UniT: Toward a Unified Physical Language for Human-to-Humanoid Policy Learning and World Modeling

Boyu Chen[1,2,*] **Yi Chen**[1,3,*] **Lu Qiu**[3] **Jerry Bai**[1] **Yuying Ge**[1,†] **Yixiao Ge**[1]
[1]XPENG Robotics [2]Tsinghua University [3]The University of Hong Kong
https://xpeng-robotics.github.io/unit/

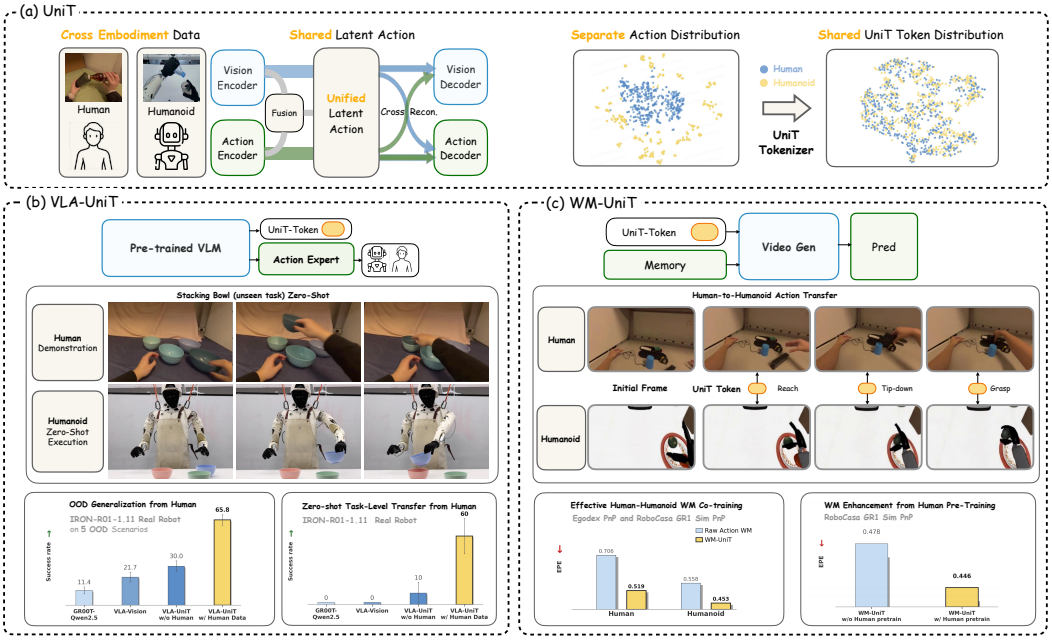

Figure 1: Overview of the UniT Framework, which establishes a **unified physical language** to bridge the human-humanoid chasm. (a) **Unified Tokenization:** A tri-branch cross-reconstruction mechanism projects heterogeneous actions into a shared discrete latent space (verified via t-SNE) for extracting unified physical intents. (b) **Policy Learning**: Through predicting these unified tokens, VLA-UniT effectively leverages diverse human data to achieve robust OOD generalization and *zero-shot task transfer* on humanoid robot. (c) **World Modeling**: By aligning cross-embodiment dynamics through unified token conditioning, WM-UniT enables direct human-to-humanoid action transfer, effectively leveraging human priors to improve controllability in humanoid video generation.

Abstract

Scaling humanoid foundation models is bottlenecked by the scarcity of robotic data. While massive egocentric human data offers a scalable alternative, bridging the cross-embodiment chasm remains a fundamental challenge due to kinematic mismatches. We introduce **UniT** (**Uni**fied Latent Action **T**okenizer via Visual Anchoring), a framework that establishes a unified physical language for human-to-humanoid transfer. Grounded in the philosophy that heterogeneous

* Equal contribution.
† Corresponding author.

Preprint.

kinematics share universal visual consequences, UniT employs a tri-branch cross-reconstruction mechanism: actions predict vision to anchor kinematics to physical outcomes, while vision reconstructs actions to filter out irrelevant visual confounders. Concurrently, a fusion branch synergies these purified modalities into a shared discrete latent space of embodiment-agnostic physical intents. We validate UniT across two paradigms: 1) **Policy Learning (VLA-UniT):** By predicting these unified tokens, it effectively leverages diverse human data to achieve state-of-the-art data efficiency and robust out-of-distribution (OOD) generalization on both humanoid simulation benchmark and real-world deployments, notably demonstrating *zero-shot task transfer*. 2) **World Modeling (WM-UniT):** By aligning cross-embodiment dynamics via unified tokens as conditions, it realizes direct human-to-humanoid action transfer. This alignment ensures that human data seamlessly translates into enhanced action controllability for humanoid video generation. Ultimately, by inducing a highly aligned cross-embodiment representation (empirically verified by t-SNE visualizations revealing the convergence of human and humanoid features into a shared manifold), UniT offers a scalable path to distill vast human knowledge into general-purpose humanoid capabilities.

1 Introduction

Scaling foundation models for humanoids in both policy learning and world modeling is fundamentally bottlenecked by scarce high-quality robotic data. Massive, structured human motion sequences from low-cost capture provide a scalable alternative rich in physical interaction priors, but leveraging them requires bridging a major cross-embodiment gap[1]. Biomechanical and hardware differences create heterogeneous state-action spaces with mismatched degrees of freedom (DoF) and control paradigms. Traditional pipelines rely on **motion retargeting**[2, 3], which uses complex kinematic solvers to map human motions to specific robots. This case-by-case process is labor-intensive, unscalable, and often physically inconsistent. We therefore need a data-driven **unified physical language** that projects heterogeneous data into a shared latent action space.

Although recent literature explores unified representations, critical limitations remain (Fig. 2). **Action-only** methods [4, 5, 6, 7, 8] rely exclusively on proprioceptive reconstruction, often suffering from severe distribution shifts between humans and robots due to the lack of external grounding. Conversely, emerging **latent action** frameworks are predominantly **vision-only** [9, 10, 11, 12], inferring intent directly from pixels. While bypassing kinematic mismatches, these representations are prone to entangling low-level appearance confounders (e.g., textures and lighting). This entanglement limits fine-grained physical execution and leaves the structural priors of human pose data underexploited. Furthermore, while some paradigms [13] incorporate **both vision and action**, they typically employ independent tokenizers for each modality. This results in disjoint vocabularies without deep representational unification, failing to establish a truly universal medium for control. Empirically, many existing latent-action systems are confined to fixed single- or dual-arm setups with simple grippers, leaving their scalability to dexterous humanoid largely underexplored.

To address these challenges, we introduce **UniT** (**Uni**fied Latent Action **T**okenizer via Visual Anchoring). Our design is driven by a critical insight into cross-embodiment alignment: while human and humanoid kinematics differ in structural DoFs and contain embodiment-specific noise, the physical outcomes of their intents share a consistent visual representation. Therefore, visual observations can serve as a universal anchor to ground and align disparate kinematic spaces. Building on this principle, UniT functions as a cross-modal information bottleneck to distill the underlying physical intent. The tokenizer concurrently extracts three coupled representations: a *temporal-visual* feature from consecutive frames, a *kinematic* feature from corresponding inter-frame actions, and a *fused visuo-motor* feature. Instead of treating these as isolated streams, we enforce a rigorous **cross-reconstruction objective**, compelling each representation to independently decode both the visual transitions and the low-level actions.

This mechanism operationalizes the concept of **visual anchoring**. By forcing kinematic features to reconstruct visual transitions, heterogeneous actions are anchored to their actual physical consequences in the environment. This constraint prevents the network from merely memorizing embodiment-specific kinematics and filters out visually unobservable artifacts. Conversely, compelling visual features to reconstruct kinematics strips away low-level appearance confounders, such as lighting

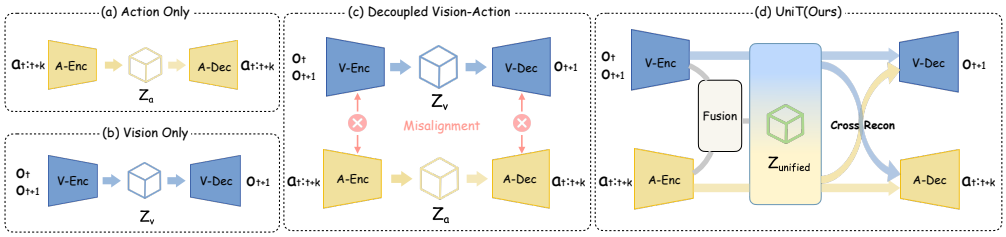

Figure 2: **Comparison of Latent Action Architectures. (a) Action-Only:** Relies on action reconstruction (Z_a), suffering from distribution misalignment without visual grounding. **(b) Vision-Only:** Infers representations (Z_v) from pixels, which entangles with low-level appearances and misses fine-grained physical details. **(c) Decoupled Vision-Action:** Encodes modalities into disconnected spaces (Z_v, Z_a), lacking the explicit alignment needed to extract shared physical concepts. **(d) UniT (Ours):** Fuses heterogeneous data into a shared space ($Z_{unified}$) via **cross-reconstruction**. This strict cross-modal alignment ensures tokens capture universal physical intentions, forming a robust unified physical language.

and irrelevant backgrounds, that do not contribute to physical motion. Uncorrelated noise from either domain is discarded during optimization, preserving only the essential intersection of both modalities: the embodiment-agnostic physical intent. As a result, UniT generates deeply integrated visuo-motor tokens that synergize visual and kinematic information into a unified latent action space. This strict alignment also ensures that the independent visual and kinematic branches extract structured cross-modal tokens, maintaining robust representations even in the absence of a modality during deployment. Ultimately, these visually-anchored discrete tokens serve as a **universal physical language**, providing a stable foundation for transferring intent across different robot morphologies.

To evaluate this unified language, we deploy UniT in two embodied-AI paradigms:

First in **Policy Learning.** We introduce **VLA-UniT** by integrating UniT into Vision-Language-Action architectures. Instead of fitting raw actions across large distribution gaps, VLA-UniT predicts UniT tokens in the shared latent space, and a lightweight flow head then generates embodiment-specific actions for execution. We evaluate on the RoboCasa GR1 benchmark and a real humanoid. VLA-UniT outperforms state-of-the-art baselines and improves data efficiency and out-of-distribution (OOD) generalization using diverse human data, and shows zero-shot task transfer with emergent upper-body coordination (e.g., waist rotation) on unseen tasks.

