot v + \Delta y$, with cropping or padding to a target resolution.

View-Invariant Motion Distribution Balancing. To prevent camera bias and ensure robust 3D generalization, we introduce this strategy that augments video-motion pairs from underrepresented sources. Our strategy varies hand pose distribution without introducing camera viewpoint and position changes. Unlike image-level augmentations like random cropping or flipping, it preserves weak-perspective consistency between hand motion and visual observations, ensuring coherent 3D understanding. Our strategy employs two complementary components: **(1) Depth Scaling.** For a hand pose in camera coordinate $m_c = \{\beta, \theta, R_c, \tau_c\}$, where τ_c denotes the wrist's 3D position and $R_c \in \mathbb{R}^{3\times3}$ is the rotation matrix, we perturb human hand's depth by randomly scaling depth $\tau_c^{z'}$ by λ_s, yielding $\tau_c^{z'} = \lambda_s \cdot \tau_c^z$. The paired image is rescaled by $1/\lambda_s$ to maintain weak-perspective consistency. λ_s is constrained to plausible ranges to avoid unrealistic perspective distortions caused by the non-negligible

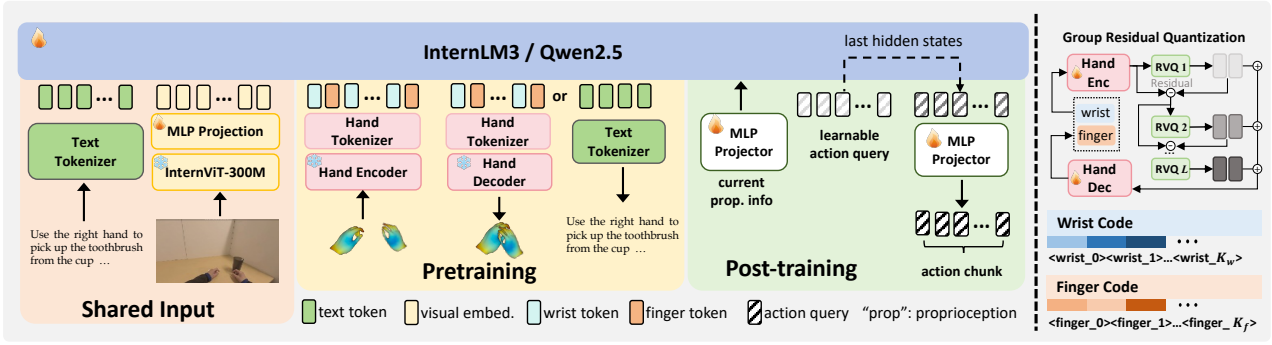

Figure 2. **Model Overview.** The text tokenizer and visual encoder are shared by both pretraining and post-training. For pretraining, a hand motion tokenizer based on GRQ is included for hand motion inputs and autoregressive outputs. For post-training and downstream tasks, Being-H incorporates a set of learnable queries, whose hidden states are converted into continuous action chunks.

physical size of the hand. **(2) In-Plane Rotation.** To improve hand position diversity in the image plane, we rotate the hand around the camera's Z-axis by a uniformly sampled angle φ, updating wrist position $\tau_c' = R_z(\varphi) \cdot \tau_c$ and global rotation $R_c{}' = R_z(\varphi) \cdot R_c$, where $R_z(\varphi)$ is the rotation transform matrix.

The image is rotated synchronously by φ, and all transformed frames are resized to a target resolution to preserve weak-perspective projection integrity. While our approach focuses on perspective alignment, its underlying motivation extends to a broader notion of *physical space alignment*, aiming to unify visual observations with richer physical cues, additionally discussed in Appendix C.

3.3. Post-training for Dexterous Manipulation

Learned rich behavior priors from hand motion pretraining, Being-H requires adaptation to bridge the human-robot kinematic gap. To better address this gap, the post-training stage serves as the point where pre-trained high-level human priors are grounded into the embodiments. For a general downstream embodiment support, we do not directly utilize hand motion generation. Instead, we map the hand motion latent to the robot action space, allowing the model to adapt high-level priors to low-level robot actuation. In this paper, we employ a **non-autoregressive** projection (Figure 2) for manipulation post-training and inference.

Added to the visual-text sequence, a chunk of learnable queries $(\mathbf{q}_1, \ldots, \mathbf{q}_{N_a})$ attend to the instructional context through the pre-trained backbone. The final hidden states of the action queries serve as action features that retain pretraining human priors and are regressed to executable controls by an MLP head f_r.

Additionally, a lightweight MLP f_p maps the robot's proprioceptive states into the VLA's embedding space for assistance. Formally, the conbined unified context is $[\texttt{visual}, \texttt{text}, f_p(\mathbf{p}_t), \mathbf{q}_{1:N_a}]$. Denote the queried output of $\mathbf{q}_i$ as h_i. We conduct imitation learning between predicted actions a and expert demonstrations $\mathbf{a}^* = \{\mathbf{a}_i^*\}$:

$$\mathbf{a}_i = f_r(h_i), \quad \mathcal{L}_{\text{imitate}} = \frac{1}{N_a} \sum_{i=1}^{N_a} \|\mathbf{a}_i - \mathbf{a}_i^*\|_1, \tag{5}$$

where N_a is the action chunk size. This approach adapts the VLA to generate executable robot controls while preserving its cross-modal capabilities. The post-training only requires the pretrained VLA backbone and two MLP projectors. We adopt this simple MLP-based projector to provide a clean evaluation of the pretrained VLA. In fact, stronger adapters could further improve downstream performance, and an alternative is discussed in Appendix B.4.

4. UniHand: Scaling Hand-Motion Instruction

To support explicit VLA pretraining from large-scale human videos, we curate ***UniHand***, a dataset aggregated from 11 sources with detailed hand motion annotations and RGB video. It contains over 440K task trajectories (130M frames, 1,100+ hours), ensuring high diversity and coverage of real-world scenarios. Due to computational constraints, we sample 2.5M instruction data points via a balanced strategy to preserve task and source diversity, which we refer to as *UniHand-2.5M* (Figure 3).

We adopt several key steps to form the curation pipeline: **(1) Hand Pose Standardization.** Our method standardizes all hand motions as MANO parameters to eliminate camera system variance and learns an explicit 2D-to-3D mapping. **(2) Task Description Labeling.** We annotate UniHand via a hierarchical labeling framework with chunk-level and per-second labels to enrich sparse texts and strengthen vision-language-motion alignment. **(3) Instructional Data Generation.** We create diverse task types to construct instruction-following data for our VLA pretraining. Each task type contains 20 base templates, further augmented by LMMs. We then use rule-based instantiation to populate these templates with grounded instructions, motion tokens, and length constraints. More curation details and dataset statistics are posted in Appendix D.1 and D.2.

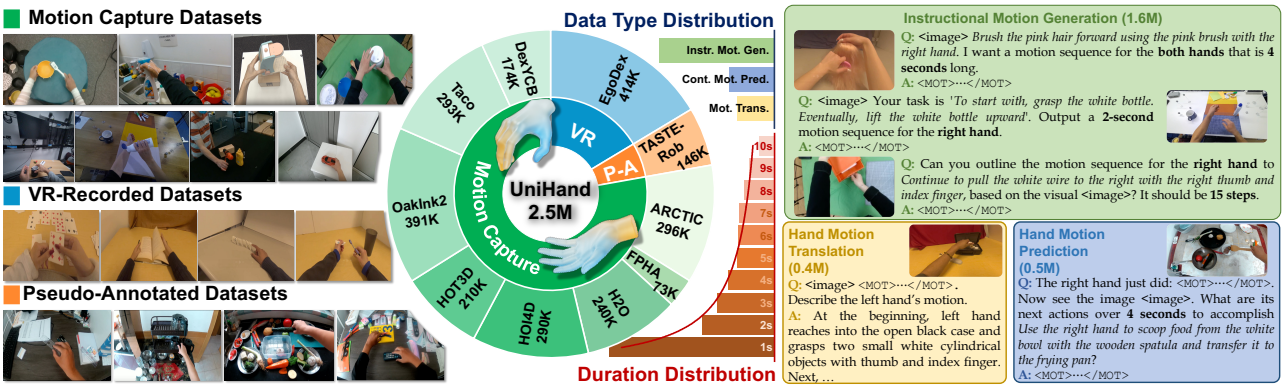

Figure 3. UniHand-2.5M Overview. **(left)** Scenes and tasks across motion-capture, VR-recorded, and pseudo-annotated data. **(middle)** Distributions of sources, data types, and clip durations. **(right)** Example samples from each data type.

5. Experiments

5.1. Experimental Setup

Implementation Details. We encode motion sequences with a temporal downsampling $\alpha = 4$. Our part-level tokenizer uses an 8-layer GRQ architecture with a group size $n = 2$, converting each one-second motion sequence into $2nL\lceil T/\alpha \rceil = 128$ tokens. Codebook sizes are K_w=K_f=4096 per part shared across all RQ layers with code dimension of $d = 512$. The tokenizer is optimized with loss weights $\lambda_1 = 0.02, \lambda_2 = 1.0$, batch size 2048, and learning rate 1×10^{-4}. For multimodal modeling, we employ *InternVL3* (1B/8B/14B) backbones trained on *UniHand-2.5M*. Each instance includes a 448×448 scene image and a camera-coordinate-aligned hand motion. Hand poses are represented using MANO-D162, sampled at 15 FPS and discretized into 128 tokens per hand per second. During pre-training, we use AdamW with a learning rate 1×10^{-5}, batch size 128, and training on 32 × A800-80G GPUs, jointly fine-tuning both ViT adapter and LLM backbone. In addition, we also include a variant noted as 'Being-H (FM)' where the MLP action head is replaced with the flow-matching head used in GR00T (Bjorck et al., 2025). Please refer to the Appendix B.4 for architectural details.

Evaluation Benchmarks. We evaluate Being-H on two kinds of benchmarks. Below, we describe the datasets, tasks, and metrics with all details provided in Appendix E.

Hand Motion Modeling: We sample 5% data of UniHand for three sub tasks: **(1) Generation** which produces a 3D hand motion sequence from a static scene image, text instruction, and duration; **(2) Prediction** which forecasts subsequent motion given an image, a short motion context, and a follow-up text instruction; **(3) Translation** which generates a text description from an image and a motion sequence. We evaluate on two splits "**head**"(EgoDex) and "**tail**"(other datasets, which constitute a sparse long-tail like TACO, HOI4D, and H2O). We assess the spatial accuracy and semantic alignment for motion generation using MPJPE, MWTE, PA-MPJPE, M2T R@3, and FID. For hand motion translation, we adopt T2M R@3 for evaluation.

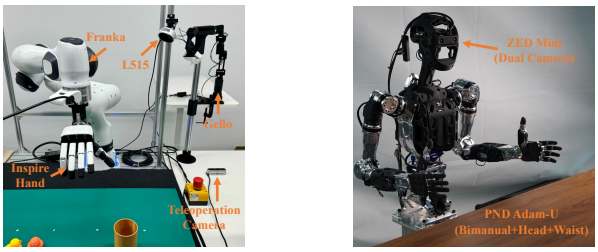

Figure 4. Hardware: stable view (Left) and mobile stereo (Right).

Dexterous Manipulation: We evaluate how the pretrained priors transfer in the downstream dexterous manipulation through simulation benchmarks and real-world tasks. Simulation benchmarks include LIBERO (Liu et al., 2023a) and RoboCasa (Nasiriany et al., 2024). In LIBERO, models are fine-tuned separately on four task suites using 50 demonstrations per task with two camera views. In RoboCasa, models are fine-tuned on all 24 atomic tasks using 50 demonstrations per task from a single left-camera view. Real-world experiments use the hardware setup in Figure 4. Under the stable single-view setting (left), we evaluate grasping and placing (*Pick-Place-Toy*), articulated-object manipulation (*Close-Toolbox*, *Close-Lid*), deformable-object manipulation (*Unfold-Clothes*), and precise motion control (*Pour-Cup*) over 20 randomized trials with binary success. We additionally include *Spray–Plant* to probe contact-rich dexterity (narrow-neck grasp, multi-finger stabilization, trigger actuation), reporting both success and a three-stage completion score. We provide 50 demonstrations for all tasks, except *Pick-Place-Toy* with 100 to cover four toy colors. We further introduce a human-like mobile stereo-view setting (right) with moving binocular input and extra head/waist DoF, which makes viewpoint coordination substantially harder. We evaluate *Bimanual-Handover* and *Arrange-Flower* with 222 and 485 demonstrations.

5.2. Comparisons on Hand Motion Modeling

Table 1. Comparison on motion generation and translation, where we use valid rate (%) and T2M R@3 (%) as the metrics.

