H2R-GROUNDER: PAIRED-DATA-FREE HUMAN-TO-ROBOT VIDEO TRANSLATION WITH REAL-ROBOT SUPERVISION

Anonymous authors
Paper under double-blind review



Figure 1: **H2R-Grounder** converts human interaction videos into temporally aligned robotic manipulation videos, maintaining motion and background consistency and ensuring physically plausible robot arm structures and interactions. RoboMaster (Fu et al., 2025) (animation-based) losees motion and background consistency. Kling (Kuaishou, 2024) and Runway Aleph (Runway, 2025) (editing-based) produce geometrically distorted robot arms.

ABSTRACT

Robots that learn manipulation skills from everyday human videos could acquire broad capabilities without tedious robot data collection. We propose a video-to-video translation framework that converts ordinary human–object interaction videos into motion-consistent robot manipulation videos with realistic and physically plausible robot–object interactions. Our approach does not require any paired human–robot videos for training – only a set of unpaired robot videos, making the system easy to scale. We introduce a transferable representation that bridges the embodiment gap: by inpainting the robot arm in training videos to obtain a clean background and overlaying a simple visual cue (a marker and arrow indicating the gripper's position and orientation), we can condition a generative model to insert the robot arm back into the scene. At test time, we apply the same process to human videos (inpainting the person and overlaying human pose cues) and generate high-quality robot videos that mimic the human's actions. We fine-tune a SOTA video diffusion model (Wan2.2) in an in-context learning manner to ensure temporal coherence and leveraging of its rich prior knowledge. Empirical results demonstrate that our approach achieves significantly more realistic and grounded robot motions compared to baselines, pointing to a promising direction for scaling up robot learning from unlabeled human videos.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Collecting large-scale, diverse robot manipulation data remains a core challenge in robotics (O'Neill et al., 2024). Recording demonstrations with physical robots is slow, costly, and constrained to lab settings (Walke et al., 2023), leaving even the largest robot datasets far smaller and less varied than those in NLP. In contrast, human interaction videos—from casual online clips to egocentric recordings—are abundant and richly depict diverse manipulation behaviors. If robots could learn directly from these human videos, data collection would be vastly accelerated. Prior efforts often rely on specialized hardware (Liu et al., 2025) to collect paired human–robot data (Kareer et al., 2024) for learning, which limits scalability. Moreover, the visual embodiment gap—human arms and hands differ significantly in appearance and motion from robot arms and grippers—makes the learning non-trivial.

Recent works attempt to "robotize" human videos by rendering a robot arm into them to fill the visual gap, enabling imitation learning (Lepert et al., 2025b) or representation learning (Li et al., 2025a) for policy improvement. For instance, Phantom (Lepert et al., 2025c) inpaints the human hand in video frames and overlays a rendered robot arm in its place based on the estimated hand pose. Masquerade (Lepert et al., 2025b) and H2R (Li et al., 2025a) extend this idea to egocentric views. Although effective, these rendering-based methods often produce physically inconsistent visuals—robots may appear to float or misalign with objects—and require accurate camera calibration and pose estimation, which hinders generalization to in-the-wild videos. See Figure 2.

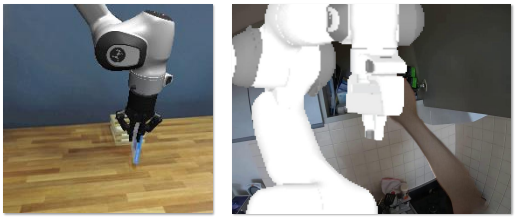


Figure 2: Issues in prior rendering-based H2R methods. (a) shows the rendered robot arm from **Phantom** (Lepert et al., 2025c), produced using their released code and provided calibrated camera parameters. Without accurate depth, the gripper appears to "float" above the book. (b) shows an overlaid robotic arm from **H2R** (Li et al., 2025a), collected from their dataset, which suffers from severe floating artifacts and camera misalignment.

In this paper, we introduce *H2R-Grounder*, a novel framework that marries the strengths of generative video models with a simple, transferable representation of manipulation, *H2Rep*. Our key insight is to remove the need for any paired human–robot videos in training by using only unpaired robot videos and an abstract conditioning signal that is common to both human and robot domains. Concretely, we take a collection of robot manipulation videos (which may be limited in scene diversity) and algorithmically strip the robot from them: we inpaint the robot arm out of each frame, yielding a clean background video of the scene and target objects. Into this background, we overlay a minimal pose indicator – a colored dot and arrow that mark the robot gripper's 2D location and orientation. This annotated video serves as the conditioning input. We then fine-tune the video generator Wan2.2 (Wan et al., 2025) to reconstruct the original robot video given this conditioned input. Through this process, the model learns to "insert" a robot arm into a scene according to the provided pose cues, effectively learning the mapping from gripper end-effector pose sequences to realistic robot imagery. Crucially, the model is learning from actual robot videos, so it observes correct physics, contacts, and occlusions during training – but it never sees a human in these videos.

At test time, we can apply the same procedure to a human demonstration video: estimate the human's hand pose, inpaint the person from the frames, and overlay the equivalent pose indicator. This produces a transferrable representation *H2Rep* of the human demonstration, to which our model can now respond by generating a robot video. The result is a robot manipulation video that follows the human's motion in the scene, with the robot properly interacting with the objects and environment (e.g. grasping and moving objects on a table, rather than hovering unnaturally). See Figure 1. Our approach offers several advantages. It eliminates the need for paired demonstrations, leverages existing robot datasets (Walke et al., 2023; O'Neill et al., 2024; Khazatsky et al., 2024; Brohan et al., 2022), and produces realistic, temporally consistent robot videos with visually plausible robot–object interactions. Moreover, our in-context fine-tuning strategy enhances temporal coherence compared to popular video-to-video pipelines such as VACE (Jiang et al., 2025). Finally, by using minimal 2D pose indicators instead of strict 3D alignment (Lepert et al., 2025c;b; Li et al., 2025a), our method

avoids calibration dependencies and generalizes robustly to diverse internet videos. To summarize, our contributions are threefold:

1. **A novel human-to-robot video translation framework — H2R-Grounder**, enabling robot video generation from human demonstrations without paired data.
2. **A simple and transferable intermediate representation — H2Rep,** unifying human and robot embodiments.
3. **An in-context fine-tuning scheme for large diffusion video models**, improving realism and temporal consistency for physically grounded generation.