Second in **World Modeling.** We propose **WM-UniT**, which uses UniT tokens as universal conditions instead of raw actions. Because these tokens absorb physical priors during joint dynamics training, they improve prediction consistency. Rollout validations show that pre-training on large-scale human data aligns fine-grained human and humanoid actions, enabling transfer of physical dynamics across embodiments and improving downstream humanoid control generation.

Beyond downstream performance, UniT offers structural benefits. It forms a shared cross-embodiment latent space and also encourages downstream models to align their internal features. In both VLA policies and world models, UniT tokens induce more aligned cross-embodiment context representations. The tokenizer also provides useful denoising: encoding and decoding noisy captured actions filters perturbations and recovers cleaner, more executable trajectories.

Our main contributions are summarized as follows:

- **The Unified Tokenizer (UniT):** We propose a visual-anchored tri-branch tokenizer with cross-reconstruction that maps heterogeneous actions into a shared discrete latent space, yielding robust cross-embodiment alignment and action denoising.
- **Generalizable Policy Execution (VLA-UniT):** We integrate UniT into VLA architectures and achieve strong data efficiency, superior OOD generalization, and zero-shot task transfer for humanoids in simulation and the real world.
- **Unified World Modeling (WM-UniT):** We use UniT tokens as universal conditions for world models, showing that co-training on human and humanoid data improves dynamics prediction and downstream control generation.

2 Related Work

Learning from Human Data. Leveraging human data for robot learning has attracted growing attention as a scalable alternative to costly robot demonstrations. A prominent line of work pre-trains visual representations from egocentric human videos [14, 15, 16, 17], showing positive transfer to downstream manipulation. However, most of these approaches focus on unsupervised visual learning without exploiting fine-grained hand or wrist pose information, limiting their utility for dexterous upper-body manipulation. More recent methods co-train unified policies on human and robot demonstrations through explicit alignment or motion information [18, 19, 20, 2, 21]. However, as illustrated in [1, 22], co-training on mixed embodiment data end-to-end forces the model to fit fundamentally different action distributions simultaneously, often leading to embodiment-specific shortcuts rather than shared representations. To mitigate this, another line of work introduces motion retargeting. EgoVLA [3] and In-n-On [1] predict unified human wrist and hand actions and then apply inverse kinematics to map them onto robot joint configurations. However, retargeted actions often misalign with the original visual observations, and the IK-based pipeline remains case-by-case and difficult to scale across diverse morphologies. UniT addresses these challenges by learning a shared latent space that absorbs embodiment differences at the representation level, bypassing the need for explicit retargeting or action-space unification.

Latent Action Representations. A growing body of work explores latent action representations for robot learning, which we categorize by the modality they encode. **Action-only** methods [4, 5, 6, 7, 8] learn discrete or continuous autoencoders over raw action trajectories to produce compact representations for policy learning. Among them, VQ-BeT [4] and FAST [5] demonstrate that structured tokenization improves behavior generation. However, without external grounding, these representations reflect embodiment-specific kinematics and struggle to align heterogeneous action distributions, limiting cross-embodiment transferability. **Vision-only** methods [9, 10, 11, 12] infer latent actions directly from visual observations to bypass kinematic mismatches. Moto [9] and LAPA [10] learn latent motion tokens from video, while UniVLA [11] uses vision-derived latent actions for cross-embodiment policy learning. While this offers cross-domain potential, such representations tend to entangle low-level appearance factors and miss fine-grained motor detail, underexploiting the structural priors available in human pose data. Villa-X [23] partially addresses this by incorporating action reconstruction as an auxiliary target, but the unidirectional vision-to-action objective still limits the precision of the learned motor representation. Concurrent works such as METIS [13] and XR-1 [24] take **both vision and action as encoder inputs** but have not achieved explicit vision-action alignment. XR-1 applies KL regularization to encourage distributional proximity, which may not fully capture the fine-grained cross-modal correspondence that cross-reconstruction provides.

Vision-Language-Action Model and Action-Conditioned World Models. Policy learning and **world modeling** represent two core paradigms for embodied AI. For policy learning, VLA models [25, 26, 5, 27] integrate vision-language backbones with action generation for closed-loop control. Cross-embodiment generalist policies such as GR00T [26], π_0 [28], RT-X [29], and Octo [30] train across multiple embodiments, generating raw actions via diffusion heads or predicting action tokens from proprioceptive data. For world modeling, action-conditioned video generation has emerged as a promising approach for simulating robotic dynamics. IRASim [31], Ctrl-World [32], and WPE [33] explore controllable generation conditioned on robot actions, building upon video foundation models such as Cosmos [34]. Across both paradigms, the heterogeneity of action spaces across embodiments remains a key bottleneck, as most existing systems are confined to single-embodiment or single-arm gripper settings with limited cross-embodiment validation. UniT provides a unified token interface for both paradigms, projecting heterogeneous actions into a shared latent space that serves as a prediction target for VLA and a conditioning signal for world models, enabling scalable human-to-humanoid transfer across both policy learning and world modeling.

3 Methodology

3.1 Overview

Our goal is to establish a unified physical language that bridges heterogeneous human and humanoid action spaces for cross-embodiment transfer. We consider human and humanoid demonstrations as sequences of observations, embodiment-specific states, and actions, (o_t, s_t, a_t). We first present **UniT** (Sec. 3.2), a visual-anchored latent action tokenizer built upon a tri-branch architecture. By

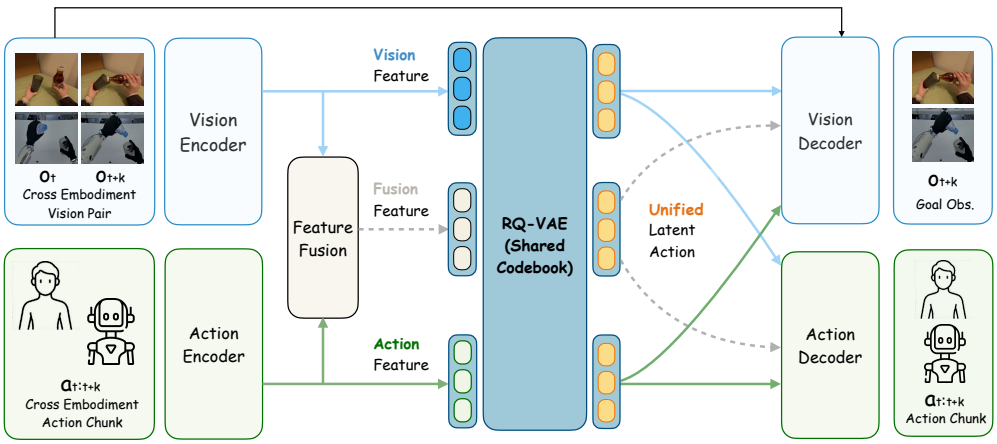

Figure 3: **Architecture of UniT.** Heterogeneous cross-embodiment vision pairs (o_t, o_{t+k}) and action chunks $a_{t:t+k}$ are encoded into vision, action, and fused features via tri-branch encoders. A **shared RQ-VAE codebook** quantizes all three branches into a unified discrete space, yielding embodiment-agnostic **Unified Latent Action** tokens. Both vision and action decoders reconstruct from these shared tokens, enforcing **cross-reconstruction** alignment.

jointly modeling visual transitions, actions, and fused visuo-motor features, UniT produces discrete tokens that capture embodiment-agnostic physical intent and serve as a unified interface across embodiments.

Leveraging this shared token representation, we deploy UniT in two complementary embodied-AI paradigms. **VLA-UniT** (Sec. 3.3) models the policy decision process by predicting future action chunks from the current observation and state through UniT token prediction and embodiment-specific action generation. **WM-UniT** (Sec. 3.4) models the dynamics prediction process by predicting future visual observations from the current observation and action conditions, using UniT features as a universal control interface instead of embodiment-specific raw actions.

3.2 UniT: Unified Latent Action Tokenizer via Vision Anchoring

Given an observation transition (o_t, o_{t+k}), current state s_t, and action chunk $a_{t:t+k}$, UniT learns a discrete latent action representation that maps heterogeneous human and humanoid behaviors into a shared token space. The resulting tokens serve as a unified embodiment-agnostic interface for downstream policy learning and world modeling. To achieve this, UniT is built upon a tri-branch architecture that jointly models visual transitions, actions, and fused visuo-motor features under cross-reconstruction (Fig. 3).

Tri-Branch Encoding. UniT uses three parallel branches. Each branch employs a transformer encoder with learnable queries to summarize modality-specific inputs into a compact latent representation before quantization.

- **Visual branch** E_v: operates as an inverse dynamics model (IDM), taking frozen DINOv2 [35] features of the observation pair (o_t, o_{t+k}) as input, following the visual feature design adopted in UniVLA [11], and producing a latent representation of the underlying physical transition. The domain-invariant nature of DINOv2 provides a stable visual anchor across embodiments.
- **Action branch** E_a: encodes current state s_t and action chunk $a_{t:t+k}$. Because human and humanoid embodiments differ in control mode, action parameterization, and degrees of freedom, raw actions are first padded to a unified maximum length and projected by embodiment-specific MLPs, then summarized into a compact latent control representation.
- **Fusion branch** E_m: takes the branch features from vision and action as input and produces a fused visuo-motor latent representation, capturing complementary cross-modal structure for more compact and robust tokens.

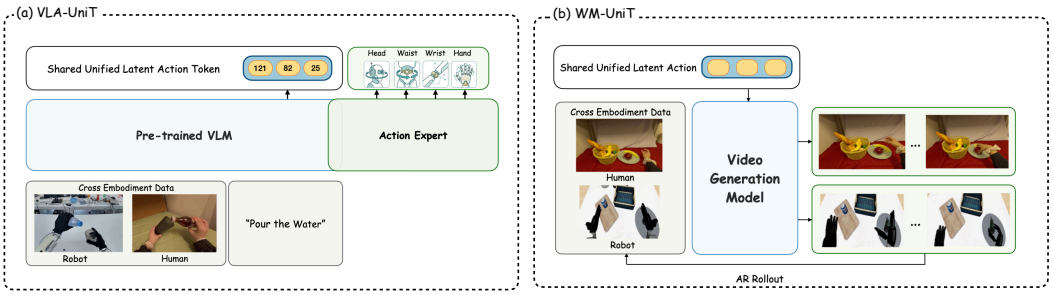

Figure 4: **Downstream Applications of UniT Tokens. (a) VLA-UniT:** Integrates UniT token prediction into a VLA architecture. Rather than direct action regression, the model predicts Unified Latent Action tokens in the shared space, while a lightweight action expert generates embodiment-specific controls from the same vision-language context, covering head pose, waist pose, wrist pose, and hand pose. **(b) WM-UniT:** Employs UniT tokens as universal action conditions for world modeling instead of embodiment-specific raw actions, generating future frames with autoregressive rollout.

