

Morphology-Consistent Humanoid Interaction through Robot-Centric Video Synthesis

Weisheng Xu^{1*}, Jian Li^{1*}, Yi Gu¹, Bin Yang¹, Haodong Chen²,
Shuyi Lin³, Mingqian Zhou⁴, Jing Tan¹, Qiwei Wu¹, Xiangrui Jiang¹,
Taowen Wang¹, Jiawen Wen¹, Qiwei Liang¹, Jiaxi Zhang¹, Renjing Xu^{1†}

¹Hong Kong University of Science and Technology (Guangzhou)

²Harbin Institute of Technology, Shenzhen ³Shenzhen University ⁴University of Cambridge

wxu421@connect.hkust-gz.edu.cn, renjingxu@hkust-gz.edu.cn

Project Page: <https://wesleyxu224.github.io/Dream2Act/>

Abstract

Equipping humanoid robots with versatile interaction skills typically requires either extensive task-specific policy training or explicit human-to-robot motion retargeting. However, learning-based policies are hindered by prohibitive data collection costs, limiting their scalability. Meanwhile, retargeting paradigms rely heavily on human-centric pose estimation (e.g., SMPL), which inevitably introduces the morphological gap. Such skeletal scale mismatches result in severe spatial misalignments when mapped to robots, compromising interaction success. In this work, we propose Dream2Act, a robot-centric framework that enables zero-shot interaction through generative video synthesis. Given an image of the target robot and target object in third-person view, our framework leverages video generation models to synthesize video sequences in which the physical robot completes the task with spatially aligned, morphology-consistent motion. We then employ a robust, high-fidelity pose extraction system to recover physically feasible, robot-native joint trajectories from these synthesized “dreams”, which are subsequently executed via a general-purpose whole-body controller. By operating strictly within the robot-native coordinate space, Dream2Act effectively avoids traditional retargeting errors and eliminates the need for task-specific policy training. We evaluate Dream2Act on the Unitree G1 across four categories of whole-body mobile interaction tasks (multi-position ball kicking, sofa sitting, bag punching, and box hugging). Dream2Act achieves an overall task success rate of 37.5%, compared to 0% for conventional retargeting

pipelines. Retargeting frequently fails to establish correct physical contacts due to the morphological gap, with errors further compounded under locomotion. In contrast, Dream2Act maintains robot-consistent spatial alignment throughout execution, enabling reliable contact formation and substantially higher task completion.

1. Introduction

The ultimate goal of embodied artificial intelligence is to deploy humanoid robots capable of executing general-purpose interaction tasks in unstructured human environments. To achieve this, acquiring dexterous interaction skills from demonstrations has emerged as a promising paradigm [9, 12, 37]. However, capturing high-quality humanoid demonstrations conventionally relies on either expensive hardware-intensive teleoperation [10] or precise whole-body pose estimation for human-to-robot mapping [26]. Both pipelines are constrained by high hardware costs, substantial human labor, and a lack of scalability.

While leveraging large-scale, in-the-wild human video data offers a more scalable alternative for skill acquisition [24, 29], it introduces a fundamental challenge: how to bridge the physical mismatch between humans and robots.

To exploit human videos, mainstream approaches rely on human-to-robot motion retargeting [1, 19]. These paradigms typically utilize human-centric kinematic models, such as SMPL [18], to extract poses from human videos and map them onto the robot-specific embodiment.

This human-centric pipeline is hindered by the morphology gap: humans and humanoids differ in skeletal proportions, degrees of freedom, and joint limits. Forcing human kinematics onto a robotic skeleton inevitably introduces se-

*Equal contribution.

†Corresponding author.

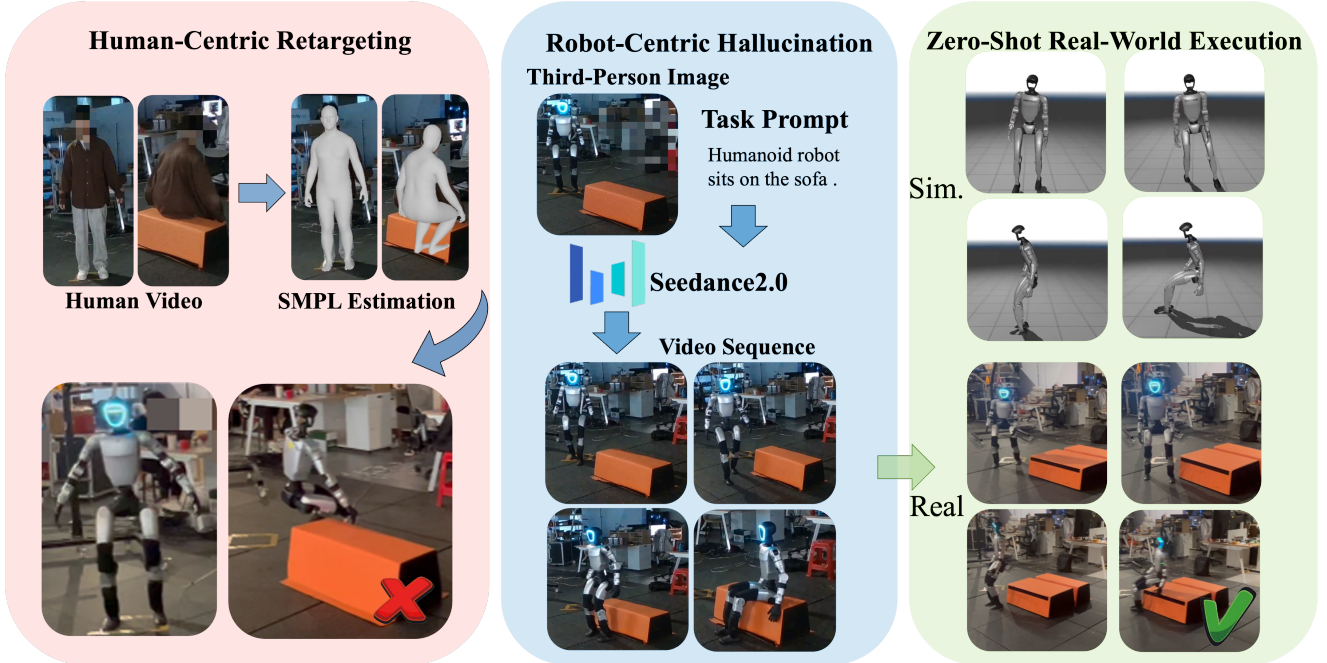


Figure 1. **Dream2Act: Zero-shot Humanoid Interaction via Robot-Centric Video Synthesis.** (A) Traditional human-centric retargeting fails due to the morphology gap, causing spatial misalignments. (B) Our Dream2Act pipeline leverages Seedance 2.0 to synthesize morphology-consistent interactions directly in the robot’s native space. (C) This visual prior translates into successful zero-shot physical execution for spatially-sensitive tasks like sitting on a sofa and kicking a ball.