2 Related Work

2.1 Intermediate Representations for Bridging Humans and Robots

Learning robot control from human videos is a long-standing challenge (Mandikal & Grauman, 2022; Shaw et al., 2023; Hafner et al., 2020). Due to the large visual embodiment gap between human and robot domains, most works (Kareer et al., 2024; Bharadhwaj et al., 2024a; Zhou et al., 2021) rely on shared intermediate representations as surrogates for joint learning. EgoMimic (Kareer et al., 2024) masks out both human hands and robot arms to minimize appearance differences. Others (Bahl et al., 2022; Zhou et al., 2021) inpaint manipulators and rely solely on background videos. Further studies leverage affordance maps (Bahl et al., 2023; Liu et al., 2025), keypoints (Bharadhwaj et al., 2024b; Haldar & Pinto, 2025), flow (Seita et al., 2023), pretrained models (Bi et al., 2025), or latent features (Xie et al., 2025). While these representations facilitate cross-domain learning, they seldom generate robot videos directly and thus remain limited by information loss or visual misalignment. Our method introduces *H2Rep*, combining pose sequences and background videos to preserve both motion and scene context. Unlike prior works that only use such representations for feature alignment, we employ them to directly synthesize robot videos from human inputs, closing the visual gap.

2.2 Translating Human Videos into Robot Videos

Recent works attempt to directly edit human videos into robot-like ones. Phantom (Lepert et al., 2025c) overlays rendered robot arms guided by estimated hand poses, while Masquerade (Lepert et al., 2025b) extends this to egocentric dataset epic-kitchen (Damen et al., 2022). H2R (Li et al., 2025a) similarly composites simulated robot arms onto inpainted egocentric frames (Grauman et al., 2022). These rendering-based pipelines exploit large-scale human data but struggle with realism—overlaid arms ignore lighting, depth, and scene geometry, leading to implausible occlusions or contacts. Moreover, they require accurate camera–robot calibration and sensor parameters (Lepert et al., 2025b;c; Wang et al., 2026), which are unavailable for in-the-wild videos. MimicDreamer (Li et al., 2025b; Team et al., 2025) narrows this embodiment gap via generative models, yet still conditions a generator on robot renderings, inheriting the same calibration requirement. In contrast, we adpot a fully generative approach, synthesizing robot videos conditioned on abstract 2D pose indicators. This design inherently models occlusion and contact learned from real robot data without calibration. HOPMan (Bharadhwaj et al., 2024a) is related, using off-the-shelf inpainting to remove robot arms and add human hands frame-by-frame (Ye et al., 2023), producing in-lab human–robot pairs. However, the reverse process—translating in-the-wild human videos into robot videos—remains infeasible due to the lack of a robot video generator. Our work fills this gap by introducing such a generator.

2.3 Cross-Robot Embodiment Transfer.

Several studies (Lepert et al., 2025a; Chen et al., 2024; Xu et al., 2025) investigate transferring across robots or exoskeletons with similar morphology, benefiting from their comparable kinematics. In contrast, our human-to-robot setting involves third-person videos with full-body humans and robotic manipulators of vastly different structures, making embodiment transfer substantially more challenging.

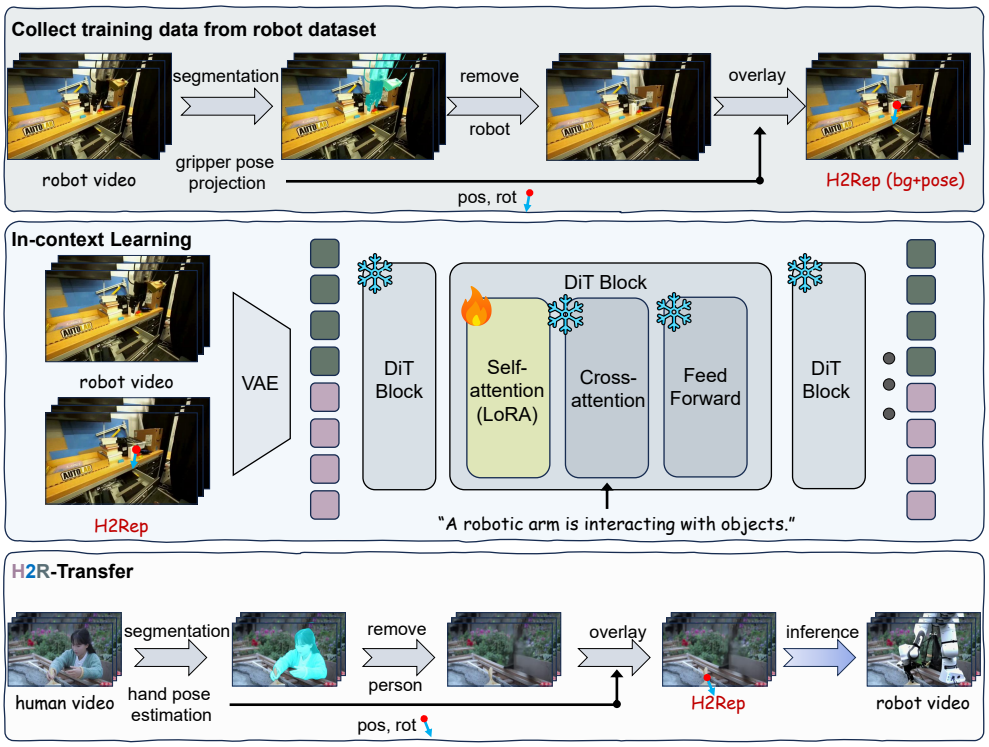


Figure 3: **Paradigm of H2R-Grounder.** The overall pipeline consists of three stages: (1) training data collection from robot video datasets, (2) in-context fine-tuning of the video generation model, and (3) transfer from in-the-wild human videos to robot manipulation videos.

3 METHODOLOGY

3.1 A SHARED ABSTRACTION FOR HUMAN AND ROBOT VIDEOS

There exist abundant human–object interaction (HOI) videos on the web and large collections of robot manipulation videos captured in labs (Khazatsky et al., 2024; O'Neill et al., 2024; Walke et al., 2023). However, collecting *frame-aligned* human–robot pairs at scale is prohibitively costly. We therefore seek a *shared representation* that bridges large-scale HOI videos and robot manipulation videos without requiring paired, frame-aligned supervision. We observe that both domains decompose naturally into: (i) a *pose trajectory* of the manipulator (human hand or robot gripper) that carries action semantics, and (ii) a *background video* that preserves scene layout and the physical state of manipulated objects. If we align human-hand and robot-gripper poses, then "pose sequence + background" becomes a common carrier of the key information in both domains. We denote this abstraction by **H2Rep**. In the following sections, we present: (1) how to extract *H2Rep*, from robot manipulation videos (Section 3.2); (2) how to train an in-context video generation model conditioned on this structured representation to synthesize robot videos (Section 3.3); and (3) how to obtain *H2Rep* from human–object interaction videos and leverage the video generator to generate frame-aligned robot videos (Section 3.4). The overall three-stage pipeline is illustrated in Figure 3.