Shared Discrete Quantization. Continuous latents from all three branches are quantized via Residual Quantization (RQ-VAE) with a shared codebook $\mathcal{C}$:

$$\hat{z}_i = \mathrm{RQ}(z_i; \mathcal{C}), \quad i \in \{v, a, m\}. \tag{1}$$

The shared codebook ensures that tokens from all branches and all embodiments reside in a unified discrete space. Residual quantization progressively refines the approximation through multiple codebook levels, capturing both coarse physical intent and fine-grained motion detail within a compact representation.

Cross-Reconstruction. The core mechanism of UniT is cross-reconstruction: every quantized token $\hat{z}_i$ is decoded by both a shared visual decoder D_v and an embodiment-specific action decoder D_a:

$$\hat{f}^{(i)}_{t+k} = D_v(\hat{z}_i,\, f_t), \qquad \hat{a}^{(i)}_{t:t+k} = D_a(\hat{z}_i,\, s_t), \tag{2}$$

where f_t denotes the DINOv2 feature of o_t. The visual decoder operates as a forward dynamics model (FDM) conditioned on the current observation, supervised by cosine similarity against f_{t+k}; the action decoder is conditioned on the current state s_t to reconstruct the action chunk. This design encourages the token to capture relative physical change, rather than memorizing absolute configurations, which benefits cross-embodiment transfer because different embodiments can still share analogous transition patterns despite vastly different state spaces.

This bidirectional constraint realizes the principle of visual anchoring: *vision provides a universal physical anchor across embodiments*. While heterogeneous action spaces are inherently incomparable, visually similar task outcomes can still be shared across embodiments. Through a shared visual decoder, tokens producing similar physical effects can converge to nearby codebook entries regardless of embodiment, anchoring heterogeneous motor representations into a unified manifold.

Training Objective. The total loss aggregates cross-reconstruction and quantization terms across all three branches:

$$\mathcal{L} = \sum_{i \in \{v,a,m\}} \Big[\lambda_v\, \mathcal{L}_{\cos}(\hat{f}^{(i)}_{t+k},\, f_{t+k}) + \lambda_a\, \mathcal{L}_{\text{act}}(\hat{a}^{(i)}_{t:t+k},\, a_{t:t+k}) \Big] + \mathcal{L}_{\text{RQ}}, \tag{3}$$

where $\mathcal{L}_{\cos}$ is the cosine similarity loss for visual feature reconstruction, $\mathcal{L}_{\text{act}}$ is the action reconstruction loss, and $\mathcal{L}_{\text{RQ}}$ is the RQ-VAE commitment loss. For downstream deployment, VLA-UniT leverages fusion-branch tokens that integrate visual and motor understanding for policy prediction (Sec. 3.3), while WM-UniT uses action-branch features as the control interface, since future frames must be generated by the world model itself and are therefore unavailable at the input side (Sec. 3.4).

3.3 VLA-UniT: Cross Embodiment Policy Learning via UniT

Given the current observation o_t, state s_t, and language instruction ℓ, VLA-UniT models the policy decision process by predicting a future action chunk $a_{t:t+H}$. Built upon the GR00T n1.5 framework [26] with Qwen2.5-VL [36] as the vision-language backbone, it uses UniT tokens as a structured

cross-embodiment prediction target for the VLM, while the action expert generates embodiment-specific controls from the same vision-language context. We therefore decompose policy learning into UniT token prediction and flow-matching action generation.

UniT Token Prediction. Learnable queries q_t are appended to the VLM sequence. Conditioned on the current observation o_t and language instruction ℓ, the VLM predicts UniT discrete codes through these queries. The target codes are obtained by encoding the ground-truth observation pair, state and action chunk through the pre-trained UniT tokenizer, $c_t = \mathrm{UniT}(o_t, o_{t+k}, a_{t:t+k}, s_t)$. Here, UniT serves as the prediction target, while the VLM vision-language context provides the conditioning signal. Since UniT tokens are discrete by construction, they naturally match the token prediction objective used for VLM training. Formally, letting $\hat{p}_t$ denote the predicted logits over UniT code indices,

$$\hat{p}_t = f_{\text{VLM}}(o_t,\, \ell,\, q_t), \qquad \mathcal{L}_{\text{token}} = \mathrm{CE}(\hat{p}_t,\, c_t). \tag{4}$$

Flow Matching Action Generation. Continuous embodiment-specific actions are generated via a lightweight flow matching head. Given action chunk $A_t = [a_t, \ldots, a_{t+H-1}]$, time variable $\tau \sim \mathcal{U}[0, 1]$, and noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we construct the interpolated path $A_t^\tau = \tau A_t + (1 - \tau)\epsilon$. The flow head learns a velocity field V_θ conditioned on vision-language features x_t from the last layer of the VLM and the current observation encoding from the VLM visual encoder, $\mathrm{Enc}(o_t)$:

$$\mathcal{L}_{\text{fm}} = \mathbb{E}_{\tau,\epsilon}\left[\left\|V_\theta(A_t^\tau \mid x_t,\, \mathrm{Enc}(o_t),\, \tau) - (A_t - \epsilon)\right\|_2^2\right]. \tag{5}$$

The total objective combines both terms: $\mathcal{L}_{\text{VLA}} = \mathcal{L}_{\text{token}} + \lambda_{\text{fm}}\, \mathcal{L}_{\text{fm}}$. Importantly, both UniT token prediction and action generation are conditioned on the same VLM vision-language features x_t. Since UniT maps heterogeneous human and humanoid behaviors into a shared physical intent vocabulary, token prediction pulls these context features into a more unified cross-embodiment space. The action expert then generates embodiment-specific actions from this unified context, enabling cross-embodiment transfer within a single policy (Sec. 5).

3.4 WM-UniT: Cross Embodiment World Modeling via UniT

Given the current observation o_t and action condition derived from $(s_t, a_{t:t+k})$, WM-UniT models the future prediction process by generating future visual observations. Built upon the Cosmos Predict 2.5 [34] action-conditioned video generation framework (Fig. 4b), it uses action-branch UniT features as a unified conditioning interface that replaces embodiment-specific raw actions. We therefore use UniT features as the control signal for action-conditioned video generation. Specifically, given state s_t and action chunk $a_{t:t+k}$, the UniT action-branch encoder E_a produces continuous pre-quantization features $\tilde{z}_t^a = E_a(s_t, a_{t:t+k})$, which are projected through an MLP and injected via cross-attention alongside the current observation o_t. The generation model is trained with flow matching (analogous to Sec. 3.3), with the velocity field applied to latent future frames X:

$$\mathcal{L}_{\text{WM}} = \mathbb{E}_{\tau,\epsilon}\left[\left\|V_\phi(X_t^\tau \mid o_t,\, \mathrm{MLP}(\tilde{z}_t^a),\, \tau) - (X_t - \epsilon)\right\|_2^2\right]. \tag{6}$$

Here, X_t denotes the latent representation of future frames. For long-horizon evaluation, WM-UniT supports autoregressive rollout by feeding generated frames back as observations for subsequent steps.

UniT uses vision as the anchor to map human and humanoid actions into a shared intent space, so the action features $\tilde{z}_t^a$ provide an embodiment-agnostic control interface while naturally carrying visual-dynamics priors learned during tokenization. The world model can therefore use UniT features as a unified conditioning signal across embodiments, enabling transfer in visual generation from human data to humanoid prediction. Because the action branch itself takes only state and action as input, this conditioning does not leak future observations at deployment time.

4 Experimental Setups

4.1 Benchmarks and Datasets

4.1.1 RoboCasa GR1 Tabletop Simulation

We evaluate on the RoboCasa benchmark with the GR1 humanoid robot [37]. The evaluation suite comprises 24 tabletop tasks: 18 pick-and-place rearrangement tasks where the robot follows language

instructions to move objects between containers, and 6 articulated tasks that involve more complex interactions such as placing objects inside and subsequently closing cabinets, drawers, or microwaves. Each task is assessed over 50 episodes in simulation.

Data Configurations. We evaluate under two data regimes. **Full Data** uses 24,000 robot trajectories (1,000 per task). **Few-Shot** uses a 10% subset of 2,400 trajectories (100 per task).

Learning from Human Data. To examine human-to-humanoid transfer for both policy learning and world modeling, we incorporate the `basic_pick_place` subset of the EgoDex dataset [38], containing 27,419 trajectories of pick-and-place interactions. The human data is combined with the few-shot robot set for co-training, followed by fine-tuning exclusively on robot data.

Generalization Scenarios. We construct three **generalization** test suites from RoboCasa assets: (1) *Unseen Appearance (18 Tasks)*, novel visual textures on familiar container and object pairs; (2) *Unseen Combinations (23 Tasks)*, seen objects in novel container pairings (14 pick-and-place, 9 articulated); (3) *Unseen Object Types (32 Tasks)*, novel object categories.

4.1.2 DROID Dataset

The DROID dataset [39] contains 95,599 diverse trajectories collected from 564 scenes, including approximately 76k successful and 19k failed trajectories. The diversity of actions and scenes provides a comprehensive testbed for evaluating action-conditioned world models.

4.1.3 Real-World Experiments

We validate on the IRON-R01-1.11 humanoid with a 50-dimensional action space covering arms, hands, waist, head, and wrist poses. As shown in Fig. 5, we design two real-world tasks corresponding to EgoDex subsets: **Pick & Place** (analogous to `basic_pick_place`, 27,419 trajectories) requires picking an object and placing it into a box, and **Pouring** (analogous to `pour`, 3,205 trajectories) requires bimanual grasping and pouring. We collect 120 robot trajectories per task. All models are pre-trained on a mixture of 32k proprietary robot trajectories and 30k EgoDex trajectories, then fine-tuned on task-specific data.