vere spatial misalignments, a problem compounded by accumulated errors during locomotion. Even in fundamental whole-body mobile interactions, these scale mismatches and accumulated locomotion drifts severely warp the absolute spatial relationship between the agent and the target, consistently compromising task success.

In this work, we advocate a shift from human-centric retargeting to a strictly robot-centric alternative. We introduce Dream2Act, a novel framework that achieves zero-shot humanoid interaction through generative video synthesis. Unlike kinematic retargeting, which struggles to natively incorporate environmental geometry, recent generative video models like Seedance 2.0 [5, 27] have demonstrated the ability to implicitly capture physical dynamics and spatial-temporal consistency directly in the visual domain. Such models have been shown to encode scene affordances and inter-object relationships through large-scale visual pretraining [3, 4, 8]. Given only a single third-person image of the humanoid and an interaction object, our framework leverages powerful image-to-video diffusion priors to hallucinate realistic and temporally coherent videos of the physical robot executing the desired task.

Subsequently, we employ a high-fidelity pose estimation and reconstruction pipeline to recover robot-native kinematic trajectories directly from these synthesized sequences. Finally, these trajectories are executed through

a general-purpose whole-body controller SONIC [20]. By entirely bypassing the the intermediate human embodiment representation, Dream2Act avoids retargeting-induced errors and sidesteps the morphology gap.

Our framework enables context-aware, task-level zero-shot interaction. By conditioning the generative process on a single real-world observation and a semantic task prompt, Dream2Act achieves the ability to execute novel tasks in a training-free manner, bypassing the need for any prior exposure to specific task dynamics or environment-specific data collection. We comprehensively validate Dream2Act across four representative categories of whole-body mobile interaction tasks deployed on the Unitree G1 humanoid. Extensive physical experiments demonstrate that while conventional optimization-based retargeting baselines suffer catastrophic failure in establishing correct physical contacts, our approach reliably achieves functional spatial alignment and successful physical engagement.

Our work suggests that the large-scale generative video model [5] has internalized sufficient world physics to serve as a high-level zero-shot motion planner, paving the way for a new class of **hallucination-to-execution** paradigms in robotics. In summary, our main contributions are three-fold:

- We propose Dream2Act, a pioneering robot-centric framework that enables task-level zero-shot humanoid interaction by leveraging generative video synthesis. This

introduces a novel paradigm for whole-body robot motion acquisition, offering a scalable alternative to costly manual teleoperation.

- We introduce a robust kinematic recovery pipeline that effectively bridges the pixel-to-physics gap. By cascading native 2D pose estimation, a domain-specific 3D lifting network, and URDF-constrained inverse kinematics, our approach successfully translates generative visual priors into physically executable joint trajectories.
- We demonstrate the efficacy of Dream2Act on a physical Unitree G1 robot across four diverse categories of whole-body mobile interaction tasks. Our results establish a scalable, data-driven path for zero-shot embodied intelligence, reliably outperforming conventional optimization-based retargeting baselines.

2. Related Work

2.1. Human-Centric Motion Acquisition and Retargeting

Imitating human demonstrations is a cornerstone for acquiring humanoid robotic skills. A prevalent paradigm relies on a multi-stage pipeline: first, human kinematics are extracted from monocular videos or motion capture data using human-centric parametric models such as SMPL [18] or GVHMR [29]. Similarly, in the Text-to-Motion domain, leading generative models, such as MDM [31], HY-Motion [33], Dart [11], and ViMogen [16] from the T2M benchmark [36], synthesize high-fidelity human kinematics directly from semantic prompts.

Irrespective of the source, these human-centric poses must be mapped to robotic embodiments through optimization-based solvers like GMT [7] or GMR [1] to bridge inherent skeletal discrepancies. These retargeted trajectories serve as foundational reference motions for training RL policies or whole-body trackers, a methodology pioneered by DeepMimic [26] and subsequently advanced by dexterous systems such as PHC [19], MimicKit [25] OmniH2O [12], and HumanPlus [9].

While effective for general locomotion, these methodologies are fundamentally bottlenecked by the morphology gap. Beyond simple discrepancies in skeletal proportions, these retargeting processes are often designed to optimize for kinematic resemblance in isolation from the physical scene. This leads to severe spatial misalignments and lost contact precision in object-centric tasks.

2.2. Robot-Centric Generative Planning

Recent breakthroughs in generative AI, particularly advanced large-scale video generative models like Wan 2.1 [32], Sora [3], Stable Video Diffusion [2], and the state-of-the-art Seedance 2.0 [5], have demonstrated unprecedented capabilities in modeling complex physical dynam-

ics. In the robotics community, extracting actionable priors from such models has gained immense traction. For tabletop robotic arms, extracting visual priors (e.g., R3m [23]) or regressing actions directly from human videos [28, 34] has proven highly effective.

For humanoids, recent works explore video world models like DreamGen [13] to unlock generalization, or frameworks like GenMimic [24] that leverage generative models to synthesize human videos, which are then retargeted into robot trajectories. However, all these approaches still struggle with the aforementioned retargeting trap due to their reliance on human intermediaries. Dream2Act pioneers a strictly robot-centric approach: by prompting Seedance 2.0 to hallucinate the physical robot itself, we obtain a native temporal prior that bypasses retargeting entirely. This allows our framework to serve as an explicit zero-shot motion planner that is inherently grounded in the environment’s geometry and the robot’s native morphology.