Notation. Let $\mathbf{V}_r$ and $\mathbf{V}_h$ be a robot video and a human video, respectively. $\mathbf{H}_r$ and $\mathbf{H}_h$ are *H2Rep* extracted from robot video and human video, respectively. We use $\mathcal{S}$ for text-prompted video segmentation (Grounded-SAM2 (Ren et al., 2024)), $\mathcal{I}$ for video object inpainting, Π for 6-DoF-to-2D pose projection using calibrated cameras, $\mathcal{R}$ for rendering a pose as graphic overlays (red dot for position and blue arrow for orientation), and $\mathrm{Blend}(\mathbf{A}, \mathbf{B}; \alpha) = (1-\alpha)\mathbf{A} + \alpha\mathbf{B}$ for alpha blending with $\alpha = 0.4$. We use a video VAE encoder/decoder $(\mathsf{Enc}, \mathsf{Dec})$, and $e(\cdot)$ for a text embedding.

3.2 TRAINING DATA CONSTRUCTION FROM ROBOT VIDEOS

Robot-arm segmentation. Given a robot video $\mathbf{V}_r$, we obtain a pixel-accurate mask sequence with a text prompt:

$$\mathbf{M}_r \;=\; \mathcal{S}(\mathbf{V}_r,\ \text{“robotic arm”}). \tag{1}$$

Gripper pose projection. Let the end-effector (EEF) 6-DoF trajectory be $\mathbf{T}_{\text{EEF}}(t) = [\mathbf{p}(t), \mathbf{R}(t)]$ and camera intrinsics/extrinsics be $(\mathbf{K}, \mathbf{R}_c, \mathbf{t}_c)$. We project to image space:

$$\mathbf{P}_r(t) \;=\; \Pi\big(\mathbf{K}, \mathbf{R}_c, \mathbf{t}_c;\ \mathbf{p}(t), \mathbf{R}(t)\big), \tag{2}$$

and render a dot/arrow overlay $\mathcal{R}(\mathbf{P}_r)$ on each frame.

Robot-arm removal (background video). We remove the arm with a video inpainting model:

$$\mathbf{V}_r^{\mathcal{I}} \;=\; \mathcal{I}\big(\mathbf{V}_r,\ \mathbf{M}_r\big). \tag{3}$$

Empirically, Minimax-Remover (Zi et al., 2025) preserves background and removes the robot arm more reliably than another popular inpainting model E2FGVI (Li et al., 2022), so we adopt it in our pipeline. See Figure 4.

Composing robot video H2Rep. We form the shared representation by blending the rendered pose with the inpainted background:

$$\mathbf{H}_r \;=\; \text{Blend}\big(\mathbf{V}_r^{\mathcal{I}},\ \mathcal{R}(\mathbf{P}_r);\ \alpha\big)\,, \quad \alpha = 0.4. \tag{4}$$

This yields training pairs $\mathcal{D}_r \;=\; \big\{(\mathbf{H}_r^{(i)}, \mathbf{V}_r^{(i)})\big\}_{i=1}^{N}$, where $\mathbf{H}_r$ carries gripper motion and scene evolution, and $\mathbf{V}_r$ is the corresponding real robot video target.

3.3 IN-CONTEXT LEARNING FOR ROBOT VIDEO GENERATION

We train a conditional video generator G_θ (Wan 2.2 backbone (Wan et al., 2025)) to synthesize $\mathbf{V}_r$ conditioned on $\mathbf{H}_r$ (and a fixed text prompt c_{text}: “A robotic arm is interacting with objects.”). Following an in-context learning design, both $\mathbf{H}_r$ and $\mathbf{V}_r$ are encoded by the same VAE and fused by self-attention; only LoRA adapters (Hu et al., 2022) on the Q/K/V projections are trainable, while all other backbone weights remain frozen:

$$\mathbf{z}_H = \mathsf{Enc}(\mathbf{H}_r),\ \ \mathbf{z}_V = \mathsf{Enc}(\mathbf{V}_r),\ \ \mathbf{c} = \big[\mathbf{z}_H;\, e(c_{\text{text}})\big]. \tag{5}$$

We adopt a flow-matching objective. Let $\mathbf{z}_0 = \mathbf{z}_V$, sample $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and linearly interpolate $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1$ with target velocity $\mathbf{v}_t = \frac{d\mathbf{z}_t}{dt} = \mathbf{z}_1 - \mathbf{z}_0$. We train the conditional velocity field u_θ:

$$\mathcal{L} \;=\; \mathbb{E}_{t\sim\mathcal{U}(0,1),\ (\mathbf{H}_r,\mathbf{V}_r)\sim\mathcal{D}_r,\ \mathbf{z}_1\sim\mathcal{N}}\Big[\big\|u_\theta(\mathbf{z}_t,\ t,\ \mathbf{c}) - \mathbf{v}_t\big\|_2^2\Big]. \tag{6}$$

At inference, robot videos $\widehat{\mathbf{V}}_r$ can be generated with the trained generator G_θ from robot video *H2Rep* $\mathbf{H}_r$:

$$\widehat{\mathbf{V}}_r = G_\theta(\mathbf{H}_r, \mathbf{z}_1, t, \mathbf{c}_{\text{text}}). \tag{7}$$

Since the target $\mathbf{V}_r$ consists of real robot interaction videos, the generator is directly supervised by realistic robot appearance, kinematic patterns, occlusions, and contact configurations, encouraging physically plausible robot–object interactions.

3.4 HUMAN VIDEO TO ROBOT VIDEO

Given a third-person HOI video $\mathbf{V}_h$, we construct its *H2Rep* and feed it to the trained generator.