Model	Valid Rate (%)	T2M R@3 (Head) ↑	T2M R@3 (Tail) ↑
ground truth	-	33.5	42.7
Being-H-1B	64.8	12.5	14.3
Being-H-8B	99.8	18.4	19.7
Being-H-14B	100.0	19.0	22.1

To begin with, we set a one-second prediction horizon. We first examine the ability to produce properly formatted motion sequences. As Table 1 shows, experiments on motion generation based on free-format decoding mode reveal substantial differences in valid generation rates across model scales. While Being-H-1B achieves only modest success in preserving motion block structure, Being-H-8B and -14B reach almost 100% validity, demonstrating that increased scale significantly improves structural format learning. Table 1 also evaluates motion understanding through hand motion translation. Results show larger models consistently achieve higher retrieval scores, confirming stronger bidirectional motion-language alignment.

Table 2 reports principal results on hand motion generation and prediction. We adapt GR00T N1.5 (Bjorck et al., 2025) as a competitive baseline following (Zhou et al., 2025), redefining its action representation as dual-hand motion sequences (padding with zeros for single-hand cases). All models use block-formatted decoding mode for consistent evaluation, enforcing generated sequences to be decoded into the correct format. Our analysis reveals three key findings. First, larger models show superior performance with lower MPJPE, MWTE, and PA-MPJPE scores, indicating enhanced spatial grounding and more plausible pose generation. Second, they achieve better M2T R@3 and FID results, showing stronger semantic consistency between generated motions and instructions. Notably, the performance advantage is particularly pronounced on the tail split, suggesting scaling substantially improves generalization across diverse motion distributions. We further evaluate long-term motion generation capabilities in Appendix F.1.

5.3. Comparison on Dexterous Manipulation

Simulation Experiments. We evaluate Being-H in both LIBERO and RoboCasa benchmarks to assess downstream transfer in gripper-based tasks. On LIBERO, Being-H surpasses prior VLA and imitation-learning policy baselines across all four task suites, achieving a 90.3% average success rate, while the stronger FM variant reaches 94.5% (Table 3). On RoboCasa, as shown in Table 4, Being-H outperforms GR00T N1.5 in all categories and achieves a higher average success rate (23.8% vs. 21.0%) despite GR00T being pretrained on extensive robot data. The model

Table 2. Comparison of hand motion generation and prediction tasks on both head and tail splits.

Model	MPJPE ↓		MWTE ↓		PA-MPJPE ↓		M2T R@3 ↑		FID ↓	
	head	tail	head	tail	head	tail	head	tail	head	tail
# Hand Motion Generation										
GR00T N1.5	9.82	15.35	8.51	11.20	1.33	1.41	13.1	14.8	11.7	14.4
Being-H-1B	9.71	17.21	8.25	12.04	1.50	1.55	12.1	15.3	12.2	13.1
Being-H-8B	7.20	9.02	5.69	8.11	1.09	1.32	15.9	18.7	11.5	13.4
Being-H-14B	6.87	8.11	5.19	7.41	1.03	1.20	17.2	20.5	10.3	11.8
# Hand Motion Prediction										
GR00T N1.5	7.14	8.55	6.65	7.93	1.11	1.25	16.1	20.5	11.2	13.3
Being-H-1B	8.73	11.34	7.88	10.67	1.17	1.38	15.4	20.6	14.7	15.6
Being-H-8B	6.67	7.98	5.03	6.93	0.90	1.03	19.7	21.4	10.1	11.7
Being-H-14B	6.21	7.33	4.89	6.52	0.92	1.04	20.1	23.5	9.8	10.1

Table 3. Results of Being-H and baselines on LIBERO manipulation tasks. We report success rates (%) across task categories and the overall average.

Model	Spatial	Object	Goal	Long	Avg.
Diffusion Policy (Chi et al., 2025)	78.3	92.5	68.3	50.5	72.4
Octo (Team et al., 2024)	78.9	85.7	84.6	51.1	75.1
OpenVLA (Kim et al., 2024)	84.7	88.4	79.2	53.7	76.5
π0-FAST (Pertsch et al., 2025)	**96.4**	96.8	88.6	60.2	85.5
GR00T N1.5 (Bjorck et al., 2025)	92.0	86.0	92.0	76.0	86.5
MolmoAct (Lee et al., 2025)	87.0	95.4	87.6	77.2	86.6
Being-H	92.6	96.8	94.4	77.4	90.3
Being-H(FM)	95.2	**97.0**	**97.8**	**87.8**	**94.5**

shows superior fine-grained manipulation, especially in high-precision tasks like *Insert* and *Button*. Compared to InternVL3, the backbone without our pretraining, Being-H achieves a 5.6% absolute gain, demonstrating a direct performance gain of our pretraining. These results demonstrate that large-scale pretraining with human videos provides robust and transferable priors for downstream tasks. We further validate that scaling post-training data on RoboCasa leads to additional performance gains (Appendix F.4).

Comparison with human-video VLA pretraining. To further compare with prior human-video VLA pretraining methods, we evaluate VITRA, the closest related baseline, on LIBERO. Since VITRA is not originally designed for the LIBERO gripper embodiment, we follow the same adaptation protocol and re-initialize a new action head on top of the pretrained VITRA backbone. As shown in Table 5, Being-H substantially outperforms VITRA under the same benchmark, and the flow-matching variant further improves the results. These results suggest that explicit hand-motion supervision, part-level motion tokenization, and action-query-based adaptation provide a more effective interface for transferring human-video priors to downstream robot control.

Stable View Experiments. In general stable view real-world experiments (Table 6), Being-H achieves the highest success rates across all benchmarks. In contrast, the *InternVL3* baseline, which lacks physical instruction tuning and hand motion priors, exhibits significantly weaker performance. While the finetuned GR00T N1.5 performs comparably on in-domain objects for the *Pick-Place-Toy* task,

Table 4. Results of Being-H and baselines on RoboCasa manipulation tasks. We report the success rate (%) across 7 task categories and the overall average on 24 tasks. *PnP.* for *Pick and Place*.

Task	PnP.	Door	Drawer	Lever	Knob	Insert	Button	Avg.
GR00T N1.5	1.3	40.5	37.0	48.0	**11.0**	6.0	26.7	21.0
InternVL3	1.3	37.0	36.0	42.0	9.0	4.0	18.0	18.2
Being-H	**2.0**	**43.5**	**40.0**	**51.3**	**11.0**	**17.0**	**30.7**	**23.8**

Table 5. Comparison with VITRA on LIBERO. Being-H achieves stronger downstream performance under the same adaptation setting, and the flow-matching action head further improves performance.

Model	Spatial	Object	Goal	Long	Avg.
VITRA	78.2	85.4	83.6	59.6	76.7
Being-H	92.6	96.8	94.4	77.4	90.3
Being-H (FM)	95.2	97.0	97.8	87.8	94.5

Table 6. Success rates (%) and Being-H vs. baselines on real-world dexterous manipulation tasks.

Task	Pick-Place-Toy			Close Toolbox	Close Lid	Pour Cup	Unfold Clothes
	Seen.	*Unseen.*	*Clutter.*				
GR00T N1.5	**0.75**	0.40	0.50	0.80	0.50	0.90	0.60
InternVL3	0.55	0.55	0.50	0.50	0.25	0.55	0.45
Being-H	**0.75**	**0.65**	**0.60**	**0.85**	**0.60**	**1.00**	**0.75**

Table 7. Results on the contact-rich *Spray–Plant* task. We report average completion score (0–1) and binary success rate (%).

Metric	GR00T N1.5	InternVL3	Being-H	Being-H (FM)
Completion	0.33	0.23	0.58	**0.63**
Success (%)	15.0	5.0	35.0	**40.0**

its generalization degrades markedly with unseen objects and cluttered scenes. Being-H's explicit motion tokenization enables superior generalization with far less data than GR00T's implicit latent action prediction. This advantage is most evident in fine-grained manipulation. For instance, our model robustly positions and closes lids, pinches cloth edges to unfold fabric, and maintains a stable grasp for smooth pouring. These results underscore Being-H's successful transfer of hand motion knowledge from physical instruction tuning to downstream real-world robot control.

To evaluate data efficiency, we further compare our model against a non-pretrained InternVL3 baseline using 25%, 50%, and 100% of demonstration data across multiple tasks. As shown in Figure 5, Being-H maintains a consistent and substantial performance advantage at all data scales, demonstrating the benefit of physical instruction tuning. The hand-motion priors from pretraining enable faster adaptation, allowing our model to match or exceed baseline performance with far less data. For example, with only 25% data, our model performs comparably to the baseline trained using 100% data on *Pick-Place-Toy*, and matches the baseline at 50% on *Close-Toolbox* and *Unfold-Clothes*. In the challenging *Close-Lid*, Being-H achieves a 15% success rate with 25% data, while the baseline fails entirely. This superior data efficiency reduces the reliance on costly teleoperated data, lowering deployment barriers for dexterous robots.

Spray-Plant is substantially more challenging than standard pick-and-place, requiring multi-finger role allocation, stable grasp under torque, and precise trigger actuation. As shown in Table 7, Being-H achieves higher completion (0.58) and success (35%) than GR00T N1.5 and InternVL3, with the largest margin in binary success, suggesting more coordinated grasp-and-trigger action and reliable contact stabilization rather than early failure. Moreover, Being-H (FM) further improves performance (0.63 / 40%), consistent with our LIBERO findings, highlighting that human-pretrained priors benefit not only high-level planning but also stable multi-finger control from limited teleoperated data.

Mobile Stereo-View Experiments. We evaluate a harder mobile stereo-view setup with ego-motion, frequent viewpoint shifts/occlusions, and increased demands on coordination. The tasks are also intrinsically challenging: *Arrange-Flower* requires precise spatial perception and state awareness to decide when to release, while *Bimanual-Handover* demands tight bimanual coordination and accurate control during transfer. Compared with $\pi_{0.5}$ (Intelligence et al., 2025) (Table 8), Being-H achieves higher success on both tasks, highlighting that the explicit learning motion-aware priors from human videos benefit temporally consistent, contact-sensitive control under dynamic viewpoints.

5.4. Ablation Study

Training Data Scale Ablation. In Figure 6, performance improves steadily with training data up to 2.5M, demonstrating the value of scaling diverse motion-language data. Noting that pose accuracy (PA-MPJPE) slightly drops when using 100% data, semantic alignment metrics (e.g., M2T R@3) continue to rise, We hypothesize that larger data volume increases diversity in task-object combinations and motion semantics, encouraging a shift toward prioritizing semantic plausibility over precise kinetics replication of finger pose detail. This reflects the model's growing emphasis on functional and contextual correctness as training data becomes abundant. We discuss more ablation about tokenizer designs and data configuration in Appendix F.2 and F.3.

5.5. Qualitative Examples

To qualitatively demonstrate the capability of Being-H in generating physically plausible hand motions and its performance in real-robot experiments, we present representative samples in Figure 7 (More in Appendix F.5).

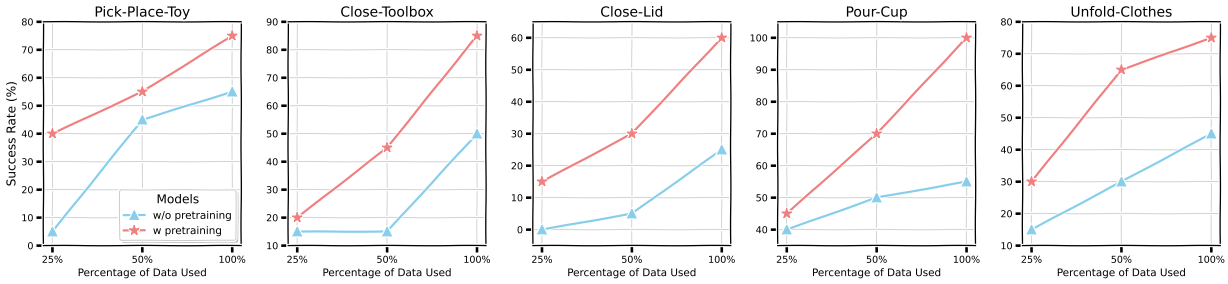

Figure 5. Data Efficiency Comparison. Success rates of Being-H and InternVL3 when fine-tuned with varying data fractions (25%-100%).

Table 8. Sucess rates(%) on mobile stereo-view real-world tasks.