2.3. Robotic Perception and General-Purpose Tracking

Translating generative visual hallucinations into physical robot actions requires a robust integration of robotic pose perception and whole-body tracking [26]. While human pose models like HRNet [30] and ViTPose [35] are mature on datasets like COCO [17], they struggle with robotic embodiments due to the morphology gap and domain-specific data scarcity [35]. Dream2Act overcomes this by leveraging simulation environments like IsaacLab [22] and motion priors such as AMASS [21] to synthesize a native dataset, enabling precise pose perception grounded in the robot’s specific kinematic skeleton.

On the execution side, the field has evolved from task-specific imitation toward general-purpose whole-body tracking. Traditional frameworks often relied on analytical solvers like Pinocchio [6] or specialized systems like TWIST [37]. While agile trackers such as BeyondMimic [15] have improved responsiveness, they remain limited to specific motion distributions. Recently, the development of general tracking foundations—such as BFM-Zero [14] and Sonic [20]—has enabled the tracking of diverse, complex motions across unstructured environments. Dream2Act synergizes with these foundations by providing morphology-consistent and kinematically plausible visual priors, achieving stable zero-shot physical execution without requiring additional task-specific training.

3. Method

The core objective of Dream2Act is to synthesize physically viable interaction trajectories directly from initial visual observations, bypassing explicit and error-prone human-centric retargeting paradigms. Operating through a strictly robot-centric pipeline, our framework comprises three pri-

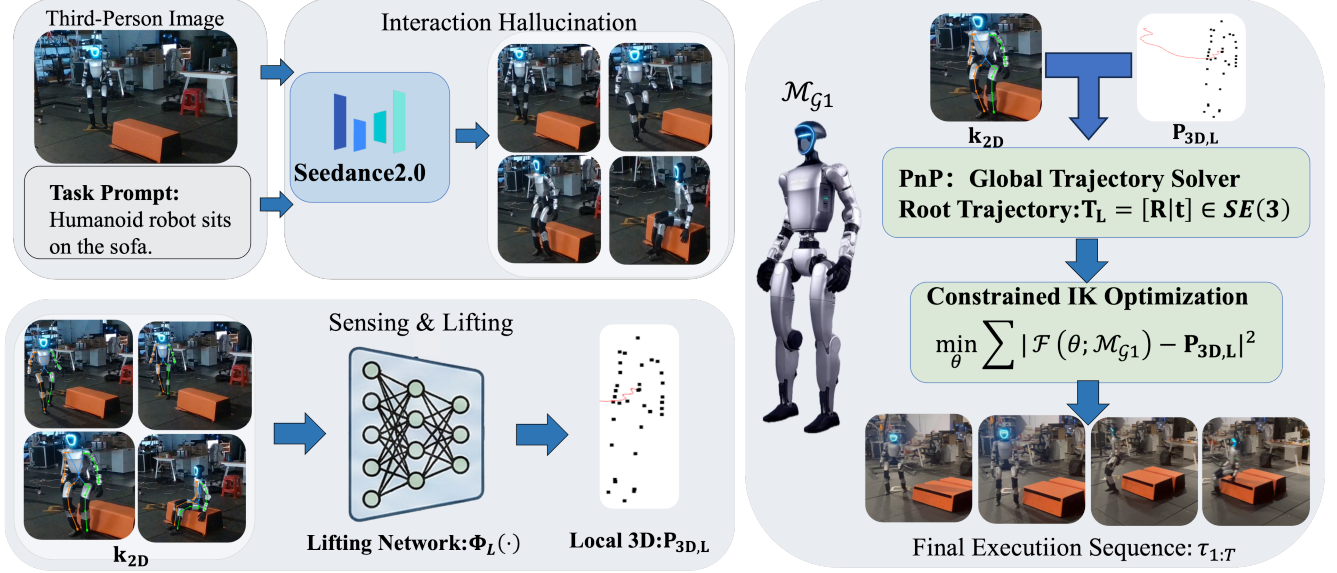


Figure 2. **Dream2Act System Architecture.** The framework comprises three core modules: (1) Interaction Hallucination via Seedance to generate morphology-consistent visual priors; (2) Native Pose Estimation and 2D-to-3D Lifting to extract precise 3D keypoints; and (3) Morphology-Aware Kinematic Recovery employing Constrained IK and PnP scale recovery to trajectories for physical execution.

mary stages: (1) Generative interaction hallucination via Seedance 2.0; (2) High-fidelity native pose estimation and 2D-to-3D lifting; and (3) Morphology-aware kinematic recovery and physical execution.

3.1. Interaction Hallucination via Seedance 2.0

To avoid the prohibitive costs associated with collecting complex interaction data in the physical world, we formulate motion planning as a video generation process. We employ the Seedance 2.0 [5] generative world model as our core predictive synthesis engine. Building upon the foundational architecture of its predecessor [27], Seedance 2.0 demonstrates unprecedented capabilities in modeling complex physical interactions and multi-subject dynamics.

We select this model over alternatives due to its state-of-the-art performance in two critical dimensions required for robotic perception: (1) Physical Plausibility: It exhibits superior adherence to real-world physics (e.g., rigid body dynamics and contact constraints) during complex interactions; and (2) Morphological Consistency: It maintains exceptional structural rigidity of the subjects without introducing distortions or artifacts, even under significant view-point shifts or prolonged task execution.

Before executing any novel interaction task, our framework only requires a single initial frame captured by a third-person camera in the real environment, containing the physical Unitree G1 robot and the target object. This frame serves as the exact spatial conditioning. By incorporating text instructions or semantic task prompts, Seedance

2.0 synthesizes a physically intuitive and temporally coherent video that shows the robot successfully completing the intended task. Crucially, the generated video effectively maintains a fixed third-person camera perspective and preserves the physical dimensions of the G1 with high morphological fidelity. This high-quality, morphology-consistent video stream thereby serves as a reliable visual prior for the subsequent native kinematic recovery.

3.2. Domain-Specific Robotic Dataset Construction

A fundamental bottleneck in deploying vision-based pose estimators on humanoid robots is the severe scarcity of domain-specific annotated datasets. To overcome this, we construct a large-scale, high-fidelity dataset for the G1 humanoid, integrating zero-cost simulation rendering with human-in-the-loop (HITL) real-world annotations.