Person segmentation and hand pose. For any given HOI video V_h, we first employ Grounded-SAM 2.1 to obtain its mask sequence M_h. Meanwhile, we use ViT-Pose (Xu et al., 2022) to estimate the human body pose and locate the hand bounding box B_h, followed by HaMeR (Pavlakos et al., 2024) to accurately estimate the hand pose P_{hand}. We then take the midpoint between the index fingertip and thumb tip as the hand position, and the direction of the thumb as its orientation, forming a surrogate



Figure 4: **Comparison of video inpainting methods** on the robot arm removal task, evaluated on a sample from the Droid (Khazatsky et al., 2024) dataset.

pose P_h that effectively represents the hand's spatial position and direction. Empirically, we find that P_h aligns well to serve as a surrogate for the projected gripper pose in robot manipulation videos.

$$\begin{aligned} \mathbf{M}_h &= \mathcal{S}(\mathbf{V}_h, \text{“person”}), \\ \mathbf{P}_h &= \mathcal{D}(\mathbf{V}_h) \quad \text{(estimate surrogate 2D hand pose).} \end{aligned} \tag{8}$$

Person removal (background video). We use Minimax-Remover to remove person from the video:

$$\mathbf{V}_h^{\mathcal{I}} = \mathcal{I}\big(\mathbf{V}_h, \mathbf{M}_h\big). \tag{9}$$

Composing human video H2Rep . *H2Rep* from the human video also follows the same format as from the robot video:

$$\mathbf{H}_h = \text{Blend}\big(\mathbf{V}_h^{\mathcal{I}}, \mathcal{R}(\mathbf{P}_h); \alpha\big), \quad \alpha = 0.4. \tag{10}$$

H2R translation . We directly condition the trained robot generator G_θ on the human video abstract $\mathbf{H}_h$ to generate the robot video from human video:

$$\widehat{\mathbf{V}}_r = G_\theta(\mathbf{H}_h, \mathbf{z}_1, t, \mathbf{c}_{\text{text}}). \tag{11}$$

Because we fine-tune only lightweight LoRA adapters and keep the base generator frozen, G_θ maintains strong OOD generalization so we can apply it to in-the-wild videos.

3.5 Design Rationale for the H2Rep

H2Rep representation overlays the abstract pose sequence onto the background video using an α-blending scheme. Another natural design is to treat pose and background as two separate video streams—one containing only the background, and the other containing only the pose rendered on a white or black canvas. This alternative preserves more disentangled information.

However, under an in-context generation framework, using dual video streams would effectively *double* the input tokens, causing both computation and memory to scale quadratically (i.e., $4\times$). To balance efficiency and expressiveness, we adopt the α-blended formulation: the pose is overlaid with controlled transparency so as to minimally affect background content while substantially reducing computational and memory costs. Moreover, this representation remains pixel-aligned with both the human reference and the final generated robot video, which facilitates learning for the video generator.

4 Experiments

4.1 Experimental Setup

Datasets. We use the Droid dataset (Khazatsky et al., 2024) for training. This dataset contains approximately 76K diverse third-person Franka arm (GmbH, 2025) manipulation videos. During training, we randomly sample from the whole dataset while reserving 50 for validation. We report SSIM, and LPIPS (Zhang et al., 2018) to evaluate motion and background consistency as well as high-level visual feature distance between generated and ground-truth videos.

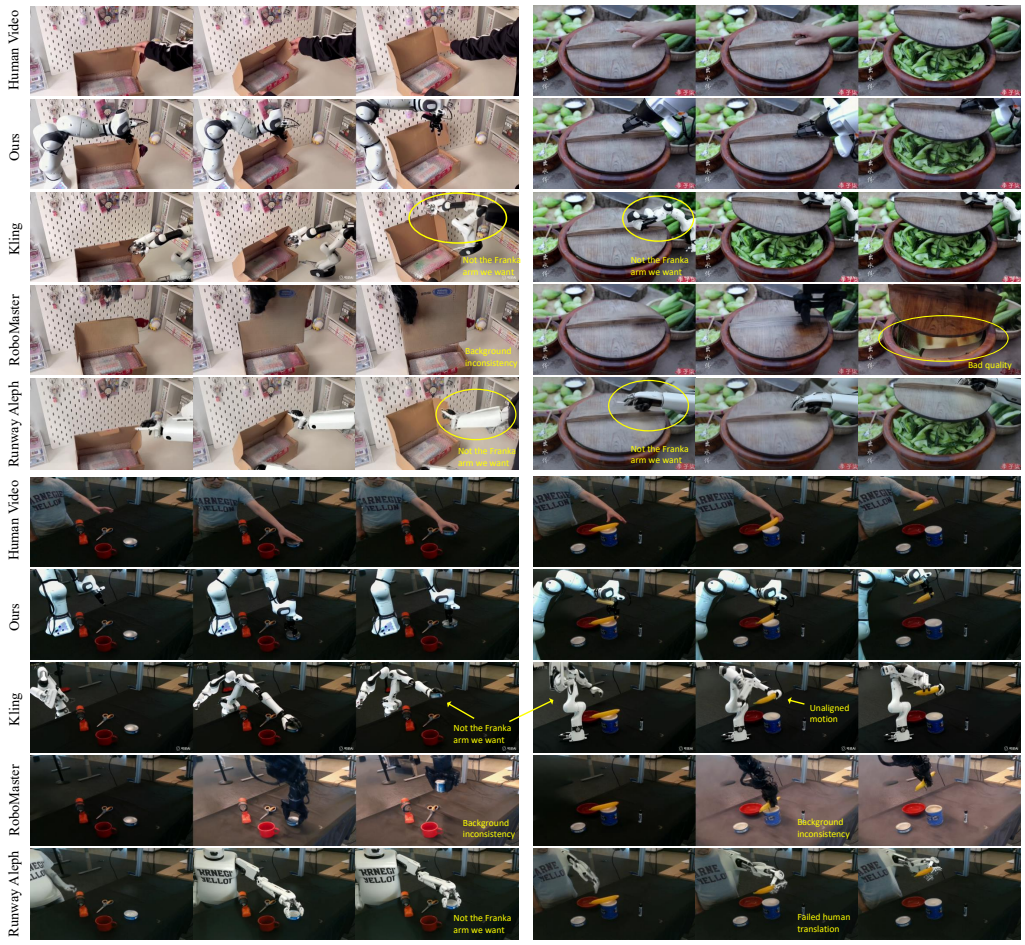


Figure 5: **OOD H2R transfer**. Top row: results on internet videos. Bottom row: results on DexYCB (Chao et al., 2021) videos.

To evaluate H2R-Grounder on out-of-distribution (OOD) human videos, we test on two types of data. First, we use the **DexYCB** human–object interaction dataset (Chao et al., 2021), which captures controlled lab-environment videos but exhibits clear domain shifts in both background and action distributions compared with Droid. It includes eight third-person camera views showing interactions between subjects and 20 distinct objects. We use the 100 videos from subject 01 under the camera 932122062010 top-down view as our test set. We do not use the ground-truth human masks or object poses provided by DexYCB; instead, we employ our automatic annotation pipeline described earlier to simulate real-world testing conditions. Since no ground-truth robot videos exist for comparison, we evaluate this set using two complementary metrics: (1) VLM-based evaluation and (2) human studies (gold standard), focusing on four aspects—motion consistency, background consistency, visual quality, and physical plausibility (robot integrity and contact realism). In addition, we collect **internet videos** featuring more diverse backgrounds, occlusions, viewpoints, and camera motions for qualitative comparisons with baseline methods.