Task	$\pi_{0.5}$	Being-H	Being-H (FM)
Bimanual-Handover	35.0	50.0	**60.0**
Arrange-Flower	15.0	45.0	**50.0**

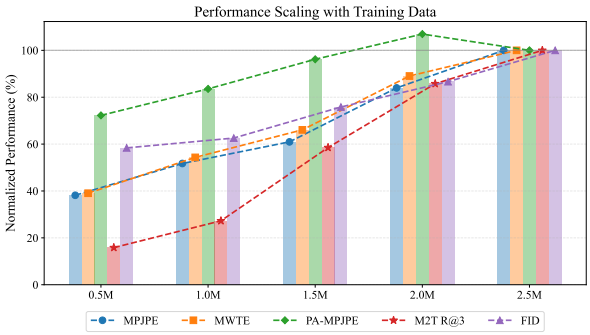

Figure 6. Being-H-8B performance scaling with training data on hand motion generation. Metrics are normalized to the final checkpoint (2.5M training samples= 100%).

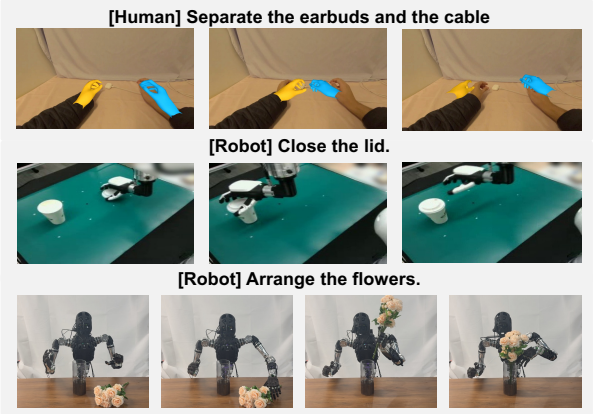

Figure 7. Examples of Being-H for generating realistic hand motions and performing dexterous real-robot tasks.

6. Conclusion

We introduce Being-H, a scalable dexterous VLA trained via physical instruction tuning. To support VLA pretraining, we curate a large-scale dataset named UniHand by integrating heterogeneous sources (motion capture, VR, RGB videos) via MANO parameter standardization. We adopt grouped residual quantization for millimeter-level accuracy and seamless integration with language models, effectively treating motion as a language. Using the human hand as a template, we transfer dexterity from human videos to robot control via mapping the VLA latent to the downstream action space, eliminating the pretraining-downstream data mismatch common in previous VLAs and achieving superior performance across downstream benchmarks. Extensive experiments show that Being-H benefits from both data and model scaling, and consistently improves data efficiency in downstream robot learning. We hope this work provides a practical step toward leveraging abundant human videos as scalable physical priors for generalizable manipulation.

Limitations. Despite the promising results, Being-H still has limitations. UniHand includes pseudo-labeled RGB videos, whose annotations may be noisy in depth-sensitive or contact-rich scenarios, and our current training uses a view-balanced subset rather than the full data pool. In addition, perspective spatial alignment assumes relatively stable viewpoints within short temporal chunks, which may limit robustness to abrupt ego-motion. Our real-world evaluation covers multiple dexterous and bimanual settings, but still remains limited relative to open-world manipulation.

Impact Statement

This paper presents work whose goal is to advance vision-language-action learning for robotic manipulation. We do not expect the work, in its current form, to have immediate broad societal impact beyond those commonly associated with existing robot learning methods. Standard concerns regarding safe deployment in physical environments apply, and any real-world use should follow appropriate safety validation and human oversight practices.

Acknowledgments

This work was supported by NSFC in part under Grant 62450001 and 62476008. The authors would like to thank the anonymous reviewers for their valuable comments and advice.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

An, S., Meng, Z., Tang, C., Zhou, Y., Liu, T., Ding, F., Zhang, S., Mu, Y., Song, R., Zhang, W., et al. Dexterous manipulation through imitation learning: A survey. *arXiv preprint arXiv:2504.03515*, 2025.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Han, S., Zhang, F., Zhang, L., Fountain, J., Miller, E., Basol, S., et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7061–7071, 2025.

Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al. $\pi 0$: A vision-language-action flow model for general robot control. corr, abs/2410.24164, 2024. doi: 10.48550. *arXiv preprint ARXIV.2410.24164*, 2024.

Brahmbhatt, S., Ham, C., Kemp, C. C., and Hays, J. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8709–8719, 2019.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Bu, Q., Yang, Y., Cai, J., Gao, S., Ren, G., Yao, M., Luo, P., and Li, H. Univla: Learning to act anywhere with task-centric latent actions. In *Robotics: Science and Systems*, 2025.

Bukschat, Y. and Vetter, M. Efficientpose: An efficient, accurate and scalable end-to-end 6d multi object pose estimation approach. *arXiv preprint arXiv:2011.04307*, 2020.

Cai, Z., Yin, W., Zeng, A., Wei, C., Sun, Q., Yanjun, W., Pang, H. E., Mei, H., Zhang, M., Zhang, L., et al. Smplerx: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023.

Cha, J., Kim, J., Yoon, J. S., and Baek, S. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1577–1585, 2024.

Chao, Y.-W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y. S., Van Wyk, K., Iqbal, U., Birchfield, S., et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9044–9053, 2021.

Chen, G. H., Chen, S., Zhang, R., Chen, J., Wu, X., Zhang, Z., Chen, Z., Li, J., Wan, X., and Wang, B. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024a.

Chen, H., Sun, B., Zhang, A., Pollefeys, M., and Leutenegger, S. Vidbot: Learning generalizable 3d actions from in-the-wild 2d human videos for zero-shot robotic manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27661–27672, 2025.

Chen, L.-H., Lu, S., Zeng, A., Zhang, H., Wang, B., Zhang, R., and Zhang, L. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024b.

Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., and Yu, G. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18000–18010, 2023.

Chen, X., Guo, J., He, T., Zhang, C., Zhang, P., Yang, D. C., Zhao, L., and Bian, J. Igor: Image-goal representations are the atomic control units for foundation model in embodied ai. *arXiv preprint arXiv:2411.00785*, 2024c.

Chen, Z., Chen, S., Arlaud, E., Laptev, I., and Schmid, C. Vividex: Learning vision-based dexterous manipulation from human videos. *arXiv preprint arXiv:2404.15709*, 2024d.

Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024e.

Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., and Song, S. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11): 1684–1704, 2025.

Christen, S., Hampali, S., Sener, F., Remelli, E., Hodan, T., Sauser, E., Ma, S., and Tekin, B. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Cutler, E., Xing, Y., Cui, T., Zhou, B., van Rijnsoever, K., Hart, B., Valencia, D., Ong, L. V. C., Gee, T., Liarokapis, M., et al. Benchmarking reinforcement learning methods for dexterous robotic manipulation with a three-fingered gripper. *arXiv preprint arXiv:2408.14747*, 2024.

Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J. S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv e-prints*, pp. arXiv–2409, 2024.

Deng, S., Yan, M., Wei, S., Ma, H., Yang, Y., Chen, J., Zhang, Z., Yang, T., Zhang, X., Cui, H., et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*, 2025.

Dong, H., Chharia, A., Gou, W., Vicente Carrasco, F., and De la Torre, F. D. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. *Advances in Neural Information Processing Systems*, 37: 2127–2160, 2024.

Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M. J., and Hilliges, O. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12943–12954, 2023.

Garcia-Hernando, G., Yuan, S., Baek, S., and Kim, T.-K. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 409–419, 2018.

Gavryushin, A., Wang, X., Malate, R. J., Yang, C., Jia, X., Goel, S., Liconti, D., Zurbrügg, R., Katzschmann, R. K., and Pollefeys, M. Maple: Encoding dexterous robotic manipulation priors learned from egocentric videos. *arXiv preprint arXiv:2504.06084*, 2025.

Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., and Slusallek, P. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, volume 42, pp. 1–12. Wiley Online Library, 2023.

Gkioxari, G., Girshick, R., and Malik, J. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1080–1088, 2015.

Gkioxari, G., Girshick, R., Dollár, P., and He, K. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8359–8367, 2018.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., 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*, pp. 18995–19012, 2022.

Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Guo, C., Mu, Y., Javed, M. G., Wang, S., and Cheng, L. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024.

He, J., Li, D., Yu, X., Qi, Z., Zhang, W., Chen, J., Zhang, Z., Zhang, Z., Yi, L., and Wang, H. Dexvlg: Dexterous vision-language-grasp model at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hoque, R., Huang, P., Yoon, D. J., Sivapurapu, M., and Zhang, J. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

Huang, M., Chu, F.-J., Tekin, B., Liang, K. J., Ma, H., Wang, W., Chen, X., Gleize, P., Xue, H., Lyu, S., et al. Hoigpt: Learning long-sequence hand-object interaction with language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7136–7146, 2025.

Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., and Chen, T. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36: 20067–20079, 2023.

Jiang, H., Liu, S., Wang, J., and Wang, X. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11107–11116, 2021.

Kareer, S., Patel, D., Punamiya, R., Mathur, P., Cheng, S., Wang, C., Hoffman, J., and Xu, D. Egomimic: Scaling imitation learning via egocentric video. *arXiv preprint arXiv:2410.24221*, 2024.

Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M. K., Chen, L. Y., Ellis, K., et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Kim, U., Jung, D., Jeong, H., Park, J., Jung, H.-M., Cheong, J., Choi, H. R., Do, H., and Park, C. Integrated linkage-driven dexterous anthropomorphic robotic hand. *Nature communications*, 12(1):7177, 2021.

Kwon, T., Tekin, B., Stühmer, J., Bogo, F., and Pollefeys, M. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10138–10148, 2021.

Lee, D., Kim, C., Kim, S., Cho, M., and Han, W.-S. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022.

Lee, J., Duan, J., Fang, H., Deng, Y., Liu, S., Li, B., Fang, B., Zhang, J., Wang, Y. R., Lee, S., et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.

Lepert, M., Fang, J., and Bohg, J. Phantom: Training robots without robots using only human videos. *arXiv preprint arXiv:2503.00779*, 2025.

Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Cahyono, J. A., Yang, J., Li, C., and Liu, Z. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Li, K., Li, P., Liu, T., Li, Y., and Huang, S. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6991–7003, 2025b.

Li, Q., Deng, Y., Liang, Y., Luo, L., Zhou, L., Yao, C., Zeng, L., Feng, Z., Liang, H., Xu, S., Zhang, Y., Chen, X., Chen, H., Sun, L., Chen, D., Yang, J., and Guo, B. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. In *International Conference on Robotics and Automation*, 2026.

Li, Z., Yuan, W., He, Y., Qiu, L., Zhu, S., Gu, X., Shen, W., Dong, Y., Dong, Z., and Yang, L. T. Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. *arXiv preprint arXiv:2410.07093*, 2024.

Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., and Stone, P. Libero: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Liu, J., Zheng, S., Karlsson, B. F., and Lu, Z. Taking notes brings focus? towards multi-turn multimodal dialogue learning. *arXiv preprint arXiv:2503.07002*, 2025.

Liu, S., Tripathi, S., Majumdar, S., and Wang, X. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3282–3292, 2022a.

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

Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., and Yi, L. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21013–21022, 2022b.

Liu, Y., Yang, H., Si, X., Liu, L., Li, Z., Zhang, Y., Liu, Y., and Yi, L. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21740–21751, 2024c.

Lu, S., Chen, L.-H., Zeng, A., Lin, J., Zhang, R., Zhang, L., and Shum, H.-Y. Humantomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978*, 2023a.

Lu, Y., Li, C., Liu, H., Yang, J., Gao, J., and Shen, Y. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*, 2023b.

Mahendran, S., Ali, H., and Vidal, R. 3d pose regression using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 2174–2182, 2017.

Mentzer, F., Minnen, D., Agustsson, E., and Tschannen, M. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023.

Nair, S., Rajeswaran, A., Kumar, V., Finn, C., and Gupta, A. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

Nasiriany, S., Maddukuri, A., Zhang, L., Parikh, A., Lo, A., Joshi, A., Mandlekar, A., and Zhu, Y. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

Niu, Y., Zhang, Y., Yu, M., Lin, C., Li, C., Wang, Y., Yang, Y., Yu, W., Zhang, T., Chen, B., et al. Human2locoman: Learning versatile quadrupedal manipulation with human pretraining. *arXiv preprint arXiv:2506.16475*, 2025.

O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., 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)*, pp. 6892–6903. IEEE, 2024.

Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., and Malik, J. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9826–9836, 2024.

Perrett, T., Darkhalil, A., Sinha, S., Emara, O., Pollard, S., Parida, K. K., Liu, K., Gatti, P., Bansal, S., Flanagan, K., et al. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23901–23913, 2025.

Pertsch, K., Stachowicz, K., Ichter, B., Driess, D., Nair, S., Vuong, Q., Mees, O., Finn, C., and Levine, S. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

Petrovich, M., Black, M. J., and Varol, G. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *International Conference on Computer Vision (ICCV)*, 2023.