Simulation-based High-Fidelity Rendering. To construct a robust motion prior, we leverage the AMASS dataset [21]. We first retarget its extensive human motion sequences to the G1 29-DoF kinematic model, yielding a massive initial trajectory collection of 5,337,514 frames. Since highly correlated adjacent frames offer marginal information gain and degrade the learning efficiency of subsequent networks, we perform a rigorous temporal downsampling. This step aggressively distills the dataset into 5,338 diverse and representative keyframes for high-fidelity rendering. With the robot anchored at the origin, we construct a hemispherical, omnidirectional multi-camera system in Isaac Lab. This array comprises nine high-resolution virtual cameras: four or-

thogonal side views, four diagonal views with a 45° downward pitch, and one strict top-down view. To bridge the coordinate system gap between the physics simulation (USD convention) and standard computer vision pipelines, we perform rigorous orthogonal frame transformations. Using the derived camera extrinsics and intrinsics, we efficiently project the 3D world joints into local camera space and subsequently into precise 2D pixel coordinates $[u, v]^T$. Beyond skeletal joints, we synchronously render depth maps and instance segmentation masks to automatically extract tight bounding boxes and center points for the robot. This pipeline furnishes a rich, multi-modal ground-truth suite for the subsequent visual generation and lifting modules.

In-the-Wild Data and Human-in-the-Loop 2D Annotation. To bridge the Sim-to-Real gap for visual perception, we collect in-the-wild videos of the physical G1 robot and temporally downsample them by a factor of 10 to reduce redundancy. We then partition these processed frames into a training set and an evaluation set. For the training set, we adopt a Human-in-the-Loop (HITL) annotation pipeline: a purely simulation-trained model first pre-annotates the 2D pixel coordinates, followed by expert manual correction of misaligned joints. Conversely, the evaluation set undergoes multi-round, fine-grained expert 2D annotation from scratch to establish a high-confidence benchmark. Crucially, this real-world dataset is exclusively utilized to fine-tune the 2D pose estimation network (detailed in Sec. 3.3) against real-world illumination and textures, whereas our 2D-to-3D lifting network relies on the exact 3D geometric ground truth provided by the simulation environment.

3.3. Native Pose Estimation and 2D-to-3D Lifting

Equipped with our hybrid dataset, we employ ViTPose [35] as the backbone architecture for Parameter-Efficient Fine-Tuning (PEFT). We freeze the first nine layers of the pre-trained ViT-Base model and fine-tune only the final three layers alongside the top-down Gaussian heatmap head. Given an input image I , this estimator accurately predicts the 2D keypoints P_{2D} for the $N = 30$ native robotic joints.

To transition from the 2D pixel space to the 3D physical space, the extracted 2D keypoints are first normalized (zero-centered and scaled to $[-1, 1]$) to isolate global scale variations. Subsequently, these normalized 2D coordinates are fed into a specifically trained 2D-to-3D Lifting Network. This network explicitly models the native 3D joint distribution of the G1 robot, directly regressing the 3D relative joint coordinates P_{3D} in the camera coordinate system, thereby reconstructing the native-scale 3D articulation required for subsequent kinematic recovery.

Formally, the entire visual perception and lifting pipeline can be summarized as:

$$P_{2D} = f_{\text{ViT}}(I), \quad P_{2D} \in \mathbb{R}^{N \times 2} \quad (1)$$

$$P_{3D} = \Phi_{\text{lift}}(\mathcal{N}(P_{2D})), \quad P_{3D} \in \mathbb{R}^{N \times 3} \quad (2)$$

where $f_{\text{ViT}}(\cdot)$ denotes the fine-tuned 2D pose estimator, $\mathcal{N}(\cdot)$ represents the spatial normalization function, and $\Phi_{\text{lift}}(\cdot)$ is the 2D-to-3D lifting network mapping the normalized 2D coordinates to the native-scale 3D domain.

3.4. Morphology-Aware Kinematic Recovery and Physical Execution

To recover the authentic motion trajectory of the G1 robot from the inferred 3D visual keypoints P_{3D} , we design a robust, decoupled pipeline that explicitly separates low-level internal joint kinematics from global root dynamics, culminating in authentic physical deployment.

Stage 1: Internal Joint Kinematics via Constrained IK.

We first compute the relative positions $\hat{P}_{3D}$ by centering all 3D points relative to the pelvis. The recovery of joint angles is formulated as a constrained Inverse Kinematics (IK) optimization problem:

$$q^* = \arg \min_q \sum_{i=1}^N \|\text{FK}_i(q) - \hat{P}_{3D}^{(i)}\|^2 \quad (3)$$

subject to joint limits $q_{\min} \leq q \leq q_{\max}$. By fixing the root node at the origin and optimizing solely on position constraints, the IK solver generates a structurally consistent internal joint configuration q^* that strictly adheres to the rigid URDF model of the G1.

Stage 2: Scale Recovery and Root Pose Optimization.

We apply Forward Kinematics (FK) to the internal state q^* (with the root fixed as an identity transformation) to obtain precise local 3D points $P_{\text{local}} \in \mathbb{R}^{N \times 3}$. Subsequently, we align these points with the un-normalized original 2D detections P_{2D} using the Perspective-n-Point (PnP) algorithm:

$$\min_{R, t} \sum_{i=1}^N \|P_{2D}^{(i)} - \pi(K(R \cdot P_{\text{local}}^{(i)} + t))\|^2 \quad (4)$$

Since P_{local} inherits the exact physical dimensions from the URDF model, the optimized translation vector t and rotation matrix R recover the true metric scale and global pose. Finally, an Exponential Moving Average (EMA) filter is applied to temporally smooth the trajectories of R and t . Subsequently, a rigorous coordinate transformation is performed to project these global trajectories back into the robot’s egocentric base frame, yielding the native motion sequence for downstream physical execution.

Stage 3: Unified Formatting and Sonic Tracker Execution.

Upon completing the kinematic recovery, we encapsulate the optimized global translation t , global rotation R (converted to quaternions), and internal joint angles q^* into the standardized General Motion Retargeting (GMR) format, yielding a unified high-dimensional state

vector (root_pos , root_rot , qpos). This standardized motion representation facilitates immediate zero-shot real-world deployment. We directly feed the generated GMR trajectory into the open-source Sonic general-purpose motion tracker. The Sonic tracker manages the low-level Whole-Body Control (WBC) and balance maintenance, driving the physical G1 robot to accurately track the extracted motion and end-to-end execute complex interaction tasks in real-world settings.