Generalization Scenarios. We construct five OOD axes (Fig. 6). For the first four, robot teleoperation covers part of the relevant variation, while EgoDex human demonstrations provide complementary conditions not sufficiently covered in robot data. We therefore co-train on both sources and evaluate on the human-introduced conditions. *(a) Geometry:* human data introduce the same affordance with new 3D shapes, such as cups of varying diameters in place of a bowl. *(b) Distractor:* human demonstrations include additional objects around the target manipulation object, creating distractor-rich scenes. *(c) Target:* human data introduce alternative placement destinations while the manipulation object remains the same. *(d) Background:* human data vary the table texture or background surface appearance. Beyond transfer, *(e) Combinational* evaluates instruction following among multiple target manipulation objects that are all seen during training, requiring the robot to disambiguate the correct object from the instruction.

4.2 Evaluated Variants

Policy Variants and Baselines. In addition to **VLA-UniT** (our full model), we consider a **GR00T baseline** denoted as GR00T-Qwen2.5, which uses the same architecture as VLA-UniT, but without UniT token prediction. We further evaluate three tokenizer ablation variants corresponding to the paradigms in Fig. 2. All variants share the same VLA architecture and differ only in the token prediction target: **VLA-Action** predicts action-only latent tokens without visual anchoring; **VLA-Vision** predicts vision-only latent tokens without motor information; and **VLA-UniT w/o Cross-Recon** predicts tokens from a tokenizer that encodes both modalities but without the cross-reconstruction objective, treating vision and action as decoupled vocabularies. We additionally compare **VLA-Villa**, which replaces UniT with a tokenizer re-implemented following the design of Villa-X [23] on our codebase, adopting a unidirectional vision-to-action (V2A) reconstruction objective rather than bidirectional cross-reconstruction. For policy learning, we compare VLA-UniT against representative methods spanning diffusion-based control, flow-matching policies, and action-token prediction paradigms: **Diffusion Policy** [25], which models actions via U-Net denoising;

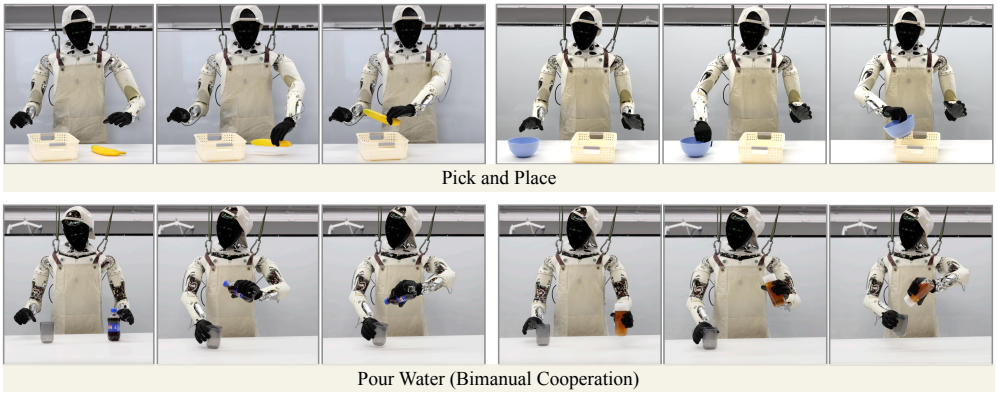

Figure 5: **Real-world in-domain tasks.** We design two tasks analogous to EgoDex subsets: **Pick & Place** (pick an object and place it into a box) and **Pouring** (grasp a bottle and a cup, then pour).

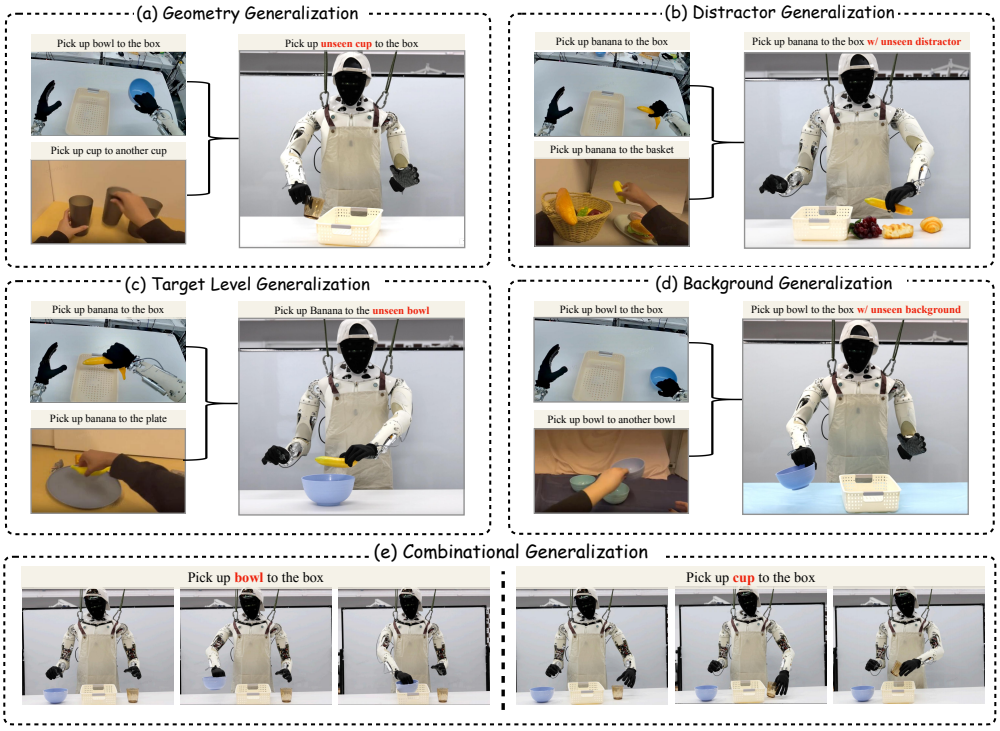

Figure 6: **Real-world OOD generalization scenarios. (a)–(d)**: human demonstrations introduce complementary variations absent from robot data; we co-train and evaluate on human-introduced conditions. **(a)** Geometry (new shapes, same affordance). **(b)** Target (new placement destinations). **(c)** Distractor (unseen objects introduced as distractors). **(d)** Visual (unseen surfaces). **(e)** Combinational (instruction-based disambiguation among multiple objects seen during training).

UWM [40], a transformer unifying action and video diffusion; **FLARE** [41], a flow-matching framework with future latent alignment; **GR00T-N1.6** [42], an upgraded GR00T variant with a larger DiT backbone; **GR00T-Qwen3** [26], combining a frozen VLM with flow-matching action generation; π**-Qwen3** [28], coupling per-layer VLM features with a flow-matching expert; **FAST-Qwen3** [5], using frequency-based BPE tokenization for autoregressive prediction; and **OFT-Qwen3** [27], an optimized fine-tuning recipe with parallel action-chunk decoding.

World Modeling Variants and Metrics. For world modeling, we compare three conditioning paradigms: **Raw Action**, which conditions the world model on embodiment-specific raw actions; **WM-Action**, which uses action-only latent tokens without visual anchoring; and **WM-UniT**, which

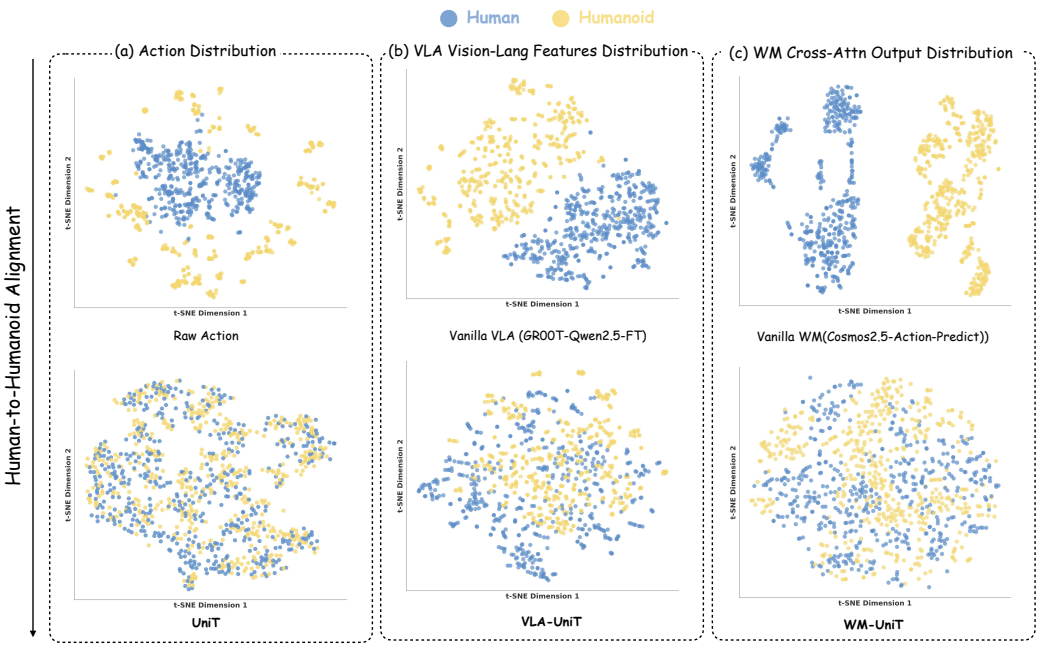

Figure 7: **Cross-embodiment representation alignment.** We plot t-SNE of human (blue) and humanoid (yellow) data across three levels: **(a)** raw actions vs. UniT token embeddings, **(b)** mean-pooled VLA vision-language hidden states, **(c)** mean-pooled WM cross-attention context embeddings. Vanilla baselines (top) show clearly separated distributions, while UniT-based models (bottom) produce highly overlapping representations, confirming that visual-anchored tokenization induces cross-embodiment alignment from the token level all the way to the internals of downstream models.

uses UniT's visually-anchored tokens (Sec. 3.4). We evaluate generation quality with PSNR, SSIM, LPIPS [43], FVD [44], and EPE. These metrics capture frame fidelity, perceptual similarity, video realism, and controllability, respectively. Here, EPE denotes End-Point Error computed from optical flow.