Qi, S., Wang, W., Jia, B., Shen, J., and Zhu, S.-C. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 401–417, 2018.

Qin, Y., Yang, W., Huang, B., Van Wyk, K., Su, H., Wang, X., Chao, Y.-W., and Fox, D. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. *arXiv preprint arXiv:2307.04577*, 2023.

Qiu, R.-Z., Yang, S., Cheng, X., Chawla, C., Li, J., He, T., Yan, G., Yoon, D. J., Hoque, R., Paulsen, L., et al. Humanoid policy˜ human policy. *arXiv preprint arXiv:2503.13441*, 2025.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. In *arxiv*. San Francisco, CA, USA, 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Radosavovic, I., Xiao, T., James, S., Abbeel, P., Malik, J., and Darrell, T. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pp. 416–426. PMLR, 2023.

Romero, J., Tzionas, D., and Black, M. J. Embodied hands: Modeling and capturing hands and bodies together. *ACM*

Transactions on Graphics, (Proc. SIGGRAPH Asia), 36 (6), November 2017.

Team, C. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Team, O. M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., and Bermano, A. H. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

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

Wan, W., Geng, H., Liu, Y., Shan, Z., Yang, Y., Yi, L., and Wang, H. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F. V., et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20270–20281, 2023.

Wang, Y., Huang, D., Zhang, Y., Ouyang, W., Jiao, J., Feng, X., Zhou, Y., Wan, P., Tang, S., and Xu, D. Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. *arXiv preprint arXiv:2410.21747*, 2024b.

Wang, Y., Zheng, S., Cao, B., Wei, Q., Zeng, W., Jin, Q., and Lu, Z. Scaling large motion models with million-level human motions. In *Forty-second International Conference on Machine Learning*, 2024c.

Wu, P., Shentu, Y., Yi, Z., Lin, X., and Abbeel, P. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12156–12163. IEEE, 2024.

Xu, B., Mei, Y., Liu, X., Zheng, S., and Jin, Q. Egodtm: Towards 3d-aware egocentric video-language pretraining. *arXiv preprint arXiv:2503.15470*, 2025.

Yang, D., Liu, S., Huang, R., Tian, J., Weng, C., and Zou, Y. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*, 2023.

Yang, R., Yu, Q., Wu, Y., Yan, R., Li, B., Cheng, A.-C., Zou, X., Fang, Y., Yin, H., Liu, S., Han, S., Lu, Y., and Wang, X. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

Ye, J., Wang, K., Yuan, C., Yang, R., Li, Y., Zhu, J., Qin, Y., Zou, X., and Wang, X. Dex1b: Learning with 1b demonstrations for dexterous manipulation. *arXiv preprint arXiv:2506.17198*, 2025.

You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.-F., and Yang, Y. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

You, T., Kim, S., Kim, C., Lee, D., and Han, B. Locally hierarchical auto-regressive modeling for image generation. *Advances in Neural Information Processing Systems*, 35: 16360–16372, 2022.

Yu, L., Lezama, J., Gundavarapu, N. B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Birodkar, V., Gupta, A., Gu, X., et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

Yu, Z., Zafeiriou, S., and Birdal, T. Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27716–27726, 2025.

Yuan, H., Bai, Y., Fu, Y., Zhou, B., Feng, Y., Xu, X., Zhan, Y., Karlsson, B. F., and Lu, Z. Being-0: A humanoid robotic agent with vision-language models and modular skills. *arXiv preprint arXiv:2503.12533*, 2025.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Zhan, X., Yang, L., Zhao, Y., Mao, K., Xu, H., Lin, Z., Li, K., and Lu, C. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 445–456, 2024.

Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., and Shan, Y. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14730–14740, 2023.

Zhang, W., Feng, Y., Luo, H., Li, Y., Yue, Z., Zheng, S., and Lu, Z. Unified multimodal understanding via byte-pair visual encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a.

Zhang, W., Xie, Z., Feng, Y., Li, Y., Xing, X., Zheng, S., and Lu, Z. From pixels to tokens: Byte-pair encoding on quantized visual modalities. In *International Conference on Learning Representations*, 2025b.

Zhang, Z., Liu, J., Shi, Y., and Wang, J. UniHM: Unified dexterous hand manipulation with vision language model. In *International Conference on Learning Representations*, 2026.

Zhao, H., Liu, X., Xu, M., Hao, Y., Chen, W., and Han, X. Taste-rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27683–27693, 2025.

Zheng, S., Zhou, B., Feng, Y., Wang, Y., and Lu, Z. Unicode: Learning a unified codebook for multimodal large language models. *arXiv preprint arXiv:2403.09072*, 2024.

Zhong, Y., Huang, X., Li, R., Zhang, C., Liang, Y., Yang, Y., and Chen, Y. Dexgraspvla: A vision-language-action framework towards general dexterous grasping. *arXiv preprint arXiv:2502.20900*, 2025.

Zhou, B., Zhan, Y., Zhang, Z., and Lu, Z. Megohand: Multimodal egocentric hand-object interaction motion generation. *arXiv preprint arXiv:2505.16602*, 2025.

Zhou, Z., Wan, Y., and Wang, B. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1357–1366, 2024.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Appendix

Roadmap. In this section, we first introduce additional references related to this work in Section A. Then, we provide additional discussion of our pretraining and physical space alignment stage in Section B and Section C, respectively. We further provide the details of dataset statistics (Section D and evaluation setups (Section E). Finally, we carry out a thorough experiments and corresponding analysis in Section F.

Table of Contents

A. Additional Related Work

This section covers related work omitted from the main text due to space constraints.

Large Multimodal Models. The transformer architecture (Vaswani et al., 2017) revolutionized language modeling (Radford et al., 2018; 2019; Brown et al., 2020), enabling powerful autoregressive text interpretation and generation. This success has extended to large multimodal models (LMMs) (Zhu et al., 2023; Wang et al., 2024a; Zheng et al., 2024; Zhang et al., 2025b;a), which combine LLM reasoning (Touvron et al., 2023a;b; Bai et al., 2023) with modality-specialized encoders (Radford et al., 2021; Zhai et al., 2023) for unified multimodal understanding. Pioneering works like Flamingo (Alayrac et al., 2022) use cross-attention for strong few-shot VQA performance. Subsequent approaches, such as the LLaVA series (Liu et al., 2024a; 2023b; Lu et al., 2023b), employee visual instruction tuning with curated datasets to enhance instruction-following capabilities. These datasets are often constructed by using vision models to label images, then generating QA pairs with an LLM (You et al., 2023; Li et al., 2025a), or by leveraging proprietary LMMs for annotation (Chen et al., 2024a; Liu et al., 2025). While leading LMMs (Team et al., 2023; Comanici et al., 2025; Achiam et al., 2023) remain closed-source, recent models show growing openness through released weights (Bai et al., 2025; Team, 2024), training details (Zhu et al., 2025), and data recipes (Deitke et al., 2024).

Human Motion Quantization. Autoregressive models, which excel in long-term motion modeling and textual reasoning, commonly rely on motion quantization to represent continuous motion as discrete tokens. Widely used techniques include VQ-VAE (Van Den Oord et al., 2017; Zhang et al., 2023), Residual Quantization (RQ) (Lee et al., 2022; Guo et al., 2024), Hierarchical Quantization (H2VQ) (You et al., 2022; Lu et al., 2023a), and more recent lookup-free methods (Mentzer et al., 2023; Wang et al., 2024c; Yu et al., 2023). By fine-tuning LLMs on these tokenized motions, such approaches (Wang et al., 2024b; Jiang et al., 2023) achieve strong performance in intention understanding and motion generation (Chen et al., 2024b; Zhou et al., 2024; Li et al., 2024). This paradigm contrasts with diffusion-based models, which focus on high-fidelity motion synthesis (Tevet et al., 2022; Chen et al., 2023).

B. Additional Details of Pretraining

B.1. Multimodal Integration

Like traditional LLMs, Being-H employs next-token prediction for generation during pretraining, unifying three modalities — RGB vision, text, and hand motion — by tokenizing each into discrete tokens. While text processing follows standard LLM practices, we detail the vision and motion tokenization below:

Vision Token. To handle variable-resolution images and dynamic content, visual inputs undergo adaptive processing. Given an input image, we first apply dynamic patching, generating N patches based on image content complexity. Following InternVL (Zhu et al., 2025), a thumbnail I_{thumb} (downsampled with a pixel-shuffle ratio of 0.5) is retained alongside the detailed patches to preserve global context. Features are extracted from patches and the thumbnail via a vision encoder, then projected into a unified embedding space using a MLP. Vision tokens are structured with boundary markers *<IMG>* and *</IMG>*, while *<IMG_CONTEXT>* acts as a placeholder dynamically replaced by actual visual embeddings during processing.

Motion Token. Motion data is quantized prior to integration. Given a motion feature sequence $\mathcal{M}$, the motion tokenizer discretizes it into a sequence of tokens $\{m_i\}$, structured with boundary tokens *<MOT>* and *</MOT>*. Each motion block contains of 128 tokens per second, ensuring motion information is clearly delineated within the token stream while maintaining compatibility with the transformer architecture.

Multimodal Fusion. All modalities are processed in a unified token space using shared embeddings and attention. During fusion, vision tokens replace *<IMG_CONTEXT>* placeholders, while motion tokens are inserted as structured blocks within the text sequence, forming a combined token sequence $\mathbf{S} = \{s_i\}$ where each element s_i may represent text, visual, or motion content. Cross-modal attention is applied simultaneously across all modalities. For the concatenated multimodal hidden states $\mathbf{H}_{v,t,m} = [\mathbf{H}_v; \mathbf{H}_t; \mathbf{H}_m]$ (representing vision, text, and motion embeddings), we compute query, key and value through shared projections:

$$\mathbf{Q}_{v,t,m} = \mathbf{W}_Q \mathbf{H}_{v,t,m}, \quad \mathbf{K}_{v,t,m} = \mathbf{W}_K \mathbf{H}_{v,t,m}, \quad \mathbf{V}_{v,t,m} = \mathbf{W}_V \mathbf{H}_{v,t,m} \tag{6}$$

where $\mathbf{W}_{\{Q,K,V\}}$ denotes the weight matrices. This design enables direct cross-modal attention, capturing rich interdepen-

dencies between modalities, such as associating visual observations with specific hand motions, or grounding language instructions in corresponding movement sequences.

As shown in Figure 1, our pretraining extends the original vision-text parameters $\Theta_{v,t}$ to include motion parameters Θ_m, facilitating unified multimodal processing via shared attention. The model thereby learns to generate coherent motion tokens conditioned on visual and linguistic context.

B.2. Hand Motion Tokenization

As we introduced, we represent hand pose using the 3D model MANO (Romero et al., 2017), parameterized as $m = \{\theta, \mathbf{r}_{rot}, \tau, \beta\}$. An effective and efficient representation of this pose is critical for modeling motion. This paper explores five alternative feature spaces derived from the MANO parameters:

- **MANO-D51**: A 51-dimensional vector $m \in \mathbb{R}^{51}$, comprising axis-angle rotations for joints $\theta \in \mathbb{R}^{15\times3}$, global rotation $\mathbf{r}_{rot} \in \mathbb{R}^3$ and and translation $\tau \in \mathbb{R}^3$.
- **MANO-D99**: A 99-dimensional vector $m \in \mathbb{R}^{99}$ which replaces the axis-angle rotations in MANO-D51 with more robust 6D rotations: $\theta \in \mathbb{R}^{15\times6}$ and $\mathbf{r}_{rot} \in \mathbb{R}^6$.
- **MANO-D109**: This 109-dimensional representation extends MANO-D99 by incorporating the shape parameters $\beta \in \mathbb{R}^{10}$.
- **MANO-D114**: This 114-dimensional representation extends MANO-D51 by adding 3D joint positions $j \in \mathbb{R}^{21\times3}$. The joint positions serve only as auxiliary features during training; at evaluation and inference, only the core 51 parameters are used.
- **MANO-D162**: This 162-dimensional representation extends MANO-D99 by adding 3D joint positions $j \in \mathbb{R}^{21\times3}$.

Our experiments reveal that 6D rotation features yield superior reconstruction quality for finger joints, while axis-angle features are more effective for the wrist. We attribute this to the distinct structural characteristics of different hand parts. The wrist exhibits larger but simpler rotations, where the compactness and computational efficiency of axis-angle formulations are advantageous (Bukschat & Vetter, 2020; Mahendran et al., 2017). In contrast, the finer, more complex motions of finger joints are better captured by the continuity and numerical stability of the 6D representation. Although the axis-angle features achieve a lower overall reconstruction error due to the dominant scale of wrist pose errors, we select the 6D rotation feature for our hand motion tokenizer based on its superior performance in Being-H training. We hypothesize that that wrist pose patterns are relatively easier for the LMM to learn, whereas accurately modeling fine-grained finger movements presents a greater challenge. Consequently, we adopt the MANO-D162 as the feature for hand motion in this work.