4. Experiments

Our experiments are systematically designed to answer the following research questions: (1) Can Dream2Act enable zero-shot physical interactions compared to human-centric retargeting? (2) Does robot-centric video synthesis offer a more physically plausible paradigm for free-space motion generation than conventional text-to-motion models? (3) how our specific system designs, namely the hybrid dataset, contribute to robust perception.?

4.1. Experimental Setup and Datasets

Datasets Construction. Training accurate native pose estimators requires domain-specific data. We construct a hybrid dataset comprising two distinct domains: (1) Simulation Dataset: We leverage AMASS motion priors, retargeting them to the G1 to yield over 5.3 million frames, which are then rigorously downsampled to 5,338 diverse keyframes. Through a 9-camera hemispherical setup in Isaac Lab, we render high-fidelity images accompanied by exact 3D geometric ground truth (P_{world}), bounding boxes, and masks. (2) In-the-Wild Real Dataset: To address the Sim-to-Real visual domain shift, we collect real-world videos of the G1, temporally downsampled to a factor of 10. These frames are partitioned into a 90% training set (refined via Human-In-The-Loop expert correction) and a 10% evaluation set (341 frames, annotated from scratch via multi-round expert labeling) to serve as a high-confidence benchmark.

Hardware and Execution. All physical deployments are conducted on the Unitree G1 humanoid (29 DoF). The extracted unified trajectories are seamlessly executed via the open-source Sonic general-purpose whole-body controller [20] running at 50 Hz.

4.2. Real-World Zero-Shot Interaction

To evaluate morphology-aware interaction, we define four diverse, spatially-sensitive tasks: (1) Ball-Kicking; (2) Box-hugging; (3) Bag-hitting; and (4) Sofa-sitting.

Baselines. To strictly isolate the impact of the morphology gap, we compare our approach against a representative optimization-based retargeting baseline: Baseline (Real Human Video): This pipeline extracts human kinematics from real-world demonstration videos using GVHMR [29],

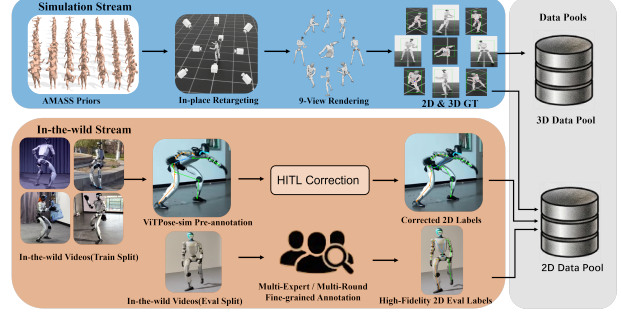


Figure 3. **G1 Domain-Specific Hybrid Dataset Construction Pipeline.** Our approach synergizes cost-free, high-fidelity 3D/2D geometry rendered from a 9-camera simulation environment (Isaac Lab) with Human-in-the-Loop (HITL) annotations to correct visual domain shifts, enabling robust native pose estimation.

Table 1. Quantitative comparison on four zero-shot interaction tasks. Task Length (in *italics*) denotes the initial distance to the target. Dream2Act significantly outperforms human-centric baselines, effectively bounding spatial alignment errors.

Method	Ball-kicking (<i>1.64m Avg</i>)		Box-hugging (<i>0.75m</i>)		Bag-hitting (<i>1.60m</i>)		Sofa-sitting (<i>1.10m</i>)	
	Succ $\uparrow$	Err $\downarrow$ (m)	Succ $\uparrow$	Err $\downarrow$ (m)	Succ $\uparrow$	Err $\downarrow$ (m)	Succ $\uparrow$	Err $\downarrow$ (m)
Baseline	0%	0.75	0%	0.35	0%	0.70	0%	0.65
Dream2Act (Ours)	40%	0.14	10%	0.11	30%	0.10	70%	0.28

which are subsequently mapped to the G1 via GMR [1] and executed by the Sonic [20] controller.

In contrast, Dream2Act (ours) operates through a strictly robot-centric paradigm. By prompting Seedance 2.0 to hallucinate the G1 robot directly, our framework bypasses the need for human-to-robot mapping. The synthesized native motion is processed through our domain-specific 2D-to-3D lifting and URDF-constrained IK optimization (Stage 2 & 3) before being tracked by the Sonic whole-body controller.

Metrics and Results. We report the Success Rate and the Average Spatial Alignment Error (calculated as the Mean Absolute Error of the end-effector distance to the intended contact point) across 20 trials per task. To evaluate robustness over distance, we also report the initial Task Length (the physical distance from the robot’s starting position to the interaction target). As shown in Table 1 and Table 2, the human-centric baseline fundamentally fails on spatially-sensitive tasks. Furthermore, we observe that traditional retargeting errors are severely compounded under locomotion. For instance, in the ‘Kick Ball (d)’ task requiring 2.18m of mobility, the human-centric baseline accumulates an overwhelming 0.79m of spatial misalignment. In contrast, Dream2Act effectively bounds the execution error (averaging 0.14m across kicking tasks) regardless of the task length. By operating natively within the G1’s embodiment, Dream2Act preserves precise spatial relationships and absolute metric scales, achieving significantly higher success rates and robust task completion.