Data preprocessing. All training videos are standardized to a resolution of 1280×720 and downsampled to 10 fps. We trim each clip to ensure its frame count n satisfies $n \bmod 4 = 1$, which is required by both Minimax-Remover and Wan. During fine-tuning, we randomly sample a clip of up to 49 frames from each training video. Thanks to Wan's strong pretraining, the fine-tuned generator generalizes well to human videos of different frame rates during inference.

Table 1: **Human preference rate on DexYCB.** Users are asked to rank the three generated videos, and our model is most frequently selected as the top choice for all aspects.

	Motion Consistency	Background Consistency	Visual Quality	Physical Plausibility
RoboMaster (Fu et al., 2025)	2.3%	2.3%	2.3%	18.2%
Runway Aleph (Runway, 2025)	22.7%	15.9%	9.1%	6.8%
Kling (Kuaishou, 2024)	9.1%	34.1%	40.9%	9.1%
Ours	**54.5%**	**56.8%**	**61.4%**	**63.6%**

Table 2: **VLM scoring on DexYCB.** We prompt Gemini (Comanici et al., 2025) to rate the generated videos across four aspects. Our model outperforms the baselines on most metrics, with a slight drop in visual quality compared to Kling (Kuaishou, 2024).

	Motion Consistency	Background Consistency	Visual Quality	Physical Plausibility
RoboMaster (Fu et al., 2025)	2.6	4.5	3.5	2.8
Runway Aleph (Runway, 2025)	**3.7**	4.5	3.6	3.9
Kling (Kuaishou, 2024)	3.5	**4.9**	**4.1**	3.6
Ours	**3.7**	**4.9**	4.0	**4.4**

Training. We fine-tune the Wan 2.2 TI2V-5B model (Wan et al., 2025) as our primary video generator. We tune with a LoRA (Hu et al., 2022) adapter for 200 steps with a mini-batch size of 4, using 8 NVIDIA H200 GPUs and a gradient accumulation factor of 2.

4.2 Baselines

Rendering-Based Methods Rendering-based approaches such as Phantom (Lepert et al., 2025c) and Masquerade (Lepert et al., 2025b) require precise hand–robot calibration to compute the transformation between the camera and robot frames, as well as accurate camera intrinsics and vertical field-of-view parameters for physically correct rendering of robot arms. Such parameters are unavailable for our in-the-wild human videos, making direct comparison infeasible. Therefore, these methods are excluded from evaluation. More detailed explanation is provided in Appendix.

Animation-Based Methods We adapt the recently proposed robot I2V method RoboMaster (Fu et al., 2025) to the human-to-robot (H2R) translation setting. The original system animates robot–object interaction videos from a static image given user-defined robot and object trajectories. To enable comparison under our setup, we construct the required inputs through a semi-manual process: (1) The first frame of the human–object video is inpainted to remove the human, serving as the reference frame; (2) hand pose trajectories are extracted following our H2R-Grounder pipeline and used as surrogates for robot trajectories; (3) the interacted object is manually selected and segmented using SAM 2.1 (Ravi et al., 2025) to obtain its mask; (4) its motion trajectory is tracked by CoTracker3 (Karaev et al., 2025); (5) the trajectory is manually divided into pre-interaction, interaction, and post-interaction phases; and (6) a textual caption describing the robot motion is written. This process allows RoboMaster to generate robot–object interaction animations, albeit with heavy manual preparation.

Commercial Video-Editing Methods Commercial video-editing systems such as Kling (Kuaishou, 2024) and Runway Aleph (Runway, 2025) can replace the subject of a video while roughly maintaining temporal coherence and background appearance. We upload an image of a Franka robotic arm and prompt Kling to replace the human in each input video with the robot arm. For Aleph, we similarly prompt it to replace the human in the video with a Franka robotic arm. This serves as a practical baseline representing appearance-level subject replacement rather than true generative translation.

4.3 Quantitative Results on DexYCB

Table 1 and Table 2 summarize the results on the DexYCB test set. We evaluate H2R-Grounder, Kling, and RoboMaster through both human studies and VLM-based scoring.

Human study. We conduct a user study with 22 participants, all holding computer-science backgrounds (bachelor's, master's, or PhD). Each participant ranks the outputs from the three methods in terms of motion consistency, background consistency, visual quality, and physical plausibility

(measured by structure integrity and contact realism). We report the first-rank rate—the percentage of participants who selected a method as best for each aspect. Ties are allowed in the ranking, so the total percentages may not sum to 100%.

As shown in Table 1, *H2R-Grounder* achieves the highest first-rank preference across all four evaluation aspects. It is most favored in visual quality (61.4%) and physical plausibility (63.6%), indicating that our generated videos are both visually convincing and physically coherent, with accurate object contacts. The high preference in motion consistency (54.5%) and background consistency (56.8%) further demonstrates that our model produces temporally stable motions while preserving contextual alignment. Kling ranks second, benefiting from its commercial editing pipeline, which yields visually appealing results (40.9%) and stable backgrounds (34.1%). However, it struggles with motion consistency (9.1%) and physical plausibility (9.1%), where the synthesized arms often lose structure or exhibit implausible interactions. Runway Aleph achieves moderate results, particularly in motion consistency (22.7%), but remains less realistic overall. RoboMaster performs the weakest, with preference rates around 2–3% across most aspects, showing that manually defined trajectories fail to capture natural motion or consistent visual quality. Overall, the human study demonstrates that H2R-Grounder achieves the best balance between motion realism, physical grounding, and visual fidelity, without relying on paired data or calibration.

VLM evaluation. We further evaluate using Gemini (Comanici et al., 2025), a multimodal visual–language model, to rate each generated video on a 1–5 scale across the same four criteria (Table 2). The VLM results align with human preferences: H2R-Grounder attains the highest or comparable scores in motion consistency (3.7), background consistency (4.9), and physical plausibility (4.4), confirming its robust understanding of scene dynamics and contact physics. Kling achieves slightly higher visual quality (4.1 vs. 4.0), likely due to its polished rendering style, but lags behind in realism-related aspects. RoboMaster again performs the worst, limited by its predefined, non-adaptive motion generation. Together, these results highlight that H2R-Grounder delivers the most balanced and physically grounded video generation among all baselines.