B.3. Training Details

Motion Tokenizer Training.

Given a hand motion sequence $m \in \mathbb{R}^{T\times D}$ represented using the MANO-D162 features, the motion tokenizer encodes each one-second window into a feature map $z \in \mathbb{R}^{T/\alpha\times d}$, followed by a multi-stage residual vector quantization (RVQ) process described in Section 3.1. For each group $g \in \{1, \ldots, n\}$, the L-stage RVQ produces quantized codes $\hat{z}_i^{(g)} = \sum_{\ell=1}^{L} q_\ell^{(g)}$ with residual updates $r_\ell^{(g)} = r_{\ell-1}^{(g)} - q_\ell^{(g)}$, where $r_0^{(g)} = z^{(g)}$ and $q_\ell^{(g)} = \mathrm{argmin}_{c\in C^{(g)}} \|r_{\ell-1}^{(g)} - c\|_2$. To train the tokenizer, we minimize a combination of reconstruction, commitment, and wrist-specific losses:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_1 \mathcal{L}_{\text{quant}} + \lambda_2 \mathcal{L}_{\text{wrist}}. \tag{7}$$

- **Reconstruction Loss.** Let $\hat{m}$ denote the decoded motion sequence obtained from the quantized codes. The reconstruction loss encourages accurate recovery of the continuous MANO parameters:

$$\mathcal{L}_{\text{recon}} = \|m - \hat{m}\|_2^2. \tag{8}$$

- **Quantization Loss.** Following standard VQ-VAE training, we quantize the motion features with a commitment loss to stabilize the codebook usage:

$$\mathcal{L}_{\text{quant}} = \frac{1}{n \cdot L} \sum_{l,g} \Big(\big\| \mathrm{sg}(r_l^{(g)}) - q_l^{(g)} \big\|_2^2 + \beta \big\| r_l^{(g)} - \mathrm{sg}(q_l^{(g)}) \big\|_2^2 \Big), \tag{9}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator.

- **Wrist Loss.** As wrist orientation and translation exhibit broader spatial variation than finger joints, we introduce an additional wrist loss that focuses explicitly on the global hand pose:

$$\mathcal{L}_{\text{wrist}} = \left\| w - \hat{w} \right\|_2^2, \tag{10}$$

where $w = [r_{\text{rot}}, \tau]$ denotes the wrist parameters and $\hat{w}$ are the corresponding decoded estimates. This term improves the numerical stability of 6D rotations and yields more reliable trajectory structure for downstream manipulation.

- **Codebook Optimization.** Each RVQ codebook $C^{(g)}$ is updated using an exponential moving average (EMA) strategy to ensure stable cluster assignments. Following prior GRQ-based tokenizers, we constrain the update magnitude to avoid codebook collapse and encourage balanced token utilization.

- **MANO Feature Preprocess.** We sample hand motion sequences at 15 FPS and tokenize them using fixed one-second windows. Since camera shift may occur within these windows and Being-H does not predict camera motion during inference, we transform each sequence into the coordinate system of its first frame. To support coherent longer-sequence generation, where each one-second segment within a multi-second output must be relative to the entire sequence's initial frame, we employ a specialized training strategy: for each one-second sample, we randomly select a reference frame from a larger 10-second window and transform the motion relative to it. This enables motion tokens to represent movements relative to varying world coordinate systems while maintaining long-term consistency.

VLA Pretraining. The model is trained using standard next-token prediction. To optimize the integrated motion codes, we introduce a dual-level masking strategy that operates at both the vocabulary and token levels:

- **Vocabulary-level Logit Masking.** Since motion codes $\mathcal{V}_{\text{motion}}$ constitute a small subset of the full vocabulary $\mathcal{V}$, we mask non-motion logits for motion labels with probability $\mathcal{P}$. This focuses gradient updates on the motion embedding space and prevents dilution by irrelevant tokens. For predicted logits $\mathbf{z} \in \mathbb{R}^{|\mathcal{V}|}$, we apply masking as:

$$\tilde{\mathbf{z}}_i = \begin{cases} \mathbf{z}_i & i \in \mathcal{V}_{\text{motion}} \\ -\infty & \text{otherwise.} \end{cases} \quad \text{(with probability } \mathcal{P}\text{)}. \tag{11}$$

- **Token-level Loss Masking.** The token-wise cross-entropy losses are computed using the masked logits $\tilde{\mathbf{z}}$. To handle variations in motion complexity (e.g., static poses vs. unpredictable jitters), we filter extreme loss values, focusing learning on moderately challenging tokens. For per-token losses $L = \{\ell_1, \ldots, \ell_N\}$, the filtered loss set as:

$$\tilde{L} = \{\ell_i \in L \mid Q_{\text{low}} \leq \ell_i \leq Q_{\text{high}}\} \tag{12}$$

where Q_{low}, Q_{high} are preset percentile thresholds. The final motion loss is the mean over the filtered losses:

$$\mathcal{L}_{\text{motion}} = \frac{1}{|\tilde{L}|} \sum_{\ell_i \in \tilde{L}} \ell_i \tag{13}$$

In practice, vocabulary-level logit masking with $\mathcal{P} = 50\%$ and token-level loss filtering within $[15\%, 95\%]$ are applied.

B.4. Downstream Adaptation Alternative

While the main paper adopts a lightweight MLP projection head f_r to map the VLA's action-query embeddings to robot actions, our framework is not restricted to this simple adapter. In settings where the embodiment gap is larger or the downstream motion dynamics are more complex, a more expressive adapter may further improve performance. For this reason, we additionally explore a *flow-matching action head* as an alternative to the MLP head during post-training.

Flow-Matching Head. Given the final hidden states of the action queries $\{h_i\}_{i=1}^{N_a}$ from the pretrained VLA backbone, we condition a diffusion-style network V on the embeddings together with the robot's proprioceptive state p_t. For a ground-truth action chunk A_t, a noised version is constructed as

$$A_t^\tau = \tau A_t + (1 - \tau)\, \epsilon, \tag{14}$$

where $\tau \in [0, 1]$ is a noise level and $\epsilon \sim \mathcal{N}(0, I)$. The flow-matching objective supervises the network to predict the denoising vector field:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\tau, A_t, \epsilon}\Big[\big\|V(\{h_i\}, A_t^\tau, p_t) - (\epsilon - A_t)\big\|_2^2\Big]. \tag{15}$$

During inference, a randomly initialized action chunk is iteratively denoised via forward Euler integration using V_θ, resulting in executable robot commands. This procedure injects high-level behavioral priors encoded in the VLA backbone into the generation of low-level robot actions.

Relationship to the MLP Adapter. The flow-matching head preserves the same pretrained VLA backbone and the action-query interface. **The only change lies in how the final action embeddings $\{h_i\}$ are mapped to continuous robot controls, which is orthogonal to our whole pipeline.** Compared with the lightweight MLP adapter, the flow-matching head introduces several conceptual advantages. (1) It offers *greater representational expressiveness*, as the diffusion-based denoising formulation can model complex, multi-modal action distributions that a simple regression head may fail to capture. (2) The iterative refinement process offers *richer temporal and force-dependent modeling*, which is particularly beneficial for tasks that involve fine-grained interaction, continuous contact modulation, or larger embodiment discrepancies between human and robot. (3) The GR00T-style flow-matching head naturally carries *significantly more parameters* than a simple MLP projector, further strengthening its ability to fit complex action manifolds and transfer the pre-trained embeddings. These benefits come at the cost of increased computational overhead and a more involved inference procedure, but they provide a practical avenue for improving downstream performance when higher-capacity adaptation is desired.

C. Additional Discussion of Physical Space Alignment

Our pretraining bridges the vision-action gap to create a foundation VLA, but faces unique alignment challenges beyond standard visual instruction tuning. The key difficulties arise from three aspects: (1) The visual inputs from multiple sources vary in camera intrinsics and are captured under dynamic world coordinates. (2) The model's backbone is initialized with 2D vision-text pretraining, leaving it without crucial 3D spatial priors. (3) Essential physical properties, like force and friction, which humans intuitively understand, are inherently missing in video data. Unlike biological vision systems that organically develop 3D perception through embodied experience, we explicitly align these disparate data sources via perspective spatial alignment — unifying observations in a consistent coordinate system to instill 3D reasoning and physical understanding.

Beyond the two strategies proposed in Section 3.2, we believe that integrating richer physical cues can further improve the model's understanding of spatial and physical environments, which will extend 'perspective spatial alignment' to broader 'physical space alignment'. For instance, incorporating visual depth information, tactile feedback, or other multi-sensory signals may provide more grounded and embodied representations of human activities. These modalities offer complementary perspectives on physical interactions and 3D structure, which are often ambiguous or underspecified in 2D visual inputs alone.

Such multi-sensory integration could address fundamental limitations inherent in vision-only approaches. Depth information from RGB-D sensors could resolve spatial ambiguities that arise from weak-perspective projection, while tactile feedback could capture crucial contact dynamics, grip forces, and material properties that are invisible in visual observations but essential for successful manipulation. Audio signals from object interactions could further disambiguate manipulation strategies that appear visually similar but involve different physical processes, such as distinguishing between gentle placement and firm pressing actions.

These enhanced alignment strategies could create more robust representations that better capture the rich physical understanding humans naturally possess during manipulation tasks. As we scale our approach to larger and more diverse datasets, incorporating such multi-modal physical cues will become increasingly important for bridging the gap between human demonstration data and reliable robotic deployment across varied real-world scenarios.

D. Additional Details of UniHand

D.1. Data Curation Steps

Hand Pose Standardization. Our model learns an explicit mapping from 2D visual observations to 3D coordinates by standardizing all annotations into the MANO parameter format, ensuring both geometric precision and visual-semantic consistency. For datasets with mocap or SLAM-tracked labels, we extract MANO parameters directly (Romero et al., 2017). When only 3D joint positions are available, we fit MANO parameters via gradient-based optimization. For datasets lacking

3D annotations, we employ HaMer (Pavlakos et al., 2024) for frame-wise pose estimation, followed by post-processing: we detect and correct left-right hand mismatches by identifying pose discontinuities and apply temporal interpolation to fill minor gaps. This fitting process incorporates joint angle constraints and temporal smoothness to ensure physically plausible motions.

Task Description Labeling. We adopt hierarchical labeling to establish strong semantic grounding between vision, language, and motion, enriching the sparse or imprecise texts in existing datasets. Videos are segmented into non-overlapping chunks with a maximum length of 10 seconds, each capturing a distinct task phase. We sample frames at 2FPS and use Gemini-2.5-pro (Comanici et al., 2025) for multi-scale annotation. (1) Chunk Level: We produce imperative instructions and concise summaries to describe overarching hand activities and object interactions. (2) Per-second Level: We further divide each chunk into overlapping 1-second windows, annotating them with precise captions detailing contact states, object attributes, motion trajectories relative to the camera perspective. For completeness, we separately annotate global two-handed and individual hand actions. This strategy ensures comprehensive, consistent coverage from high-level objectives to fine-grained interactions.

Instructional Data Generation. Building on our systematic annotations, we construct instruction-following data to establish explicit vision-language-motion alignment for our VLA. Our instruction tasks focus on multiple grounding aspects for hand motion understanding, including spatiotemporal alignment of hand trajectories with visual context, object attributes and contact states, action intentions, and consistency between high-level instructions and fine-grained motion. Guided by these principles, we develop three complementary task types. (1) Instructional motion generation: producing step-by-step motion sequences from scene images and instruction; (2) Hand motion translation: converting motions and visual cues into language descriptions; and (3) Hand motion prediction: anticipating subsequent motions given prior history, current observation, and optional instructions or task goals. For implementation, we design 20 base templates per task type, using Gemini-2.5-Pro to generate diverse variants. Each template incorporates explicit duration specifications for variable temporal granularity. We use rule-based instantiation to populate these templates with grounded instructions, motion tokens, and length constraints.