5 Experiments

We validate UniT through the following questions:

- **Q1 (Unified Representation):** Does UniT establish a shared physical language that aligns heterogeneous embodiments and remains robust to noise? (Sec. 5.1)
- **Q2 (Efficient Policy Learning):** Does UniT enable data-efficient and generalizable policy learning for humanoids? (Sec. 5.2.1, 5.2.2, 5.2.3)
- **Q3 (Effective World Modeling):** Does UniT conditioning improve action-conditioned generation and enable cross-embodiment dynamics transfer? (Sec. 5.3.1, 5.3.2)
- **Q4 (Design Soundness):** How does UniT's visual-anchored design compare to alternative tokenizer paradigms? (Sec. 5.4)

5.1 Unified Representation: Alignment and Robustness

We first examine whether UniT establishes the *unified physical language* claimed in Sec. 3.2, and whether this alignment propagates into downstream models. We perform t-SNE [45] analysis on human and humanoid samples from the RoboCasa GR1 and EgoDex co-training mixture (Fig. 7). For downstream VLA, we compare against GR00T-Qwen2.5-FT, which uses Qwen2.5-VL as backbone and fine-tunes the core language modeling blocks to predict raw action. For downstream world modeling, we compare against Cosmos Predict 2.5 with raw action conditioning. Both baselines are trained on the same human-humanoid data mixture.

Cross-Embodiment Token Alignment. We compare the distributions of raw action trajectories and UniT token embeddings (Fig. 7a). In the raw action space, human and humanoid data form

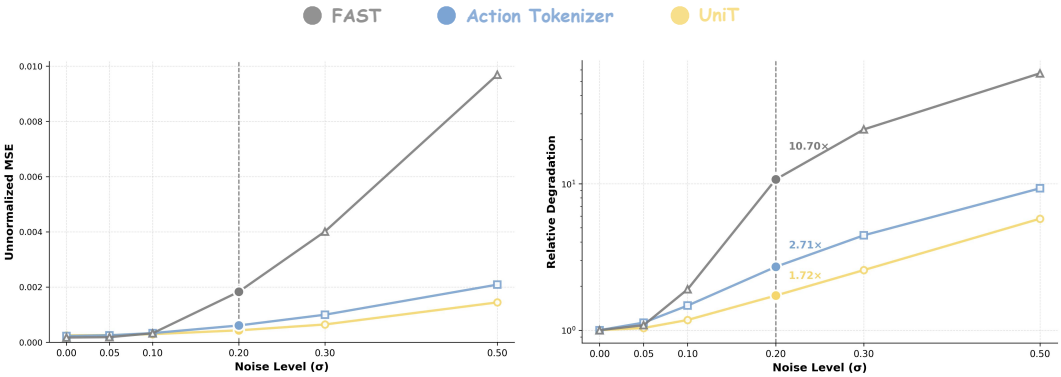

Figure 8: **Robustness to action noise.** Gaussian noise of varying intensity σ is injected into EgoDex action trajectories. Here, σ denotes the relative noise level normalized by the global action standard deviation of the dataset, so the injected noise magnitude is $\sigma \times$ global std. Reconstruction quality is evaluated using mean squared error (MSE) between reconstructed actions and the original clean actions. **(Left)** Absolute (unnormalized) MSE, zoomed in to $\sigma \leq 0.5$. **(Right)** Relative degradation, defined as $\text{MSE}_{\text{noisy}}/\text{MSE}_{\text{clean}}$, shown on a log scale. At $\sigma = 0.2$ (vertical reference line), FAST degrades by $10.7\times$, the action-only tokenizer by $2.7\times$, and UniT by only $1.7\times$, showing that visual grounding in UniT provides effective denoising.

clearly separated clusters, reflecting the inherent distribution gap between heterogeneous kinematics. After encoding through UniT, the token embeddings become highly overlapping, confirming that the visual-anchored cross-reconstruction (Sec. 3.2) successfully projects disparate action spaces into a shared manifold.

Robustness to Action Noise. In-the-wild human motion capture data inevitably contains noise from sensor jitter and annotation artifacts. We evaluate whether the cross-reconstruction mechanism in UniT (Sec. 3.2), which grounds actions in visual transitions, provides implicit denoising. We inject Gaussian noise of varying intensity σ into EgoDex action trajectories, where σ denotes the relative noise level normalized by the global action standard deviation of the dataset, and compare the reconstruction quality of three tokenizers: FAST [5], a frequency-based BPE action tokenizer; Action Tokenizer [4], which uses the same RQ-VAE architecture as UniT but is trained on action data alone; and UniT, which jointly leverages visual and action information. Reconstruction quality is measured by mean squared error (MSE) between reconstructed actions and the original clean actions. As shown in Fig. 8, UniT exhibits the lowest absolute MSE (left) and the lowest relative degradation $\text{MSE}_{\text{noisy}}/\text{MSE}_{\text{clean}}$ (right) across noise levels. At $\sigma = 0.2$, FAST degrades by $10.7\times$ and the action-only tokenizer by $2.7\times$, while UniT degrades by only $1.7\times$, maintaining near-clean reconstruction quality. The gap between UniT and the action-only tokenizer confirms that visual grounding provides complementary information that regularizes the latent space, filtering out kinematic noise that lacks visual correspondence and yielding more robust representations for downstream deployment.

Downstream Representation Alignment. We further examine whether token-level alignment propagates into downstream model internals.For VLA, we extract mean-pooled last-layer vision-language features after the action expert self-attention; for WM, we extract mean-pooled last-layer cross-attention outputs after UniT feature injection (Sec. 3.4). The vanilla VLA (GR00T-Qwen2.5VL-FT, an LLM-tuned GR00T variant based on Qwen2.5-VL without UniT token prediction supervision) and vanilla WM (Cosmos Predict 2.5) using raw actions as conditioning follow the same architectures introduced in Sec. 3.3 and Sec. 3.4. As shown in Fig. 7(b), the vanilla VLA maintains clearly separated human and humanoid feature distributions, while VLA-UniT produces substantially more interleaved representations. The effect is even more pronounced in world modeling (Fig. 7c): the vanilla WM exhibits fully disjoint clusters, whereas WM-UniT brings them into a single unified distribution. These results confirm that UniT tokens not only form a shared cross-embodiment latent space, but also induce downstream models to develop embodiment-agnostic internal representations, providing a structural basis for the performance gains observed in the following sections.

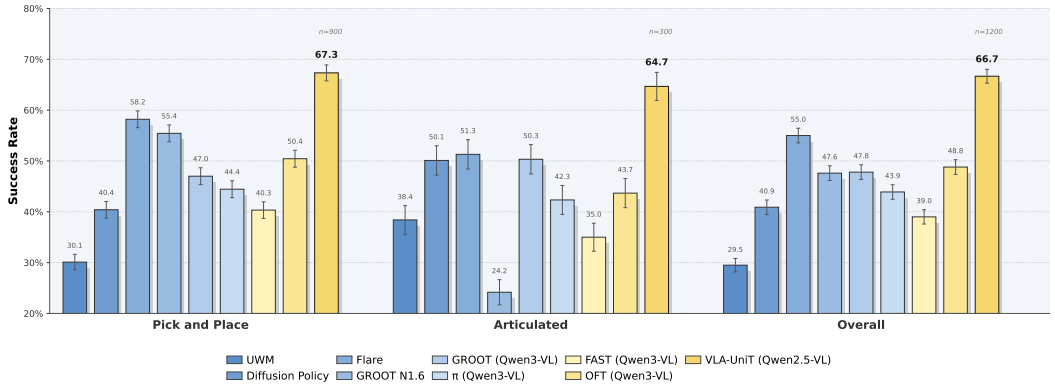

Figure 9: **Overall performance** on RoboCasa GR1 Tabletop Simulation with full training data. VLA-UniT achieves the strongest overall policy performance among all compared methods.

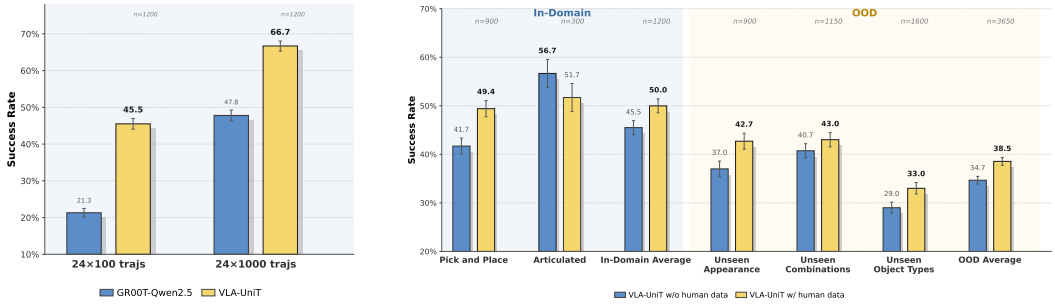

Figure 10: **(Left)** Data efficiency results on RoboCasa GR1 under full-data and few-shot settings. **(Right)** Impact of incorporating EgoDex `basic_pick_place` human demonstrations on few-shot performance in RoboCasa GR1. UniT improves both sample efficiency and the utility of human co-training.

5.2 Policy Learning

5.2.1 Benchmark Performance and Data Efficiency

As established in Sec. 3.3, UniT token prediction provides the VLM with a compact, visually-anchored learning objective that encodes physical intent. We evaluate whether this translates to improved policy performance and data efficiency on the RoboCasa GR1 benchmark, compared against the policy baselines described in Sec. 4.2.

Overall Performance. As shown in Fig. 9, VLA-UniT achieves a 66.7% overall success rate on the full-data RoboCasa benchmark, outperforming all baselines by a substantial margin. In Pick & Place tasks, which require grounding semantic instructions into precise object rearrangement behaviors, VLA-UniT attains 67.3%. In Articulated tasks, where the robot must manipulate articulated fixtures such as cabinets and microwaves through multi-step interactions, VLA-UniT reaches 64.7%, maintaining consistently strong performance across both categories. VLA-UniT surpasses all baselines, outperforming the previous best FLARE (55.0%) by 11.7%. Notably, compared to the GR00T baseline (47.8%), which shares the same architecture without UniT token prediction, the improvement of 18.9% highlights the value of introducing UniT as a learning objective (Sec. 3.3). By jointly grounding visual transitions and motor commands through cross-reconstruction (Sec. 3.2), UniT tokens provide the VLM with a compact prediction target that strengthens visuo-motor synergy, enabling more effective policy learning across diverse task types.