Data usability rate. Beyond quantitative scores, we additionally measure the practical usability of the generated robot videos, defined as the proportion of videos with negligible artifacts or abnormal visual quality. We manually inspected all transferred videos on the DexYCB dataset through human evaluation. Approximately 72% of the videos exhibit little to no visible issues and are deemed directly usable. Most failure cases arise when the manipulated object is significantly larger than the robot gripper, leading to physically implausible grasp configurations or minor geometric distortions. These results indicate that H2R-Grounder achieves a relatively high practical usability rate in real-world transfer scenarios.

4.4 Qualitative Results

Figure 5 presents qualitative comparisons of H2R-Grounder against existing baselines on both internet videos and DexYCB sequences. Although our video generator is fine-tuned only on the DROID indoor dataset, it generalizes well to in-the-wild videos, maintaining consistent backgrounds, accurate motion alignment, and sharp visual quality across different viewpoints. In contrast, Kling and Runway Aleph often produces structurally inconsistent robot arms that deviate from real-world kinematics, while RoboMaster significantly distorts the background and fails to follow the demonstrated motion precisely. As shown in the bottom-right example, H2R-Grounder accurately positions the gripper to grasp the banana tip, faithfully following the human hand trajectory.

5 Conclusion

We presented H2R-Grounder, a paired-data-free framework that translates human interaction videos into physically grounded robot manipulation videos. By introducing the unified representation *H2Rep*, our method bridges the visual embodiment gap between humans and robots without requiring calibration, paired supervision, or explicit action labels. Extensive experiments demonstrate that H2R-Grounder generates motion-consistent, physically plausible robot videos with strong generalization to in-the-wild human videos. Beyond direct control learning, the generated robot videos can serve as scalable pretraining data for robot world models, safety or reward modeling, and vision encoder adaptation, offering a new pathway for leveraging abundant human videos in robotics.

AI USE STATEMENT

In this work, we used generative AI tools for assist with translation, support qualitative and thematic data analysis, and interpret results. We have not used generative AI tools for generate synthetic data sets, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, and help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, are not applicable to this work. Additionally, we used generative AI tools for draft parts of a research paper, sourcing/searching for information, edit a research paper to improve readability, suggest a structure for a research paper, propose a title or keywords for a research paper. We have reviewed all AI-assisted work. All AI-assisted content was carefully reviewed and revised by the authors. AI-generated text was checked for factual accuracy, technical correctness, and consistency with the original research. Information obtained through AI-assisted searches was independently verified against the relevant literature and primary sources. The authors also reviewed and refined AI-assisted interpretations of experimental results to ensure that they were supported by the actual findings. All scientific claims, conclusions, and final manuscript content were critically evaluated and approved by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of our method and experimental results. Section 3 provides a detailed description of the H2Rep representation, the construction of training data from unpaired robot videos, and the in-context video generation framework, including its training objective. Section 4 describes the datasets, preprocessing procedures, model configurations, training settings, evaluation protocols, and baseline implementations.

REFERENCES

Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13778–13790, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Homanga Bharadhwaj, Abhinav Gupta, Vikash Kumar, and Shubham Tulsiani. Towards generalizable zero-shot manipulation via translating human interaction plans. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6904–6911. IEEE, 2024a.

Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision*, pp. 306–324. Springer, 2024b.

Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-rdt: Human manipulation enhanced bimanual robotic manipulation. *arXiv preprint arXiv:2507.23523*, 2025.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9044–9053, 2021.

Danze Chen, Yanzhe Chen, Qiming Huang, Zhijun Cao, Chen Gao, and Mike Zheng Shou. Supervise what survives: Geometry-guided vla adaptation from synthetic robot videos. *arXiv preprint arXiv:2606.24448*, 2026.

Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Muhammad Zubair Irshad, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. *arXiv preprint arXiv:2409.03403*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. URL https://doi.org/10.1007/s11263-021-01531-2.

Xiao Fu, Xintao Wang, Xian Liu, Jianhong Bai, Runsen Xu, Pengfei Wan, Di Zhang, and Dahua Lin. Learning video generation for robotic manipulation with collaborative trajectory control. *arXiv preprint arXiv:2506.01943*, 2025.

Franka Robotics GmbH. Franka robotics — homepage. https://franka.de/, 2025. Accessed: 2025-11-14.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination, 2020. URL https://arxiv.org/abs/1912.01603.

Siddhant Haldar and Lerrel Pinto. Point policy: Unifying observations and actions with key points for robot manipulation. *arXiv preprint arXiv:2502.20391*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing, 2025. URL https://arxiv.org/abs/2503.07598.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6013–6022, 2025.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video, 2024. URL https://arxiv.org/abs/2410.24221.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

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

Kuaishou. Kling. https://klingai.com/, 2024. Accessed: 2025-11-08.

Marion Lepert, Ria Doshi, and Jeannette Bohg. Shadow: Leveraging segmentation masks for cross-embodiment policy transfer. *arXiv preprint arXiv:2503.00774*, 2025a.

Marion Lepert, Jiaying Fang, and Jeannette Bohg. Masquerade: Learning from in-the-wild human videos using data-editing, 2025b. URL https://arxiv.org/abs/2508.09976.

Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos, 2025c. URL https://arxiv.org/abs/2503.00779.

Guangrun Li, Yaoxu Lyu, Zhuoyang Liu, Chengkai Hou, Jieyu Zhang, and Shanghang Zhang. H2r: A human-to-robot data augmentation for robot pre-training from videos, 2025a. URL https://arxiv.org/abs/2505.11920.

Haoyun Li, Ivan Zhang, Runqi Ouyang, Xiaofeng Wang, Zheng Zhu, Zhiqin Yang, Zhentao Zhang, Boyuan Wang, Chaojun Ni, Wenkang Qin, et al. Mimicdreamer: Aligning human and robot demonstrations for scalable vla training. *arXiv preprint arXiv:2509.22199*, 2025b.

Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17562–17571, 2022.

Vincent Liu, Ademi Adeniji, Haotian Zhan, Siddhant Haldar, Raunaq Bhirangi, Pieter Abbeel, and Lerrel Pinto. Egozero: Robot learning from smart glasses, 2025. URL https://arxiv.org/abs/2505.20290.

Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, pp. 651–661. PMLR, 2022.

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)*, pp. 6892–6903. IEEE, 2024.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Runway. Runway aleph. https://runwayml.com/research/introducing-runway-aleph, 2025. Accessed: 2025-11-08.