D.2. Data Statistics

As we mentioned in the main paper, we curate our dataset from three primary sources, each offering distinct advantages and trade-offs: **(1) Motion capture datasets** (Fan et al., 2023; Garcia-Hernando et al., 2018; Kwon et al., 2021; Liu et al., 2022b) provide high-precision 3D annotations from multi-view systems in controlled environments (e.g., studios or labs), though they often lack diversity. For instance, OAKINK2 (Zhan et al., 2024) offers multi-view, object-centric recordings of real-world bimanual manipulation involving complex tasks. **(2) VR-recorded datasets** use devices like the Apple Vision Pro with calibrated cameras and SLAM-based tracking to capture natural hand–object interactions in less constrained settings while maintaining reliable 3D ground truth. A notable example is EgoDex (Hoque et al., 2025), which includes up to 194 household manipulation tasks such as tying shoelaces and folding laundry. **(3) Pseudo-annotated datasets** (Perrett et al., 2025) employ off-the-shelf hand motion predictors (Perrett et al., 2025) to generate pseudo 3D labels from in-the-wild videos. Although noisier, these datasets offer superior scalability and diversity, as demonstrated in large-scale applications (Cai et al., 2023). For example, Taste-Rob (Zhao et al., 2025) contains approximately 100K egocentric videos with aligned language instructions recorded from a fixed viewpoint. The recipe of our dataset UniHand is shown in Table 9. Our dataset is aggregated from diverse sources thus provides high diversity and broad coverage of real-world scenarios. Using an instructional labeling pipeline, we generate over 165M motion–instruction pairs for dexterous VLA learning. Our pipeline supports both scalable VR-recorded data and diverse pseudo-annotated in-the-wild videos. Due to computational constraints, we sample 2.5M instruction pairs via view-invariant motion distribution balancing across tasks and sources, forming the *UniHand-2.5M* subset. The view-balanced sample proportions are illustrated in Figure 8.

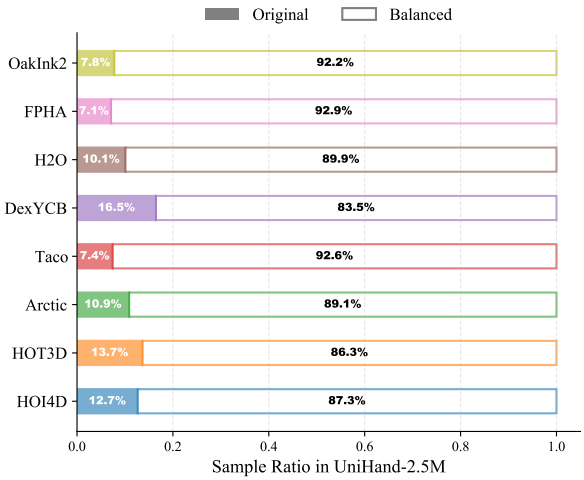

Figure 8. Comparison of instruction sample proportions in *UniHand-2.5M*: original data versus view-balanced data.

Table 9. Statistics of UniHand. Our dataset is the largest egocentric hand motion dataset to date, aggregating hand motions from 11 benchmarks across three sources: motion capture, VR-recording, and pseudo-annotation. By treating human hands as the template for dexterous manipulators or grippers, we anticipate UniHand will serve as a foundational resource for VLA learning. **#Inst** refers to the number of generated instructional samples.

Dataset	#Inst	#Seq	#Avg len	#Frames	#Hours	Hand joint	Hand Pose	Ann Granularity
ARCTIC (Fan et al., 2023)	17.9M	0.3K	725.3	245K	2.3	✓	✓	Action
FPHA (Garcia-Hernando et al., 2018)	798K	1.2K	89.8	105K	1.0	✓	✗	Action
HoloAssist (Wang et al., 2023)	8.0M	2.2K	8081.7	17.1M	166	✓	✗	Segment
H2O (Kwon et al., 2021)	3.7M	0.9K	121.5	115K	1.1	✓	✓	Action
HOI4D (Liu et al., 2022b)	21.2M	3.0K	273.0	825K	7.6	✓	✓	Action
HOT3D (Banerjee et al., 2025)	8.7M	2.8K	150.0	420K	3.9	✓	✓	N/A
OAKINK2 (Zhan et al., 2024)	18.5M	2.8K	244.4	695K	6.5	✓	✓	Action
TACO (Liu et al., 2024c)	11.5M	2.2K	154.0	340K	3.2	✓	✓	Action
DexYCB (Chao et al., 2021)	3.6M	5.6K	72.8	410K	3.8	✓	✓	N/A
Taste-Rob (Zhao et al., 2025)	1.9M	85K	164.1	14M	130	✗	✗	Trajectory
EgoDex (Hoque et al., 2025)	70.6M	338K	264.8	89.6M	829.4	✓	✗	Trajectory
Total	166.5M	444.1K	-	130M	1155	✓	✓	Fine-Grained

E. Additional Evaluation Setups

E.1. Hand Motion Modeling

Dataset Setups. We reserve 5% of videos along with paired annotations from UniHand for evaluation. Wrist translation distributions vary across sources, with EgoDex as the dominant contributor. We augment other sources to balance translation space coverage, resulting in a distribution where EgoDex forms the central mode and other datasets constitute a sparse long-tail. To reflect this structure, we evaluate on two distinct splits: the "**head split**"(held-out EgoDex samples) and the "**tail split**"(TACO, HOI4D, H2O, and OakInk2), assessing both dominant pattern capture and generalization to less frequent motion contexts.

Evaluation Metrics. We employ multiple metrics to holistically evaluate the model's ability to generate physically plausible, temporally coherent, and instruction-faithful hand motions, which are detailed below:

- **MPJPE (Mean Per Joint Position Error).** measures overall spatial accuracy by computing the mean Euclidean distance between each generated joint and its ground-truth in absolute 3D space.
- **MWTE (Mean Wrist Translation Error).** evaluates global trajectory fidelity through the mean Euclidean distance between predicted and ground-truth wrist positions across the sequence.
- **PA-MPJPE (Procrustes Aligned MPJPE).** isolates relative pose accuracy by aligning predicted joints to the ground truth via rigid transformation (including scaling, rotation, and translation).
- **M2T R@3 (Motion-to-Text Retrieval Top-3 Accuracy).** assesses semantic alignment by embedding generated motion into a shared representation space and retrieving the top-3 matching descriptions using a dataset-specific text-motion retrieval model (TMR (Petrovich et al., 2023)).
- **FID (Fréchet Inception Distance).** quantifies the distribution similarity by comparing the generated and real motion embeddings in the dataset-specific latent space of the TMR model, measuring how well synthesized motions match the true data distribution.
- **T2M R@3 (Text-to-Motion Retrieval Top-3 Accuracy).** reports how well generated descriptions retrieve corresponding motions from a database, which verifies whether the model's text output accurately captures the semantic content of motion.
- **Valid Rate.** We supplement T2M R@3 with the valid generation rate for free-form generation, which quantifies how consistently the model produces motion sequences adhering to the required structural format.

E.2. Real-world Dexterous Manipulation

Robot System. For our stable single-view setting, we use a Franka Research 3 arm (7-DoF), Inspire hand (6-DoF), and RealSense L515 camera for RGB streaming. To collect demonstrations for imitation learning, we introduce an improved teleoperation system that integrates a Gello exoskeleton (Wu et al., 2024) for arm control with a RealSense D435i camera for hand pose estimation and retargeting (Qin et al., 2023; Yuan et al., 2025) (Figure 4, Left). For mobile stereo-view setting, we use PND Adam-U with 31-DoF in total for jointly bimanual control, head control, and waist control. Each of the dexterous hands of PND Adam-U is 6-DoF. A set of ZED Mini is on the exact head to provide a mobile stereo view (Figure 4, Right).

Stable View Tasks. We design a suite of real-world manipulation tasks — including grasp-and-place, articulated object interaction, and deformable object manipulation — to evaluate fundamental skills, generalization, and precision. For each task, we collect 50–100 teleoperated trajectories to post-train our VLA. The policy maps egocentric RGB images and proprioception to action chunks containing end-effector poses and hand joint positions.

- **Grasping and Placing (*Pick-Place-Toy*):** This task comprises three scenarios of increasing difficulty. The **Seen** scenario evaluates foundational grasping of a familiar toy. The **Unseen** scenario tests object generalization by introducing a novel toy with distinct properties (e.g., color). Finally, the **Clutter** scenario assesses advanced perception and planning, requiring the robot to locate and retrieve the target from a cluttered array of distractors.
- **Articulated Object Manipulation (*Close-Toolbox*, *Close-Lid*):** These tasks require precise closure actions on a toolbox and a cup lid. They rigorously evaluate the model's capability for accurate end-effector positioning, orientation, and stable interaction with articulated object mechanics.
- **Deformable Object Manipulation (*Unfold-Clothes*):** This task challenges the robot to unfold a piece of cloth, testing its ability to perform fine-grained, multi-finger manipulation and reason about the dynamic state of non-rigid objects.
- **Precise Motion Control (*Pour-Cup*):** This task assesses motion planning and dynamic control, requiring the generation of continuous, stable trajectories to pour liquid from one cup to another while maintaining temporal action coherence.
- **Contact-Rich Dexterous Manipulation (*Spray–Plant*)**: This task evaluates the policy's ability to execute coordinated multi-finger dexterity under tight embodiment constraints. The robot must grasp a spray bottle by its narrow neck using a stable multi-finger configuration, reposition it towards the plant, and actuate the trigger to spray water. Unlike simple grasp-and-place tasks, this setting requires precise finger-role allocation (e.g., thumb and ring/pinky for stabilization, index finger for actuation), continuous force modulation, and accurate 3D spatial control. The task poses significant challenges involving contact-rich manipulation, fine-grained hand articulation, and dynamic interaction with deformable mechanisms (i.e., the trigger).

Mobile Stereo-View Tasks. To further increase task difficulty beyond the stable-view setting, we introduce a mobile stereo-view evaluation with ego-motion and additional head/waist DoF, which induces frequent viewpoint changes, partial self-occlusions, and depth-dependent scale variations. These factors stress a policy's ability to maintain temporally consistent visuomotor grounding and to coordinate actions under continuously changing perspectives. We evaluate two intrinsically challenging bimanual tasks.

- ***Arrange-Flower*** requires accurate 3D spatial perception to align the stem with the container opening, continuous state tracking of the insertion progress, and correct timing for release to avoid dropping or colliding with the vase.
- ***Bimanual-Handover*** demands tight bimanual coordination throughout the transfer, including stable pre-grasping, precise relative pose alignment between two hands, and smooth release/regrasp transitions under contact, where small errors can cause slips or failed handoffs.

We collect 222 teleoperated demonstrations for *Bimanual-Handover* and 485 for *Arrange-Flower* for post-training.

Evaluation Protocol. We use success rate as our primary evaluation metric and benchmark our Being-H against two baselines: GR00T N1.5 (Bjorck et al., 2025) and InternVL3 (Zhu et al., 2025). GR00T is selected as it is a large-scale VLA uniquely pre-trained on egocentric human videos for dexterous manipulation, contrasting with gripper-centric models like OpenVLA (Kim et al., 2024). InternVL3 provides a direct architectural and scale-matched comparison but lacks our method's hand motion pre-training and physical alignment. All models undergo identical post-training on the same

teleoperation datasets to assess the benefits of our pre-training on human hand data. To better understand model capabilities and failure modes, we perform a qualitative analysis of model behavior across three dimensions: motion precision (e.g., *Close-Lid*), semantic understanding of instructions (e.g., *Pick the white duck*), and robustness in complex tasks (e.g., *Unfold-Clothes*). To further assess the model's capability in contact-rich dexterous manipulation, we additionally report the completion score on *Spray–Plant* task. This setting stresses fine-grained control and embodiment adaptation, as successful execution requires coordinated multi-finger grasping, stable object reorientation, and precise trigger actuation. Thus, we report an average completion score by decomposing the task into three sequential sub-stages (grasping, positioning, spraying), allowing a more granular analysis of the manipulation progress. In the mobile stereo-view setting, we additionally compare with $\pi_{0.5}$ (Intelligence et al., 2025) under the same success-rate metric. This setting highlights whether a policy maintains temporally consistent, contact-sensitive control under ego-motion and occlusions.

Table 10. Comparison results of motion generation and prediction tasks upon long-range sequences. We adopt the soft-formatted decoding mode and report short-term (2–5s) and long-term (6–10s) results, respectively.