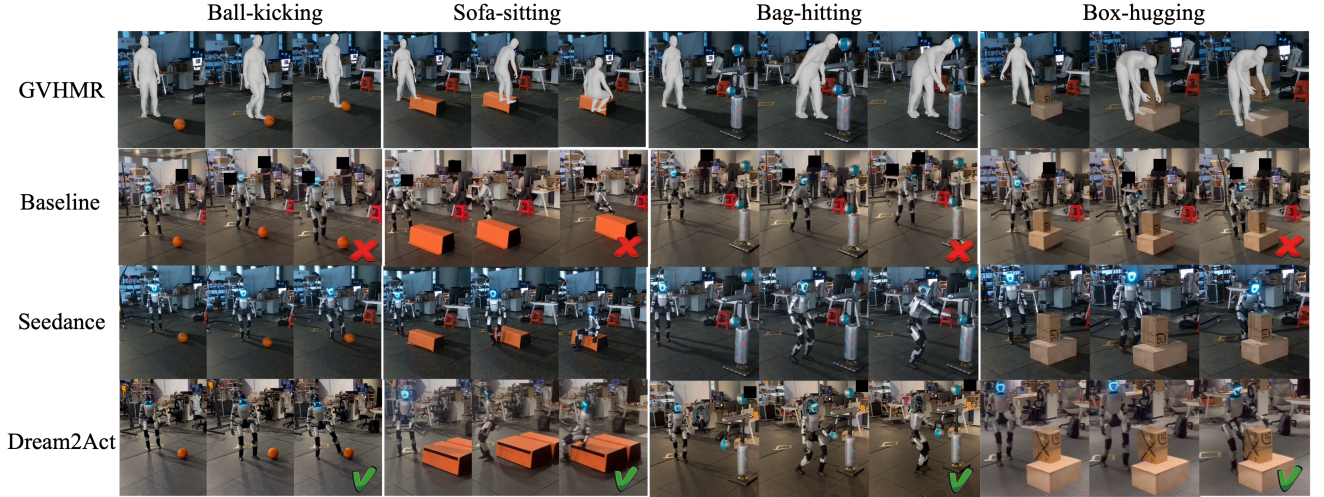


Figure 4. **Qualitative Comparison on Diverse Spatially-Sensitive Tasks.** We evaluate zero-shot performance across Kicking, Hugging, Punching, and Sitting. (Top) Traditional human-centric retargeting suffers from spatial misalignments and kinematic violations due to the morphology gap. (Bottom) Dream2Act’s robot-centric hallucination ensures precise contacts and dynamic stability.

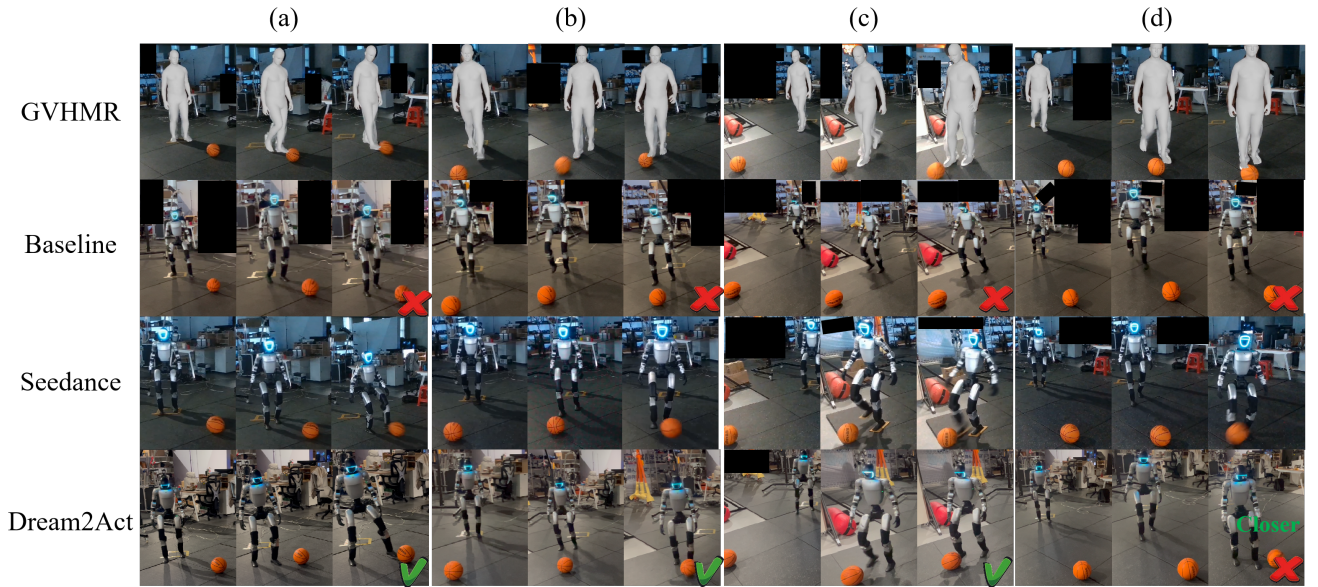


Figure 5. **Spatial Generalization Analysis on Multi-Location Ball Kicking.** To verify scale adaptability, we evaluate Dream2Act across four distinct ball positions. (Top) Human-centric retargeting fails to replicate spatial variations due to the morphology gap. (Bottom) Dream2Act successfully hallucinates and executes kicks for all targets, demonstrating robust spatial diversity.

Table 2. Quantitative comparison on multi-location ball-kicking. Task lengths vary from 0.90m to 2.18m. Dream2Act maintains precise spatial alignment across all targets, avoiding the error accumulation seen in baselines.

Method	Kick Ball (a) (0.90m)		Kick Ball (b) (2.05m)		Kick Ball (c) (1.45m)		Kick Ball (d) (2.18m)	
	Succ ↑	Err ↓(m)	Succ ↑	Err ↓(m)	Succ ↑	Err ↓(m)	Succ ↑	Err ↓(m)
Baseline	0%	0.53	0%	0.82	0%	0.84	0%	0.79
Dream2Act (Ours)	60%	0.19	20%	0.09	60%	0.09	20%	0.19

4.3. Text-Driven Free-Space Motion Generation

Beyond object-centric interaction, we qualitatively evaluate the efficacy of our robot-centric synthesis against conventional human-centric Text-to-Motion (T2M) pipelines. Given identical semantic prompts, baselines utilize state-of-the-art T2M models, such as ViMogen [16], MDM [31], and Dart [11], to generate human kinematics, which are subsequently retargeted via GMR for robot deployment.