Daniel Seita, Yufei Wang, Sarthak J Shetty, Edward Yao Li, Zackory Erickson, and David Held. Toolflownet: Robotic manipulation with tools via predicting tool flow from point clouds. In *Conference on Robot Learning*, pp. 1038–1049. PMLR, 2023.

Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *Conference on Robot Learning*, pp. 654–665. PMLR, 2023.

GigaWorld Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jiagang Zhu, Kerui Li, Mengyuan Xu, et al. Gigaworld-0: World models as data engine to empower embodied ai. *arXiv preprint arXiv:2511.19861*, 2025.

Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

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.

Ye Wang, Pei Lin, Xiong-Hui Chen, Haoqi Yuan, Zhixuan Liang, Yiyang Huang, Anzhe Chen, Zixing Lei, Jie Zhang, Tao Zhang, et al. Ego2robot: Scalable robot data synthesis from egocentric human data. *arXiv preprint arXiv:2608.02580*, 2026.

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, 2025. URL https://arxiv.org/abs/2502.16587.

Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation. *arXiv preprint arXiv:2505.21864*, 2025.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022.

Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22479–22489, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Yuxiang Zhou, Yusuf Aytar, and Konstantinos Bousmalis. Manipulator-independent representations for visual imitation. *arXiv preprint arXiv:2103.09016*, 2021.

Bojia Zi, Weixuan Peng, Xianbiao Qi, Jianan Wang, Shihao Zhao, Rong Xiao, and Kam-Fai Wong. Minimax-remover: Taming bad noise helps video object removal. *arXiv preprint arXiv:2505.24873*, 2025.

APPENDIX

A MOTIVATION OF H2REP

In this paper, our H2Rep representation overlays the abstract pose sequence onto the background video using an α-blending scheme. Another natural design is to treat pose and background as two separate video streams—one containing only the background, and the other containing only the pose rendered on a white or black canvas. This alternative preserves more disentangled information.

However, under an in-context generation framework, using dual video streams would effectively *double* the input tokens, causing both computation and memory to scale quadratically (i.e., $4\times$). To balance efficiency and expressiveness, we adopt the α-blended formulation: the pose is overlaid with controlled transparency so as to minimally affect background content while substantially reducing computational and memory costs. Moreover, this representation remains pixel-aligned with both the human reference and the final generated robot video, which facilitates learning for the video generator.

B CLARIFICATION ON THE COMPARABILITY OF PHANTOM/MASQUERADE ON DEXYCB

Phantom/Masquerade require a calibrated camera-to-robot-base transformation, rather than only standard camera intrinsics and extrinsics, to render the robot arm at a physically plausible pose in the scene. In practice, this entails constructing a real robot–camera setup that approximates the viewpoint of the human demonstration, selecting a feasible robot base pose, performing eye-to-hand calibration,

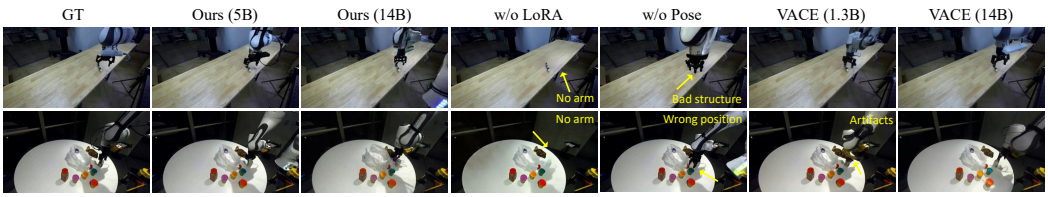


Figure 6: **Qualitative ablation on Droid**. Examples obtained by extracting frames at the same timestep.

Table 3: **Quantitative ablation on Droid.** ↑ indicates higher is better; ↓ indicates lower is better.

	SSIM ↑	LPIPS ↓
HR-Grounder 5B (ours)	**0.82**	**0.22**
w/o pose indicator	0.80	0.23
w/o LoRA	0.80	0.26
w/ 14B backbone	0.79	0.23
w/ VACE (1.3B) (Jiang et al., 2025)	0.68	0.30
w/ VACE (14B) (Jiang et al., 2025)	0.71	0.27

and handling depth-aware compositing. Such requirements make these pipelines difficult to apply to unconstrained Internet videos or datasets such as DexYCB, where the original camera–robot geometry is unavailable. Their applicability to egocentric datasets such as EPIC-KITCHENS is facilitated by the relatively consistent human–camera geometry induced by head-mounted cameras, which allows a calibrated setup to be reused with approximate spatial adjustments. In contrast, our pipeline operates without camera–robot calibration or scene depth, enabling direct application to in-the-wild Internet videos and other unconstrained human demonstrations without requiring a replicated physical setup.

C Inference Efficiency

Our 5B model runs at 13.22 seconds per frame, taking about 648 seconds to generate a 49-frame 704×1280 video on a single H200 GPU, with a peak memory consumption of 63 GB.

D Ablation Study

We ablate the key design choices of H2R-Grounder, including the pose representation, generator conditioning mechanism, parameter-efficient adaptation, and model scale. These experiments examine which components are necessary for accurately transferring the motion in a human demonstration to a robot video while preserving scene appearance and temporal coherence. Table 3 and Figure 6 provide quantitative and qualitative comparisons of these variants.

Effect of the generation architecture. Our framework is not restricted to a particular video generator: in principle, the proposed *H2Rep* can be combined with different conditional video generation architectures. We therefore compare our in-context video generator with VACE (Jiang et al., 2025), a representative framework that injects structural conditions through a ControlNet-style mechanism (Zhang et al., 2023). This comparison examines whether explicitly injecting *H2Rep* through a ControlNet is sufficient, or whether in-context conditioning provides a better mechanism for jointly interpreting the background, object configuration, and desired robot motion. Because VACE relies strongly on text conditioning, we additionally use Qwen2.5-VL (Bai et al., 2025) to automatically generate detailed captions for every training and testing video. We train VACE for two full epochs over the entire training set to ensure convergence. Replacing the in-context generator with VACE leads to lower SSIM and higher LPIPS in Table 3. Since both variants receive the same underlying *H2Rep* information, the performance gap suggests that the conditioning mechanism itself is important. In particular, in-context conditioning is better able to jointly preserve the scene context

and translate the provided motion cues into coherent robot dynamics, whereas the ControlNet-based alternative exhibits weaker motion–background consistency.

Effect of the pose indicator. Removing the pose indicator from *H2Rep* causes noticeable motion drift, showing that appearance cues alone are insufficient to precisely specify the desired robot trajectory. As illustrated in Figure 6 (second row, w/o Pose), the generated gripper deviates from the target white bottle and instead moves toward a nearby green bottle. The explicit pose indicator therefore serves as a compact spatial control signal that anchors the generated robot motion to the intended human interaction.