Model	Short-Term (2–5s)						Long-Term (6–10s)					
	MPJPE ↓		MWTE ↓		PA-MPJPE ↓		MPJPE ↓		MWTE ↓		PA-MPJPE ↓	
	head	tail	head	tail	head	tail	head	tail	head	tail	head	tail
# Hand Motion Generation												
Being-H-1B	8.97	9.96	7.01	8.75	1.43	1.67	9.12	11.24	7.13	9.91	1.60	1.81
Being-H-8B	7.55	8.45	5.78	7.51	1.10	1.30	8.21	9.98	6.12	8.34	1.22	1.36
Being-H-14B	7.43	8.39	5.65	7.39	1.11	1.28	7.98	9.72	5.88	8.01	1.18	1.32
# Hand Motion Prediction												
Being-H-1B	8.44	9.52	6.71	7.99	1.20	1.45	9.01	10.98	6.98	8.75	1.35	1.50
Being-H-8B	7.67	8.20	5.81	7.13	1.01	1.22	8.23	9.67	6.23	7.83	1.14	1.27
Being-H-14B	7.39	8.51	5.77	7.21	1.05	1.25	8.01	9.45	6.02	7.67	1.18	1.30

F. Additional Experiments

F.1. Comparisons on Long-range Motion Generation

We evaluate the long-term motion generation capabilities of different Being-H's variants in Table 10. To mitigate the inherent error accumulation that causes trajectory drift and degraded pose quality in longer sequences, we employ the soft-formatted decoding mode to constrain outputs within plausible ranges relative to ground-truth distributions. Results are categorized into **short-term** (2–5 seconds) and **long-term** (6–10 seconds) ranges to precisely examine the quality degradation. Using MPJPE, MWTE, and PA-MPJPE as the metrics for spatial accuracy and trajectory stability, we observe that generation quality deteriorates with longer sequences, evidenced by elevated MPJPE and MWTE. However, larger models maintain more stable spatial accuracy, as they better leverage the partial ground-truth context to anchor the trajectory under these soft constraints.

F.2. Ablation Study of Hand Motion Tokenization

In this section, we provide additional exploration about the configuration of hand motion tokenization.

MANO Feature Choice. We analyze the impact of different motion features through GRQ reconstruction. Axis-angle features yields superior overall reconstruction accuracy (e.g., MANO-D99 vs. MANO-D51; MANO-D162 vs. MANO-D114), while 6D rotation achieves better PA-MPJPE scores, indicating its potential advantage at modeling the finger actions. As Table 11 shows, the 6D rotation-based MANO-D162 feature is most effective for Being-H. We also find that incorporating auxiliary joint positions (j) enhances performance, but modeling hand shape parameters (β) degrades it. Consequently, we hold shape parameters constant from each sequence's initial frame, focusing the tokenizer exclusively on motion dynamics.

Impact of Motion Tokenizer on Hand Motion Generation. We adopt the MANO-D162 + part-level tokenizer as our default configuration for training Being-H. To verify this choice, we benchmark against three high-performing alternatives (MANO-D114 + 4-groups, MANO-D162 + 4-groups, and MANO-D114 + part-level). As demonstrated in Table 12, our default tokenizer consistently outperforms others in generation tasks, despite a slight higher reconstruction error with larger MPJPE shown in Table 11. We attribute its effectiveness to the 6D rotation representation and part-level decomposition, which better facilitate temporal modeling and autoregressive generation of fine-grained motions.

Table 11. Performance of different motion tokenization practices on the hand reconstruction task, including motion features and part-level tokenizing. Results are reported in centimeters (cm).

Feature	Part-Level		4-Groups		16-Layers	
	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
MANO-D51	0.556	0.209	1.165	0.184	1.466	0.243
MANO-D99	0.584	0.149	1.093	0.148	1.510	0.170
# with shape parameters β						
MANO-D109	0.592	0.160	1.107	0.140	1.602	0.201
# with auxiliary joint positions j						
MANO-D114	0.523	0.167	0.810	0.202	0.996	0.253
MANO-D162	0.573	0.129	0.704	0.138	1.054	0.226

Table 12. Ablation of motion tokenizer variants and data recipes, where "Trans" and "Pred" denote hand motion translation and prediction task respectively, while "Balance" represents view-invariant motion distribution balancing, a strategy adopted for physical space alignment. Evaluations are carried out on the hand motion generation benchmark.

Variants	MPJPE ↓		MWTE ↓		PA-MPJPE ↓		M2T R@3 ↑		FID ↓	
	head	tail	head	tail	head	tail	head	tail	head	tail
# Base										
Being-H-8B	**7.20**	**9.02**	5.69	**8.11**	**1.09**	**1.32**	**15.9**	**18.7**	11.5	**13.4**
# Tokenizer Variants										
MANO-D114 + 4-Groups	8.31	10.35	6.52	9.11	1.14	1.35	13.1	14.7	13.7	15.6
MANO-D162 + 4-Groups	7.98	9.71	5.58	8.98	1.09	1.38	15.4	17.1	11.9	14.3
MANO-D114 + Part-Level	7.74	9.92	6.11	8.83	1.16	1.41	12.3	16.1	13.1	12.3
# Data Recipe										
w/o Trans	7.22	9.11	**5.51**	8.12	1.27	1.46	13.2	11.6	13.1	15.7
w/o Pred	8.01	10.97	6.34	9.03	1.11	1.52	13.7	15.2	**11.1**	14.9
w/o Balance	8.54	12.13	7.74	10.04	1.24	1.57	11.3	10.3	15.8	16.3

F.3. Ablation of Data Configuration

To optimize Being-H, we construct UniHand with diverse instruction types. We also adopt the view-invariant motion distribution balancing (Section 3.2) to refine our dataset. We systematically ablate these components to understand their impact.

Impact of view-invariant motion distribution balancing. As described in Section 3.2, our balancing strategy equalizes motion-view coverage by augmenting hand poses under consistent weak-perspective projection. We evaluate its effect from two perspectives: **(1) Tokenizer Learning.** As shown in Figure 9, balancing significantly reduces GRQ reconstruction errors on the held-out test set of UniHand, even for datasets without explicit augmentation (e.g., EgoDex, Taste-Rob). This indicates that more precise encoding of global wrist and fine-grained finger motion due to evenly distributed motion-view coverage. **(2) Pretrained VLA Learning.** We compare Being-H with a variant trained on the dataset without the balancing strategy. Table 12 shows that removing balancing causes substantial performance degradation on the tail split. Without it, the model overfits to dominant camera configurations and fail to generalize to underrepresented perspectives. These results underscore the critical role of view-invariant balancing in enhancing both the tokenizer's representational robustness and the pretrained VLA's generalization for generating accurate and semantically grounded motions across diverse views.

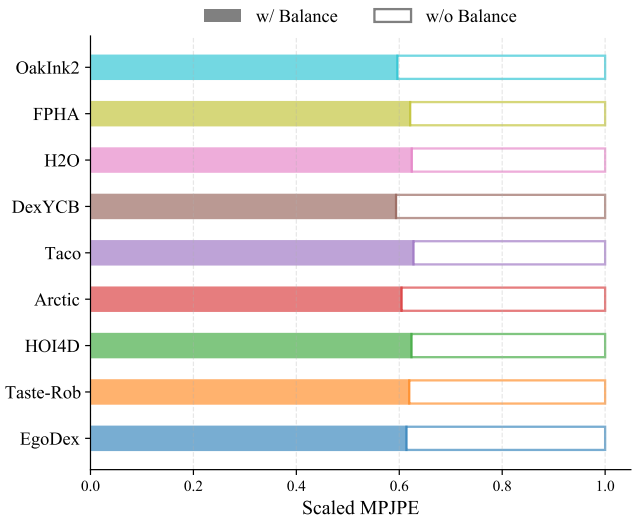

Figure 9. Ablation of view-invariant motion distribution balancing ("Balance") on the motion reconstruction task.

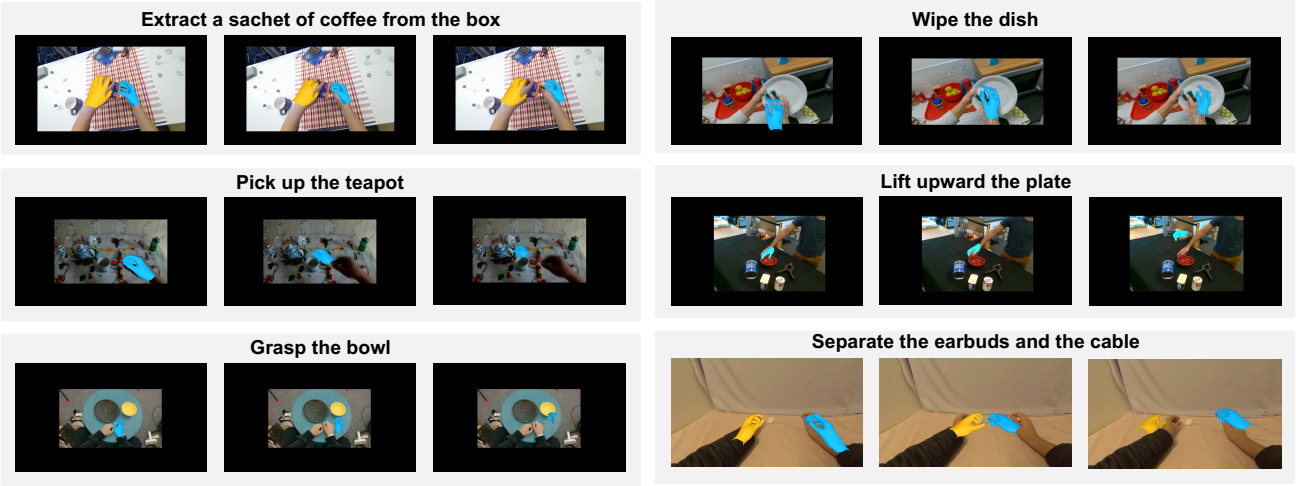

Figure 10. Qualitative examples of hand motions generated by Being-H-8B across diverse tasks, scenes, and viewpoints. Each block shows simplified instruction and three temporal frames of the motion, rendered in the first-frame camera coordinates and overlaid on the input image. Black padding ensures consistent weak-perspective projection.

Benefits of auxiliary supervision tasks. As introduced in Section D.1, we incorporates two auxiliary supervision types in UniHand in addition to instructional motion generation data: hand motion translation and contextual motion prediction. Evaluation on the core motion generation task reveal their distinct benefits as shown in Table 12. First, removing translation data yields marginal changes in global wrist metrics (MPJPE, MWTE), but leads to clear degradation in PA-MPJPE, M2T R@3, and FID, underscoring its role in generating semantically aligned, detailed articulations. Second, removing motion prediction data leads to uniform drops across all metrics, highlighting its central role in ensuring temporally coherent and global context awareness. Thus, auxiliary supervision not only enables task-specific capabilities but also significantly strengthens the core generation model through improved semantic and temporal grounding.

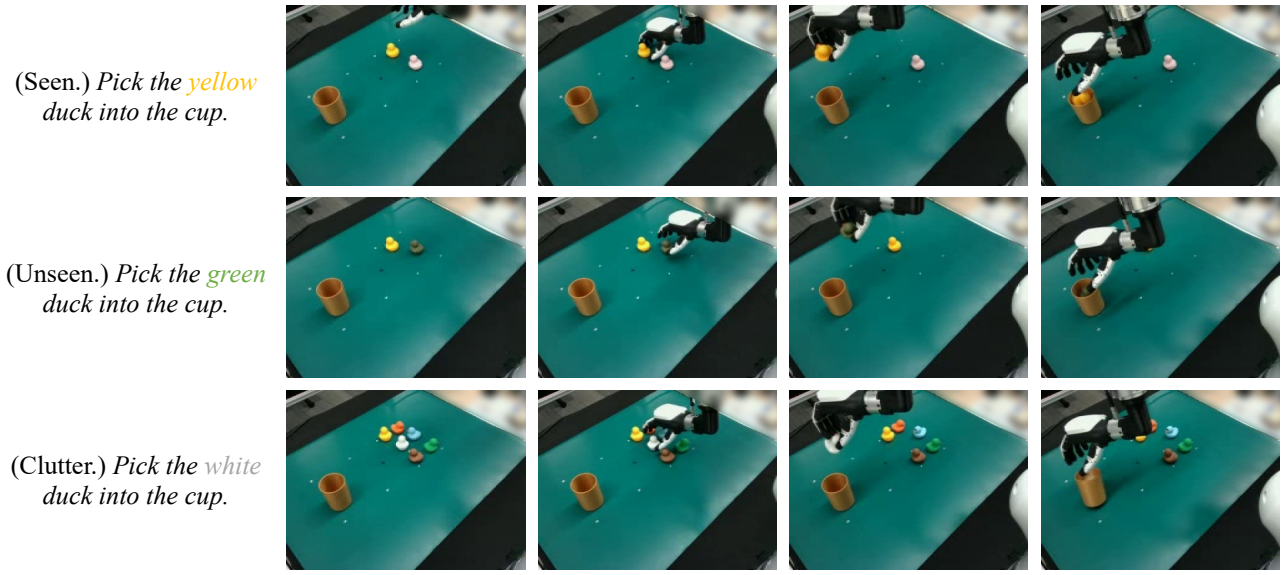

Figure 11. Qualitative examples of real-world task generation: Being-H performing the *Pick-Place-Toy* task under three conditions, including seen objects, unseen objects, and cluttered scenes.