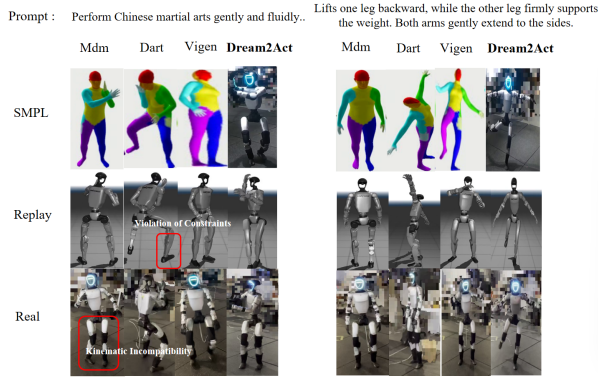


Figure 6. **Qualitative Comparison of Text-to-Motion Pipelines in Free Space.** Human-centric T2M models amplify inherent artifacts during retargeting, causing dynamic constraint violations and hardware failure. Dream2Act strictly enforces physical consistency, achieving stable zero-shot deployment.

Qualitative Results and Analysis. Our qualitative assessment, illustrated in Fig. 6, reveals that traditional human-centric pipelines suffer from severe compounding errors that manifest in three critical failure modes. First, kinematic incompatibility frequently arises because generated human poses often involve joint configurations or degrees of freedom that do not exist on the G1 robot, causing the mapping process to fail and resulting in frozen motor execution. Furthermore, these baselines consistently violate physical support constraints by hallucinating aerial trajectories where the agent appears to float or translate in mid-air without appropriate ground contact, a phenomenon that is physically unrealizable for a ground-based humanoid. Finally, minor artifacts inherent in human motion generation, such as foot sliding or jitter, are catastrophically amplified during retargeting, leading to violations of the G1’s dynamic balance constraints and causing the physical robot to fall. In contrast, Seedance 2.0 generates motions that are implicitly grounded in the G1’s native physical embodiment. By synthesizing morphology-consistent trajectories from the start, Dream2Act ensures that the resulting motion flawlessly maintains balance and physical stability while strictly adhering to the semantic intent of the text prompt.

4.4. Module Evaluation I: Native Pose Perception

Robust 2D perception is the prerequisite for kinematic recovery. We evaluate our 2D detection module on both the Simulation Test Set and the In-the-Wild Internet Test Set. **Baselines and Ablation.** We compare against zero-shot human pose estimators (ViTPose [35] pre-trained on COCO). We ablate our training strategy by comparing a variant fine-tuned solely on simulation data (*Ours-Sim Only*) against our model trained on the hybrid dataset (*Ours-Hybrid*).

Results. As detailed in Table 3, off-the-shelf human detectors fail catastrophically on the G1 morphology. While the *Sim Only* variant performs flawlessly on simulated images, its Average Precision (AP) and 2D PCK drop precipitously on the Internet test set due to lighting and texture domain shifts. Our *Hybrid* model effectively closes this Sim-to-Real gap, achieving robust native joint localization across all environments.

Table 3. 2D Pose Estimation Performance on 12 Common Joints. We compare zero-shot generalization of standard human estimators (COCO-17) vs. our domain-specific model on shared kinematic joints.

Method	Simulation Test Set (2402 frames)		Internet Test Set (341 frames)	
	AP $\uparrow$	2D PCK@0.05 $\uparrow$	AP $\uparrow$	2D PCK@0.05 $\uparrow$
Zero-shot ViTPose	0.234	64.99%	0.260	66.06%
Ours (Sim Only)	0.940	99.76%	0.497	91.61%
Ours (Hybrid)	0.785	95.78%	0.613	97.02%

4.5. Module Evaluation II: 2D-to-3D Lifting

The precision of converting 2D pixel coordinates k_{2D} into 3D joint positions $P_{3D,L}$ is fundamental to the fidelity of the final humanoid motion. To quantify the performance of our 2D-to-3D lifting network $\Phi_L(\cdot)$, we conduct a rigorous evaluation on the simulation test set, which provides absolute 3D ground truth for the G1 robot $\mathcal{M}_{G1}$. Using Mean Per Joint Position Error (MPJPE) as the primary metric, our lifting network achieves a low average error of 29mm on the simulation benchmark. This high precision confirms the network’s capability to accurately regress the robot’s native-scale 3D articulation from monocular keypoints, effectively overcoming the depth ambiguity inherent in single-view perception by explicitly modeling G1-specific joint distributions. Such high-fidelity 3D reconstruction provides a geometrically consistent prior for the subsequent URDF-constrained IK solver, ensuring that the recovered trajectories $\tau_{1:T}$ not only visually align with the generative hallucinations but also strictly adhere to the robot’s physical dimensions for stable real-world deployment.

5. Conclusion

In this paper, we presented Dream2Act, a pioneering robot-centric framework that enables true zero-shot humanoid interaction through generative video synthesis. By shifting from traditional human-centric retargeting to directly synthesizing interactions in the robot’s native spatial coordinate system, our approach fundamentally eliminates the persistent morphological gap and spatial misalignments that have long bottlenecked dexterous skill acquisition. To seamlessly bridge the pixel-to-physics gap, we introduced a robust, domain-specific kinematic recovery pipeline—cascading native 2D pose estimation, 2D-to-3D lifting, and URDF-constrained inverse kinematics—to

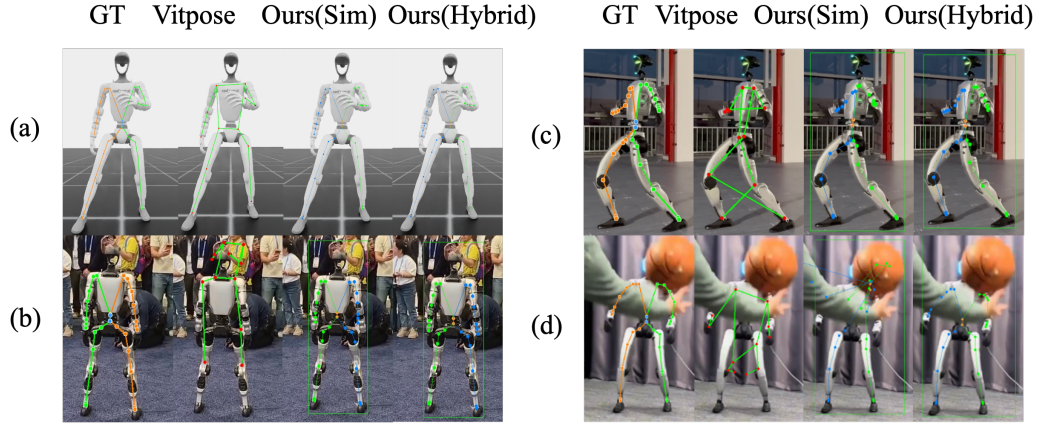


Figure 7. **Qualitative Ablation on Native G1 Pose Estimation.** Given a challenging real-world image: (a) Zero-shot ViTPose (COCO-only) completely misidentifies robot features; (b) A Sim-Only variant suffers from local drift due to domain shift; (c) Our Dream2Act native estimator, fine-tuned on the hybrid dataset, achieves pixel-perfect alignment.