Data Efficiency. To assess sample efficiency, we compare VLA-UniT and the GR00T baseline under both full-data and few-shot regimes (Fig. 10, left). With only 10% of the training data (100 trajectories per task), VLA-UniT achieves 45.5% success rate, already approaching the GR00T baseline trained on full data (47.8%). This $\sim$10$\times$ reduction in data requirements highlights the

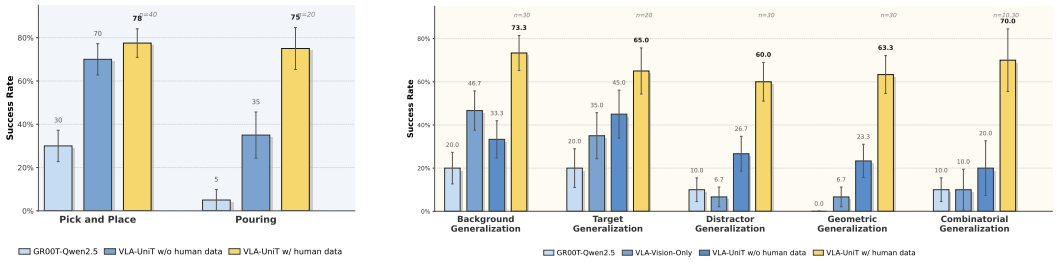

Figure 11: **(Left)** In-domain results on the IRON-R01-1.11 robot. **(Right)** OOD generalization results on the IRON-R01-1.11 robot. Human Demonstration improves VLA-UniT's real-world execution and OOD robustness.

advantage of UniT's compact, cross-modality aligned representation: by operating in a structured discrete latent space rather than regressing raw actions, the VLM extracts task-relevant intent more efficiently from limited demonstrations.

5.2.2 Human-to-Humanoid Transfer

We investigate whether UniT's shared latent space enables leveraging large-scale human demonstrations to improve humanoid policy learning. Under the few-shot regime in simulation, we co-train VLA-UniT on robot data and EgoDex human demonstrations, then fine-tune on robot data alone (Fig. 10, right).

Incorporating human data improves both in-domain and OOD performance. The in-domain average performance increases from 45.5% to 50.0%, with the largest gain in Pick & Place (41.7% $\rightarrow$ 49.4%), which directly corresponds to the EgoDex `basic_pick_place` domain. Across all three OOD scenarios, human co-training brings consistent improvements: Unseen Appearance (37.0% $\rightarrow$ 42.7%), Unseen Combinations (40.7% $\rightarrow$ 43.0%), and Unseen Object Types (29.0% $\rightarrow$ 33.0%), yielding an OOD average gain from 34.7% to 38.5%. These results confirm that UniT's shared latent space enables effective human-to-humanoid transfer. As shown in Sec. 5.1, VLA-UniT produces highly overlapping vision-language representations for human and humanoid data (Fig. 7b), indicating that the VLM builds a unified internal feature space across embodiments. This representational alignment provides the structural basis for transfer: human demonstrations, mapped into the same token vocabulary, broaden the VLM's task coverage and directly benefit humanoid policy generalization even in out-of-distribution settings.

5.2.3 Real-World Generalization

We deploy VLA-UniT on the real-world IRON-R01-1.11 humanoid to validate whether the performance, human-to-humanoid transfer, and generalization gains observed in simulation also carry over to physical deployment.

In-Domain Performance. We evaluate VLA-UniT on two real-world tasks (Fig. 11, left). With robot data alone, VLA-UniT already substantially outperforms the GR00T baseline on both Pick & Place (70% vs. 30%) and Pouring (35% vs. 5%). Incorporating EgoDex human data further improves performance to 78% and 75% respectively, with the gain particularly pronounced in Pouring, a task requiring coordinated dual-arm control that benefits from the rich bimanual interaction patterns in human demonstrations. UniT's shared latent space enables the VLM to directly leverage human bimanual coordination experience for humanoid execution.

OOD Generalization. As described in Sec. 4.1.3, the first four OOD axes are designed so that robot data provide partial coverage while human demonstrations introduce the complementary variation; we evaluate on the human-introduced conditions. Across all five categories (Fig. 11, right), VLA-UniT with human co-training consistently achieves the strongest performance. In Geometry and Distractor Generalization scenarios — where human videos introduce novel object shapes and visual clutter respectively — the improvement is most pronounced (23.3%$\rightarrow$63.3% and 26.7%$\rightarrow$60.0%), confirming that human data effectively fills the variation gap left by limited robot demonstrations. Background and Target-Level Generalizations show similar trends, indicating that the relevant visual and goal-conditioned knowledge also transfers through UniT's shared space. Notably, VLA-

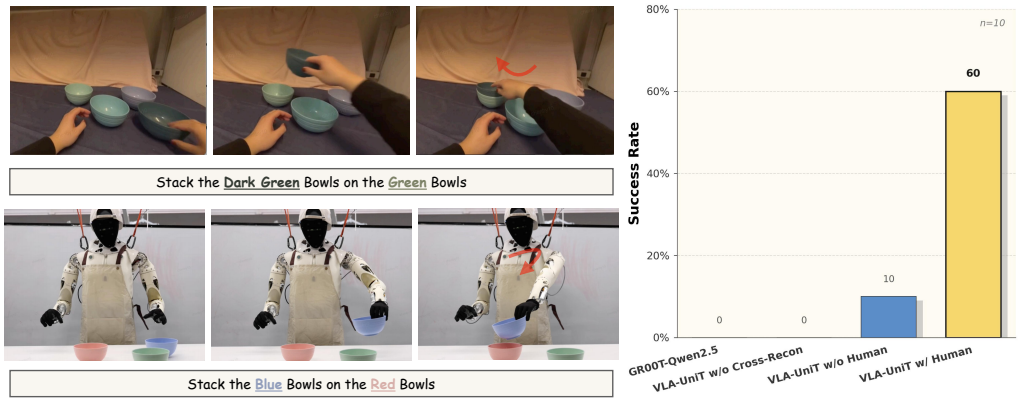

Figure 12: **Zero-shot task transfer** on the IRON-R01-1.11 robot. (Left) Task illustration. (Right) Success rates on the unseen stacking task. UniT with human co-training shows the clearest transfer to the unseen task.

Table 1: **Controllable generation** results on DROID and EgoDex + RoboCasa-GR1 co-training.

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	EPE ↓
Single-embodiment (DROID)						
DROID	Raw Action	21.02	0.820	0.097	**76.38**	0.2662
	WM-Action	20.86	0.819	0.102	80.30	0.2593
	WM-UniT	**21.32**	**0.823**	**0.095**	76.44	**0.2588**
Human-humanoid co-training (EgoDex + RoboCasa-GR1)						
EgoDex	Raw Action	24.84	0.800	0.164	171.37	0.706
	WM-UniT	**28.06**	**0.858**	**0.086**	**130.87**	**0.519**
RoboCasa-GR1	Raw Action	13.45	0.590	0.259	237.13	0.558
	WM-UniT	**17.66**	**0.718**	**0.142**	**166.50**	**0.453**

Vision shows limited OOD robustness in these scenarios, confirming the limitations of vision-only representations discussed in Sec. 3.2. The Combinational setting, which tests language-guided disambiguation, further improves from 10% to 70%, suggesting that the broader interaction diversity from human co-training also strengthens compositional generalization.

Zero-Shot Task Transfer. We evaluate on a stacking task that is not covered by robot training demonstrations (Fig. 12). Robot data only include pick-and-place of individual bowls, while EgoDex human videos contain stacking sequences performed with view switching and upper-body coordination. The GR00T baseline and VLA-UniT without cross-reconstruction both score 0%, indicating that neither raw action fitting nor tokenization without cross-modal alignment can bridge the task gap. VLA-UniT without human data achieves 10%. With human co-training, VLA-UniT reaches 60%, transferring the stacking logic from human demonstrations and exhibiting emergent upper-body coordination — waist rotation and head turning to adjust viewpoint — that mirrors the coordination patterns observed in human videos. This demonstrates that UniT's visual-anchored cross-reconstruction creates a representational bridge strong enough to transfer not just task semantics but also fine-grained coordination patterns across embodiments.

5.3 World Modeling

5.3.1 Controllable Generation

As described in Sec. 3.4, WM-UniT conditions video generation on UniT's continuous pre-quantization features. We then examine whether this conditioning interface improves controllability and cross-embodiment transfer under a fixed video generation backbone. In all world-model experiments, we evaluate 10-step autoregressive rollouts over 10-second video generation; DROID is evaluated at resolution 192×320, while RoboCasa-GR1 and EgoDex are evaluated at 192×336. We compare UniT Tokens against raw actions and action-only latent tokens.

Table 2: **Human data pre-training** for world modeling. Pre-trained on EgoDex `basic_pick_place`, fine-tuned on RoboCasa-GR1 pick-and-place. Human pre-training through UniT improves downstream humanoid controllability.

Configuration	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	EPE ↓
WM-UniT w/o Human Pre-training	16.34	0.678	0.168	180.51	0.478
WM-UniT (Full)	**18.06**	**0.713**	**0.135**	**153.31**	**0.446**

As shown in Table 1, WM-UniT achieves the strongest controllability on DROID, as reflected by the best EPE. It also improves PSNR, SSIM, and LPIPS over both baselines, while remaining competitive on FVD. In contrast, WM-Action does not yield a similarly reliable gain, indicating that latent tokenization alone is insufficient for controllable video generation. The results demonstrate that UniT provides a more compact and more faithful conditioning signal for world modeling.

5.3.2 Human-Humanoid Transfer

We further investigate whether UniT's shared latent space enables human-to-humanoid dynamics transfer for world modeling. As established in Sec. 5.1, WM-UniT produces unified context embeddings for human and humanoid data (Fig. 7c), suggesting that the world model builds a shared internal dynamics representation across embodiments.

Co-Training. Under joint training on EgoDex and RoboCasa-GR1 (Table 1), WM-UniT consistently outperforms Raw Action on both the human and humanoid subsets. The gain is reflected not only in reconstruction quality, but more importantly in stronger controllability, indicating that UniT provides a shared conditioning space that supports cross-embodiment dynamics modeling. Combined with the aligned context embeddings in Fig. 7(c), this result shows that UniT enables the world model to co-train on human and humanoid data without collapsing into embodiment-specific dynamics.