F.4. Additional Post-training Scaling Experiments

In the main paper, we demonstrate that Being-H achieves strong downstream performance under *few-shot* post-training. To investigate whether our pretrained VLA continues to improve beyond the few-shot regime, we conduct a scaling study

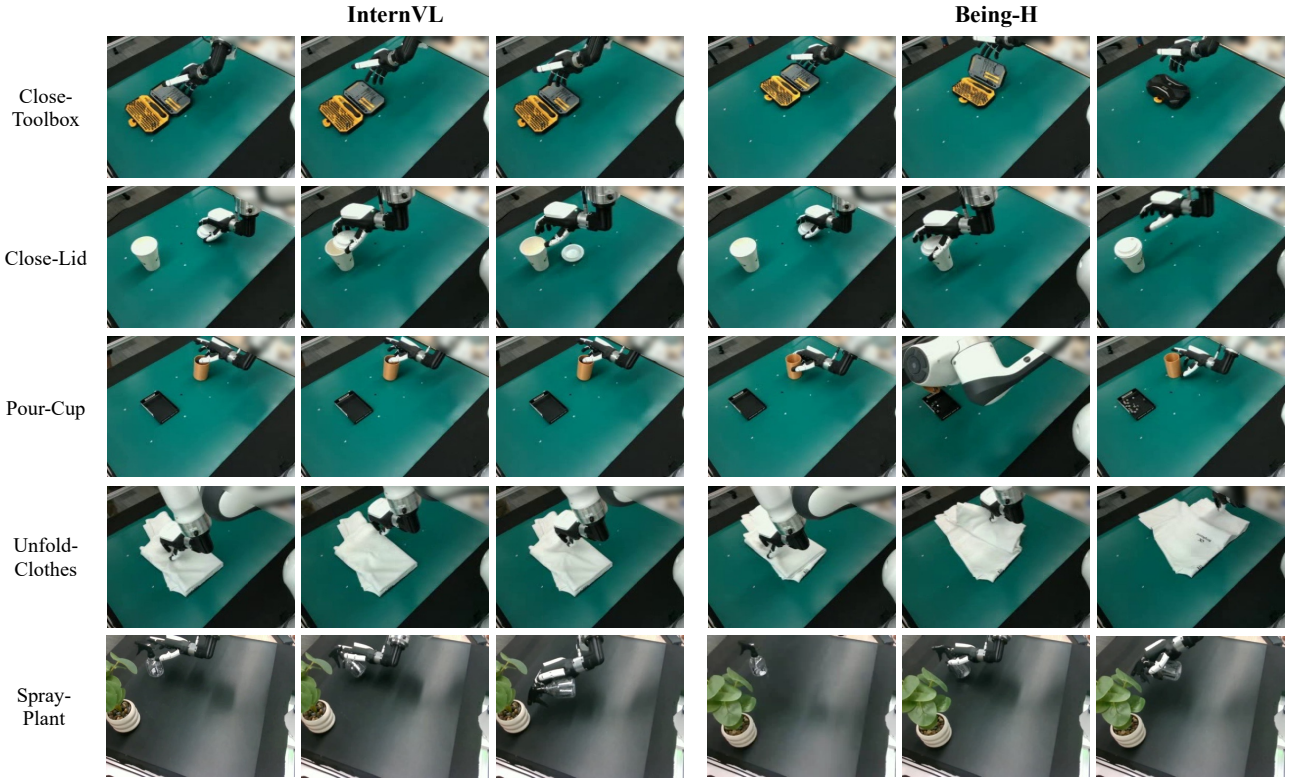

Figure 12. Qualitative comparisons between Being-H and InternVL3 baseline without VLA pretraining.

Table 13. Scaling post-training demonstrations on RoboCasa (3-view, Being-H(FM)). We report success rates (%).

Tasks	Pick & Place	Door	Drawer	Lever	Knob	Insert	Button	Avg.
50×24 demos	10.0	64.5	48.0	58.0	21.0	26.0	34.0	33.5
100×24 demos	16.0	71.5	55.0	60.7	25.0	27.0	38.0	38.5
200×24 demos	18.0	76.0	63.0	66.7	24.0	31.0	41.3	42.0

on RoboCasa using the information-rich three-view setup and the higher-capacity flow-matching action head (Being-H(FM)). This setting provides the most expressive inputs and output head, enabling us to examine whether additional robot demonstrations can further ground the pretrained human priors and push the policy performance to a higher level.

We fine-tune Being-H(FM) with three demonstration budgets per task: 50, 100, and 200 demonstrations across the 24 RoboCasa atomic tasks, corresponding to a total of 1200, 2400, and 4800 demonstrations, respectively. Policies receive three-view RGB observations and proprioception as input. We report success rates (%) across seven task categories and the overall average.

As is shown in Table 13, scaling the amount of post-training demonstrations yields consistent and meaningful improvements. The overall average success increases from 33.5% to 38.5% and further to 42.0% as the demonstration budget grows from 50 to 200, indicating that the pretrained human priors can be progressively grounded into the robot embodiment with additional supervision.

F.5. Additional Qualitative Examples

Hand Motion Generation. To demonstrate Being-H's capability in generating physically plausible hand motions across diverse tasks, scenes, and viewpoints, we present additional samples in Figure 10 for visualization. Each example illustrates a generated motion sequence rendered over the first frame in its respective camera coordinate system, with hands color-coded for clarity (yellow for the left hand, blue for the right hand). To enable consistent visualization under a unified

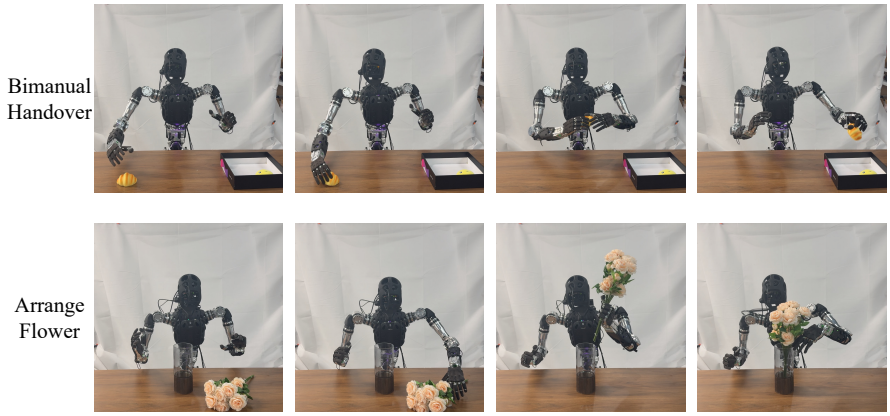

Figure 13. Mobile stereo-view qualitative examples: *Bimanual-Handover* (top) and *Arrange-Flower* (bottom), illustrating coordinated bimanual transfer and precise insertion/release under ego-motion.

weak-perspective projection, we apply a standardized transformation to these frames, which introduces black borders around the images. The effective region — excluding the black padding — captures variations in hand-object interaction depth, where closer interactions manifest as larger apparent sizes.

Real-world Task Generalization. Our model adeptly handles both single-hand and dual-hand manipulations across a broad spectrum of tasks, showcasing robust generalization to varied viewpoints and physical contexts. For instance, Figure 11 highlights how Being-H not only successfully picks up the seen yellow duck but also generalizes to the unseen green duck. Even more impressively, in a cluttered environment with multiple distractors, Being-H accurately adheres to the instruction "Pick the white duck", precisely identifying and retrieving the target white duck. This underscores the model's seamless integration of visual perception, language comprehension, and action generation.

Being-H vs. InternVL3 on Fine-grained Tasks. The advantages of Being-H become especially evident in tasks requiring fine-grained manipulation. In contrast, the baseline model InternVL3, which lacks physical instruction tuning and prior knowledge related to hand motion dynamics, exhibits markedly weaker performance. A qualitative comparison presented in Figure 12 clearly reveals several characteristic failure modes of InternVL3:

- *Close-Toolbox*: The motion trajectory from InternVL3 baseline lacks precision, frequently missing contact with the edge of the toolbox lid, thereby failing to generate sufficient force to close it.
- *Close-Lid*: The InternVL3 demonstrates positional deviation, often misaligning the lid beside the cup's rim rather than seating it correctly.
- *Pour-Cup*: The grasp of InternVL3 baseline is unstable, occasionally failing to securely hold the cup, which compromises the stability of the subsequent pouring motion.
- *Unfold-Clothes*: The InternVL3 baseline misjudges the operational height, causing the fingers to close at an incorrect elevation and miss the cloth's edge, resulting in a failed unfolding attempt.
- *Spray-Plant:* The InternVL3 baseline struggles to establish a stable grasp on the bottle's narrow neck, often making contact too low or at an incorrect angle, which prevents the model from allocating proper finger roles for support. As a result, the bottle frequently slips or rotates during the trigger-pressing phase, leading to incomplete or failed watering attempts.

Mobile stereo-view behaviors. Figure 13 shows that Being-H executes temporally consistent, contact-aware behaviors under ego-motion. In *Bimanual-Handover*, the policy forms a stable pre-grasp, brings both hands into accurate relative alignment, and performs a smooth release–regrasp transition without interrupting the object support. In *Arrange-Flower*, it maintains a secure grasp while approaching the container, aligns the stem with the opening, and releases only after insertion is completed, reflecting stronger state awareness and fine-grained control compared to the baselines.

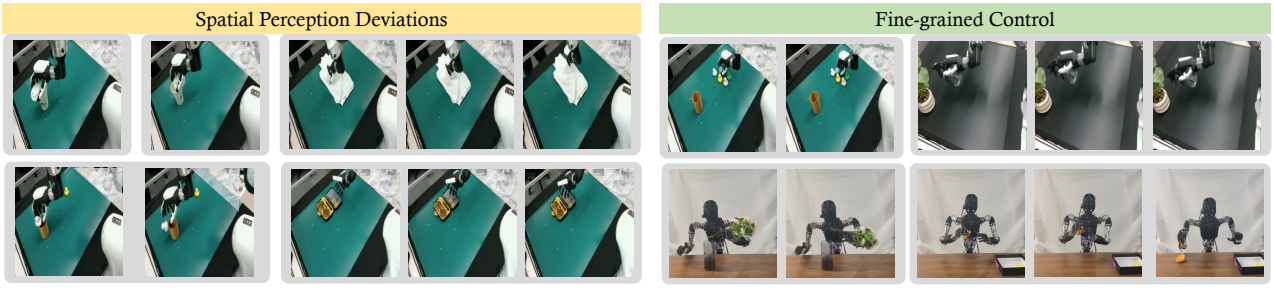

Figure 14. Failure case modes across real-world dexterous manipulation tasks.

F.6. Failure Cases Study

Across the evaluated tasks, we observe that the failure cases (Figure 14) of Being-H fall into two primary categories.

Spatial perception deviations. A number of errors originate from small but impactful inaccuracies in estimating 3D contact locations from monocular RGB observations. In tasks such as *Unfold-Clothes* and *Close-Toolbox*, the end-effector visually appears to be correctly aligned, yet the actual contact point remains slightly offset, preventing effective interaction with the cloth edge or toolbox lid. Similar but more subtle deviations occur in *Close-Lid* and *Pick-Place-Toy*, where the gripper occasionally misses the cup rim or toy by a narrow margin. These errors reflect the inherent ambiguity in depth and fine-scale geometry under single-view RGB input, which can cause near-contact states to result in failure despite seemingly correct approach trajectories.

Fine-grained control. Another typical failure arises when the global approach and grasp region are correct, but the task hinges on precise finger placement or subtle contact modulation. This pattern appears in duck-toy grasping and is most evident in *Spray-Plant*: the robot often stabilizes the bottle in a largely correct pose (substantially better than InternVL3), yet small deviations in finger position or closure timing can prevent the index finger from reliably actuating the trigger. Similar fine-control bottlenecks also occur in the two mobile stereo-view tasks: in *Arrange-Flower*, minor misalignment during insertion or an imperfect release timing can snag the stem or drop the bouquet, while in *Bimanual-Handover*, slight pose errors during the handoff can lead to slips or failed transfer. Overall, the pretrained model provides strong *high-level* behavioral priors, but these do not fully guarantee *fine-grained* dexterous control when contact geometry and multi-finger coordination are critical. The lightweight MLP adapter further compounds this issue, as its single-step mapping offers limited expressiveness for modeling delicate force adjustments.

Although these errors occur infrequently, they suggest that further gains may be achieved by enriching 3D perception (e.g., multi-view or depth cues) and exploring more expressive continuous-action heads to better support fine-grained dexterous control.