translate visual priors into physically viable joint trajectories. Extensive physical validations on the Unitree G1 humanoid across a diverse set of spatially-sensitive interactions (e.g., ball kicking, sofa sitting, and bag punching) demonstrate that Dream2Act reliably achieves correct physical contacts and significantly outperforms conventional optimization-based baselines in both spatial alignment and task success rates. While currently functioning as an offline zero-shot motion planner due to the inherent inference latency of diffusion models, Dream2Act bypasses prohibitive data collection and task-specific policy training, establishing a highly scalable, data-driven pathway for advancing embodied humanoid intelligence, with future work focusing on accelerating inference for real-time closed-loop control.

References

- [1] Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C. Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking, 2025. 1, 3, 6
- [2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3
- [3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024. 2, 3
- [4] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments, 2024. 2
- [5] ByteDance Seedance Team. Seedance 2.0: Advanced multimodal video generation foundation model. https://seed.bytedance.com/zh/seedance2_0, 2026. 2, 3, 4
- [6] Justin Carpentier, Guilhem Saurel, Gabriele Buondonno, Joseph Mirabel, Florent Lamiraux, Olivier Stasse, and Nicolas Mansard. The pinocchio c++ library: A fast and flexible implementation of rigid body dynamics algorithms and their analytical derivatives. In *2019 IEEE/SICE International Symposium on System Integration (SII)*, pages 614–619. IEEE, 2019. 3
- [7] Zixuan Chen, Mazeyu Ji, Xuxin Cheng, Xuanbin Peng, Xue Bin Peng, and Xiaolong Wang. Gmt: General motion tracking for humanoid whole-body control. *arXiv preprint arXiv:2506.14770*, 2025. 3
- [8] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023. 2
- [9] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. *arXiv preprint arXiv:2406.10454*, 2024. 1, 3
- [10] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024. 1
- [11] Jiatao Gu, Yuyang Wang, Yizhe Zhang, Qihang Zhang, Dinghuai Zhang, Navdeep Jaitly, Josh Susskind, and Shuangfei Zhai. Dart: Denoising autoregressive transformer for scalable text-to-image generation, 2025. 3, 7
- [12] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024. 1, 3
- [13] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through neural trajectories. *arXiv e-prints*, pages arXiv–2505, 2025. 3
- [14] Yitang Li, Zhengyi Luo, Tonghe Zhang, Cunxi Dai, Anssi Kanervisto, Andrea Tirinzoni, Haoyang Weng, Kris Kitani, Mateusz Guzek, Ahmed Touati, et al. Bfm-zero: A promptable behavioral foundation model for humanoid control using unsupervised reinforcement learning. *arXiv preprint arXiv:2511.04131*, 2025. 3
- [15] Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025. 3
- [16] Jing Lin, Ruizi Wang, Junzhe Lu, Ziqi Huang, Guorui Song, Ailing Zeng, Xian Liu, Chen Wei, Wanqi Yin, Qingping Sun, et al. The quest for generalizable motion generation: Data, model, and evaluation. *arXiv preprint arXiv:2510.26794*, 2025. 3, 7
- [17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3
- [18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6), 2015. 1, 3
- [19] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10895–10904, 2023. 1, 3
- [20] Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Sirui Chen, Fernando Castaneda, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025. 2, 3, 6
- [21] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 3, 4
- [22] Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Munoz, Xinjie Yao, René Zurbrugg, Nikita Rudin, et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025. 3
- [23] 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. 3
- [24] James Ni, Zekai Wang, Wei Lin, Amir Bar, Yann LeCun, Trevor Darrell, Jitendra Malik, and Roei Herzig. From gen-

erated human videos to physically plausible robot trajectories. *arXiv preprint arXiv:2512.05094*, 2025. 1, 3

- [25] Xue Bin Peng. Mimickit: A reinforcement learning framework for motion imitation and control. *arXiv preprint arXiv:2510.13794*, 2025. 3
- [26] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018. 1, 3
- [27] Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025. 2, 4
- [28] Kenneth Shaw, Shikhar Bahl, Aravind Sivakumar, Aditya Kannan, and Deepak Pathak. Learning dexterity from human hand motion in internet videos. *The International Journal of Robotics Research*, 43(4):513–532, 2024. 3
- [29] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1, 3, 6
- [30] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 3
- [31] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3, 7
- [32] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3
- [33] Yuxin Wen, Qing Shuai, Di Kang, Jing Li, Cheng Wen, Yue Qian, Ningxin Jiao, Changhai Chen, Weijie Chen, Yiran Wang, et al. Hy-motion 1.0: Scaling flow matching models for text-to-motion generation. *arXiv preprint arXiv:2512.23464*, 2025. 3
- [34] Sicheng Xie, Haidong Cao, Zejia Weng, Zhen Xing, Haoran Chen, Shiwei Shen, Jiaqi Leng, Zuxuan Wu, and Yu-Gang Jiang. Human2robot: Learning robot actions from paired human-robot videos. *arXiv preprint arXiv:2502.16587*, 2025. 3
- [35] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022. 3, 5, 8
- [36] Bin Yang, Rong Ou, Weisheng Xu, Jiaqi Xiong, Xintao Li, Taowen Wang, Luyu Zhu, Xu Jiang, Jing Tan, and Renjing Xu. T2mbench: A benchmark for out-of-distribution text-to-motion generation. *arXiv preprint arXiv:2602.13751*, 2026. 3
- [37] Yanjie Ze, Zixuan Chen, Joao Pedro Araújo, Zi-ang Cao, Xue Bin Peng, Jiajun Wu, and C Karen Liu. Twist:

Teleoperated whole-body imitation system. *arXiv preprint arXiv:2505.02833*, 2025. 1, 3