Pre-Training. We pre-train WM-UniT on EgoDex `basic_pick_place` human demonstrations, then fine-tune on RoboCasa-GR1 pick-and-place data (Table 2). Human pre-training brings consistent gains across all metrics, with the most meaningful improvement reflected in controllability. This indicates that the dynamics learned from human data remain usable after transfer to humanoid prediction, rather than being tied to human-specific kinematics. Together with the shared representation analysis in Sec. 5.1, these results confirm that UniT provides a transferable dynamics interface for world modeling.

Cross-Embodiment Conditioning. Beyond co-training and pre-training, we directly test whether UniT tokens from one embodiment can condition video generation for the other — without any domain-specific adaptation. We condition the world model with actions from a source embodiment and generate videos for the target embodiment, comparing UniT against Raw Action conditioning (Fig. 14, 13).

In the human-to-robot setting (Fig. 13), the human reference includes a reach, tip-down, and grasp sequence, as well as varying magnitudes — a slight exploratory reach, a significant forward extension, and a slight retraction. UniT-conditioned generation distinguishes between these scales and preserves the non-monotonic trajectory (reach then retract), while Raw Action produces uniform motion that neither reflects magnitude differences nor captures directional reversals.

In the robot-to-human setting (Fig. 14), the robot reference shows a multi-phase sequence — forward approach, vertical descent, and internal wrist rotation before grasping. UniT-conditioned generation faithfully reproduces each phase in the human domain, preserving both the grasp semantics and fine-grained pose adjustments such as the terminal rotation and tip-down motion. Raw Action conditioning captures the coarse trajectory but collapses these atomic actions into a flat reach, losing the rotational and postural details.

Together, these examples highlight three capabilities of UniT's cross-embodiment encoding: (1) *fine-grained action semantics* — atomic actions such as internal rotation, tip-down, and grasp transfer across embodiments; (2) *magnitude sensitivity* — the distinction between slight and significant motions is preserved; and (3) *temporal coherence* — non-monotonic trajectories including retraction are faithfully reproduced.

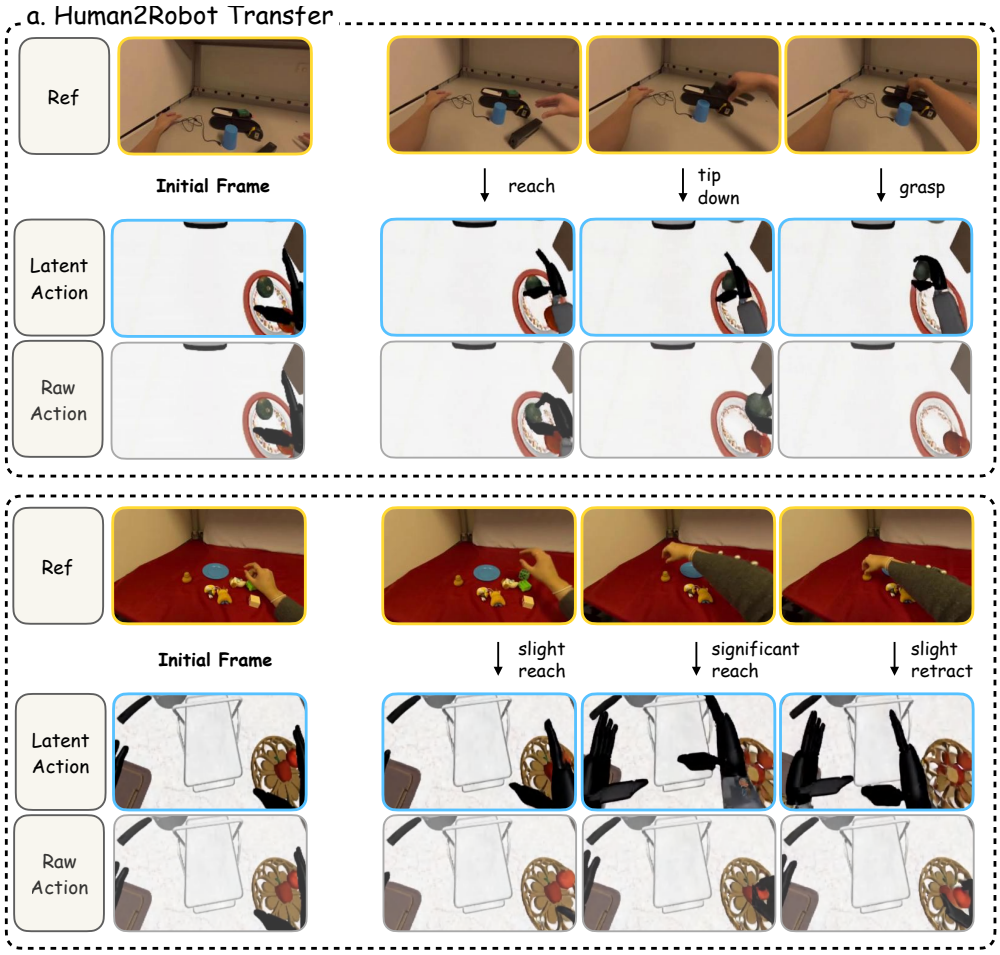

Figure 13: **Human-to-robot conditioning.** Top: human reference video. Bottom: human reference actions condition robot video generation via UniT vs. Raw Action. UniT yields more faithful cross-embodiment conditioning than raw actions.

Table 3: **Cross-embodiment conditioning consistency.** MLLM-based evaluation (Gemini-3-Pro) on EgoDex and RoboCasa-GR1. UniT improves semantic, temporal, and geometric consistency in both transfer directions.

Method	Semantic ↑	Temporal ↑	Geometric ↑	Overall ↑
Robot-to-Human				
Raw Action	2.96	3.12	2.74	2.92
WM-UniT	**3.91**	**3.98**	**3.66**	**3.84**
Human-to-Robot				
Raw Action	2.98	3.16	2.72	2.95
WM-UniT	**3.28**	**3.43**	**3.09**	**3.27**

To quantify these observations, we evaluate cross-embodiment conditioning consistency using Gemini-3-Pro as an automated judge. For each generated video paired with its reference, the model scores three dimensions on a 1–5 scale: *Semantic* consistency (whether the intended action is preserved), *Temporal* consistency (whether the motion timing and sequencing match), and *Geometric* consistency (whether spatial trajectories and pose details are faithful).

As shown in Table 3, WM-UniT consistently outperforms Raw Action across all three evaluation dimensions in both directions. The Semantic score reflects the preservation of action intent (grasp, rotate, retract); the Temporal score captures sequencing fidelity and the ability to reproduce non-monotonic trajectories; and the Geometric score measures spatial precision including magnitude

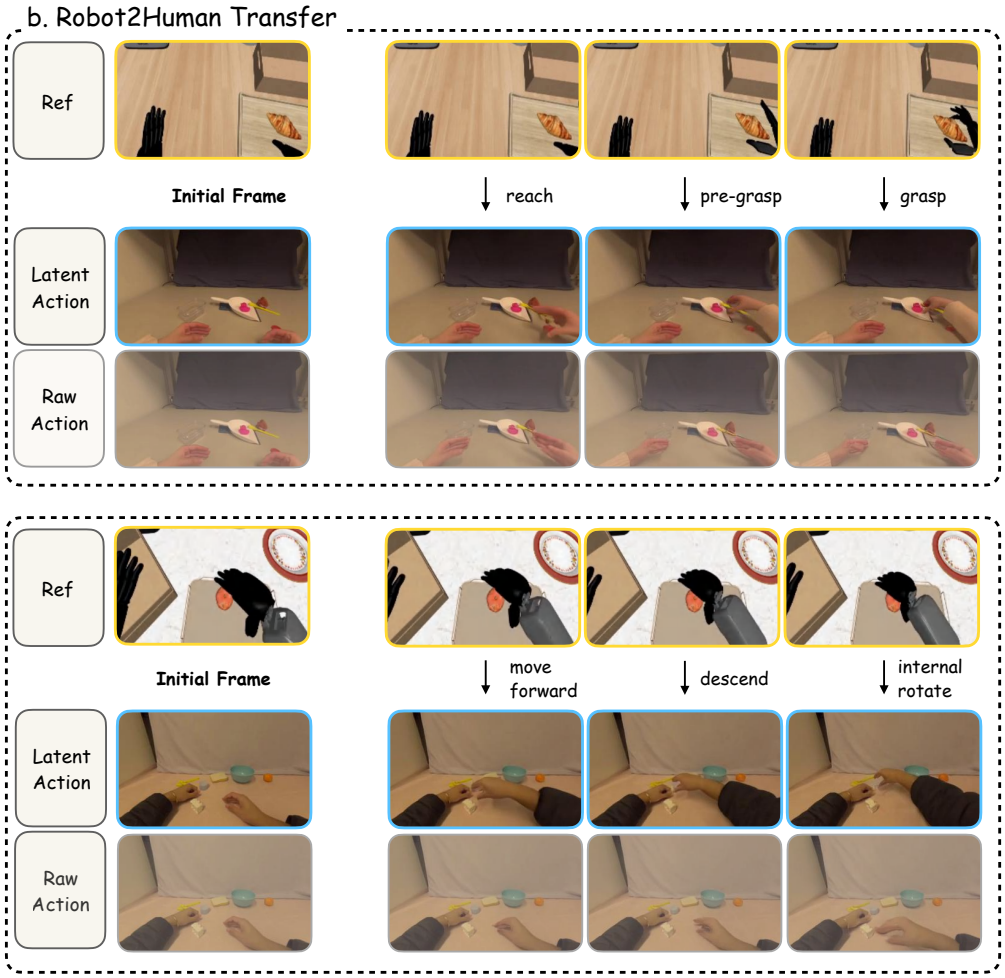

Figure 14: **Robot-to-human conditioning.** Top: robot reference video. Bottom: robot reference actions condition human video generation via UniT vs. Raw Action. UniT yields more faithful cross-embodiment conditioning than raw actions.

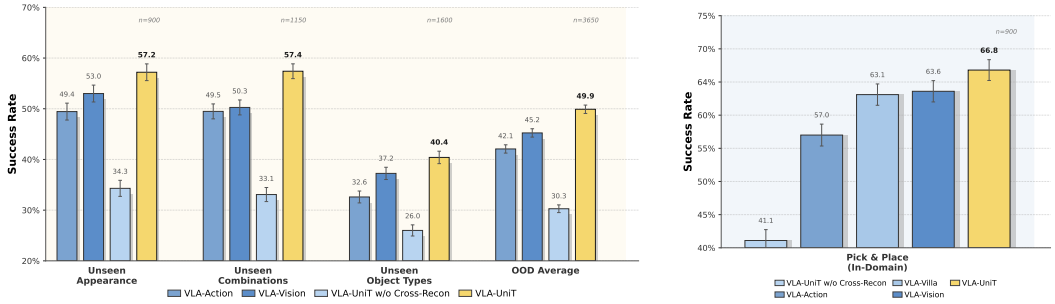

Figure 15: **Tokenizer paradigm ablation** on RoboCasa GR1 (full data with human co-training). **(Left)** OOD generalization. **(Right)** In-domain Pick & Place performance.

sensitivity and pose detail. WM-UniT's consistent improvement across all three confirms that it encodes a precise and transferable action representation for cross-embodiment world modeling.

5.4 Tokenizer Design Ablation

UniT's design rests on two key claims made in Sec. 3.2: (1) both vision and action are needed — action-only methods suffer cross-embodiment distribution misalignment without visual grounding,

while vision-only methods entangle low-level appearance and miss fine-grained motor detail; and (2) the two modalities must be explicitly aligned through cross-reconstruction, rather than treated as disconnected vocabularies. We validate both claims under the human-humanoid co-training setup (EgoDex + RoboCasa pre-train, RoboCasa fine-tune), where the tokenizer's ability to bridge embodiments is directly tested. We compare paradigms corresponding to the architectures in Fig. 2.

Vision-Action Synergy Enables Transfer. As shown in Fig. 15 (left), VLA-UniT (OOD average 49.9%) consistently outperforms both single-modality variants across all OOD scenarios. VLA-Vision (45.2%) provides a transferable visual signal but lacks fine-grained motor detail; VLA-Action (42.1%) captures motor intent but struggles with the cross-embodiment distribution gap without visual grounding. The joint encoding of both modalities in VLA-UniT combines the embodiment-invariant nature of vision with the precision of action, forming a more complete *unified physical language* for cross-embodiment transfer.

Cross-Reconstruction Produces Aligned Representations. VLA-UniT w/o Cross-Recon (30.3%) falls below even the single-modality variants despite having access to both modalities, showing that multi-modal input alone does not guarantee alignment. VLA-UniT's cross-reconstruction objective addresses this by enforcing mutual reconstruction between vision and action (Sec. 3.2), transforming disconnected modalities into a coherent shared vocabulary. The resulting 19.6% gain over the ablation without cross-reconstruction confirms that explicit cross-modal alignment is the key enabler for human-to-humanoid transfer.

Bidirectional vs. Unidirectional Reconstruction. On in-domain performance (Fig. 15, right), we further include VLA-Villa, which uses unidirectional V2A reconstruction. VLA-UniT (66.8%) consistently outperforms VLA-Villa (63.1%), confirming that bidirectional cross-reconstruction is more effective than unidirectional alternatives for producing aligned cross-embodiment tokens.

6 Conclusion and Discussion

We presented UniT, a visual-anchored latent action tokenizer that establishes a unified physical language for human-to-humanoid transfer through cross-reconstruction. In VLA-UniT, UniT improves policy performance and data efficiency, while enabling effective human-to-humanoid transfer with OOD generalization and zero-shot task transfer. In WM-UniT, UniT provides a stronger conditioning interface for cross-embodiment dynamics modeling and human-to-humanoid transfer. Ablation studies confirm that both vision-action synergy and bidirectional cross-reconstruction are essential for these gains.

Looking forward, the visual branch of UniT encodes physical transitions from observations alone, without requiring paired action annotations. This opens a path toward absorbing the vast and largely untapped reservoir of internet video, where humans perform diverse physical tasks without motor labels. Such data could serve as an additional source of physical priors that enriches the shared latent space. Furthermore, the fact that UniT serves as a unified interface for both policy and world model suggests a deeper possibility: policies can propose latent actions, world models can simulate their visual consequences, and the resulting imagined rollouts can flow back as reward signals for reinforcement learning or enable test-time planning through search over the latent space. This closed-loop co-evolution, mediated entirely within the shared token space, may be a compelling route toward scalable embodied intelligence. On the data side, UniT's data-driven alignment readily scales to broader internet-scale human motion data without manual kinematic correspondence, enabling the framework to fully leverage massive heterogeneous datasets. This scalability also holds the potential to learn upper-body coordination and dexterous control directly from diverse human demonstrations.

Acknowledgments

We thank Chuan Ma and Lu Qiu for sharing the codebases for world model experiments, and Hui Zhou for his help with the real robot infrastructure and teleoperation data.

References

[1] Xiongyi Cai, Ri-Zhao Qiu, Geng Chen, Lai Wei, Isabella Liu, Tianshu Huang, Xuxin Cheng, and Xiaolong Wang. In-n-on: Scaling egocentric manipulation with in-the-wild and on-task data. *arXiv preprint arXiv:2511.15704*, 2025.

[2] Tony Tao, Mohan Kumar Srirama, Jason Jingzhou Liu, Kenneth Shaw, and Deepak Pathak. Dexwild: Dexterous human interactions for in-the-wild robot policies. *arXiv preprint arXiv:2505.07813*, 2025.

[3] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

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

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

[6] An Dinh Vuong, Minh Nhat Vu, Dong An, and Ian Reid. Action tokenizer matters in in-context imitation learning. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13490–13496. IEEE, 2025.

[7] Atharva Mete, Haotian Xue, Albert Wilcox, Yongxin Chen, and Animesh Garg. Quest: Self-supervised skill abstractions for learning continuous control, 2024. *URL https://arxiv. org/abs/2407.15840*, 2024.

[8] Yating Wang, Haoyi Zhu, Mingyu Liu, Jiange Yang, Hao-Shu Fang, and Tong He. Vq-vla: Improving vision-language-action models via scaling vector-quantized action tokenizers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11089–11099, 2025.

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

[10] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.

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

[12] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Xindong He, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025.

[13] Yankai Fu, Ning Chen, Junkai Zhao, Shaozhe Shan, Guocai Yao, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Metis: Multi-source egocentric training for integrated dexterous vision-language-action model. *arXiv preprint arXiv:2511.17366*, 2025.

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

[15] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

[16] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pages 416–426. PMLR, 2023.

[17] Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *Advances in neural information processing systems*, 37:124420–124450, 2024.

[18] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J Yoon, Ryan Hoque, Lars Paulsen, et al. Humanoid policy˜ human policy. *arXiv preprint arXiv:2503.13441*, 2025.

[19] Chengbo Yuan, Rui Zhou, Mengzhen Liu, Yingdong Hu, Shengjie Wang, Li Yi, Chuan Wen, Shanghang Zhang, and Yang Gao. Motiontrans: Human vr data enable motion-level learning for robotic manipulation policies. *arXiv preprint arXiv:2509.17759*, 2025.

[20] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13226–13233. IEEE, 2025.

[21] Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-rdt: Human manipulation enhanced bimanual robotic manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18135–18143, 2026.

[22] Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. Emergence of human to robot transfer in vision-language-action models. *arXiv preprint arXiv:2512.22414*, 2025.

[23] Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, et al. Villa-x: enhancing latent action modeling in vision-language-action models. *arXiv preprint arXiv:2507.23682*, 2025.

[24] Shichao Fan, Kun Wu, Zhengping Che, Xinhua Wang, Di Wu, Fei Liao, Ning Liu, Yixue Zhang, Zhen Zhao, Zhiyuan Xu, et al. Xr-1: Towards versatile vision-language-action models via learning unified vision-motion representations. *arXiv preprint arXiv:2511.02776*, 2025.

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

[26] NVIDIA, Johan Bjorck, Nikita Cherniadev Fernando Castañeda, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. In *ArXiv Preprint*, March 2025.

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

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

[29] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, 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)*, pages 6892–6903. IEEE, 2024.

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

[31] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9834–9844, 2025.

[32] Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.

[33] Julian Quevedo, Percy Liang, and Sherry Yang. Evaluating robot policies in a world model. *arXiv e-prints*, pages arXiv–2506, 2025.

[34] NVIDIA, Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, Prithvijit Chattopadhyay, Mike Chen, Yongxin Chen, Yu Chen, Shuai Cheng, Yin Cui, Jenna Diamond, Yifan Ding, Jiaojiao Fan, Linxi Fan, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Ruiyuan Gao, Yunhao Ge, Jinwei Gu, Aryaman Gupta, Siddharth Gururani, Imad El Hanafi, Ali Hassani, Zekun Hao, Jacob Huffman, Joel Jang, Pooya Jannaty, Jan Kautz, Grace Lam, Xuan Li, Zhaoshuo Li, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Yen-Chen Lin, Huan Ling,

Ming-Yu Liu, Xian Liu, Yifan Lu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Seungjun Nah, Yashraj Narang, Abhijeet Panaskar, Lindsey Pavao, Trung Pham, Morteza Ramezanali, Fitsum Reda, Scott Reed, Xuanchi Ren, Haonan Shao, Yue Shen, Stella Shi, Shuran Song, Bartosz Stefaniak, Shangkun Sun, Shitao Tang, Sameena Tasmeen, Lyne Tchapmi, Wei-Cheng Tseng, Jibin Varghese, Andrew Z. Wang, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Jiashu Xu, Dinghao Yang, Xiaodong Yang, Haotian Ye, Seonghyeon Ye, Xiaohui Zeng, Jing Zhang, Qinsheng Zhang, Kaiwen Zheng, Andrew Zhu, and Yuke Zhu. World simulation with video foundation models for physical ai, 2025.

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

[36] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[37] NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots, 2025.

[38] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

[39] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. *RSS*, 2024.

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

[41] Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, et al. Flare: Robot learning with implicit world modeling. *arXiv preprint arXiv:2505.15659*, 2025.

[42] NVIDIA GEAR Team, Allison Azzolini, Johan Bjorck, Valts Blukis, et al. Gr00t n1.6: An improved open foundation model for generalist humanoid robots. `https://research.nvidia.com/labs/gear/gr00t-n1_6/`, December 2025.

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[44] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

[45] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.