Effect of LoRA adaptation. Removing LoRA fine-tuning substantially degrades the structure and temporal consistency of the generated robot arm (Figure 6, w/o LoRA). This indicates that, although the pretrained video model provides a strong generative prior, task-specific adaptation is still necessary to learn the correspondence between *H2Rep* and robot embodiment. LoRA provides this adaptation efficiently while retaining the general video-generation capability of the pretrained model.

Effect of model scale. Finally, scaling the backbone from 5B to 14B parameters does not provide a clear improvement in generation quality, while substantially increasing inference cost and reducing the maximum supported sequence length from 49 to 17 frames. This suggests that model capacity is not the primary bottleneck for the proposed representation and conditioning scheme. Considering both generation quality and computational efficiency, we therefore adopt the 5B in-context model as the default configuration of H2R-Grounder.

E Benefit for Real Robot Manipulation

We further investigate whether H2R-Grounder can benefit downstream real-robot manipulation, beyond generating visually plausible robot videos. Human demonstrations provide abundant task-level motion information, but H2R-translated videos do not contain reliable low-level robot action labels. Rather than recovering pseudo-actions from generated pixels, we exploit H2R-translated videos through future 2D trajectory prediction. Given a translated robot frame, an auxiliary head predicts the next K end-effector waypoints, providing geometric supervision for learning task-relevant spatial dynamics. We then fine-tune the policy on real robot demonstrations with ground-truth action labels, while retaining the waypoint prediction objective as a spatial anchor. In this way, H2R-Grounder transfers human demonstration data into robot-domain visual supervision without requiring action annotations for the translated videos.

Two-stage policy training. We build our manipulation policy on OpenVLA-OFT (Kim et al., 2025) and follow the two-stage training strategy of GRA (Chen et al., 2026). Let

$$y_t = (p_{t+1}, p_{t+2}, \ldots, p_{t+K}), \qquad p_\tau \in \mathbb{R}^2, \tag{12}$$

denote the future K-step end-effector trajectory in normalized image coordinates.

In **Stage 1**, we use H2R-translated videos for geometric pretraining. We use the position points of *H2Rep* derived from human demonstrations as the targets. An auxiliary head h_{2D} predicts the future trajectory from the OpenVLA-OFT's visual representation $\phi_v(I_t^{\text{H2R}})$:

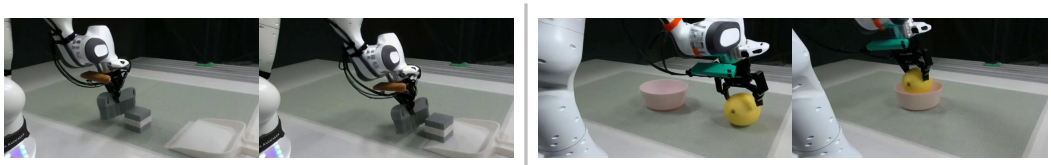
$$\mathcal{L}_{\text{2D}}^{\text{H2R}} = \mathbb{E}_{(I_t^{\text{H2R}}, y_t^{\text{H2R}})} \left[\left\| h_{\text{2D}} \left(\phi_v(I_t^{\text{H2R}}) \right) - y_t^{\text{H2R}} \right\|_2^2 \right]. \tag{13}$$

This stage updates the visual encoder ϕ_v and the auxiliary trajectory head h_{2D}, encouraging the visual representation to encode task-relevant motion geometry without imposing potentially inaccurate action supervision on generated videos.

In **Stage 2**, we fine-tune the policy using real robot demonstrations only. The action head learns ground-truth EEF robot actions, while the auxiliary trajectory objective is retained on real observations to preserve the spatial representation acquired in Stage 1:

$$\mathcal{L} = \mathcal{L}_{\text{act}} + \lambda \mathcal{L}_{\text{2D}}^{\text{real}}, \tag{14}$$

The real-video waypoint targets are obtained by projecting the recorded end-effector trajectory into the camera view and therefore require no additional annotation. Following GRA, we use $K = 8$ and $\lambda = 0.5$.



Sweep the block | *Pick up the lemon and place it in the bowl*

Figure 7: **Real-robot evaluation.** Human demonstrations translated into robot-domain videos by H2R-Grounder and used for future 2D trajectory supervision.

Table 4: **Real-robot manipulation results.** Closed-loop success rate (%) over 30 trials per task. Both settings use the same 25 real robot demonstrations; H2R-Grounder additionally provides 75 translated videos for co-training.

Method	Training Videos		Success Rate (%) ↑		
	Real	H2R	Sweep Block	Pick Lemon	Avg.
OpenVLA-OFT (Kim et al., 2025)	25	–	60.0	46.7	53.4
+ H2R-Grounder	25	75	**66.7**	**53.3**	**60.0**

Experimental setup. Experiments are conducted on a 7-DoF Franka Research 3 with a fixed third-person RealSense D435 camera. We consider two manipulation tasks: *Sweep the block* and *Pick-and-place lemon into bowl*, shown in Figure 7. For each task, we collect 25 teleoperated robot demonstrations with action labels and additionally use H2R-Grounder to translate 75 human demonstrations into robot-domain videos. Thus, both the baseline and our method use the same 25 action-labeled real robot demonstrations, while our method additionally leverages the 75 translated videos only through geometric supervision. Each policy is evaluated over 30 closed-loop trials, and we report task success rate.

Results. As shown in Table 4, incorporating H2R-Grounder translated videos consistently improves closed-loop manipulation performance under the same real-robot data budget. The success rate increases from 60.0% to 66.7% on *sweep block* and from 46.7% to 53.3% on *pick lemon*, resulting in an average improvement of 6.6 percentage points. These results demonstrate that H2R-Grounder is useful not only for synthesizing robot videos, but also for transferring task-relevant motion geometry from human demonstrations to downstream robot policies.

F Analysis of the 2D Hand Surrogate

Failure modes and limitations. Our 2D hand surrogate provides only an approximate gripper position and in-plane orientation, while finer-grained interaction cues, such as out-of-plane wrist orientation, gripper roll, and contact state, must be inferred by the video generator. Consequently, failures are more likely to occur for uncommon objects, complex hand configurations, or interaction patterns that are underrepresented in the robot training data or the pretrained video model. As shown in Figure 8, besides errors from hand-pose estimation, a most common failure mode on DexYCB is that the generated gripper reaches approximately the correct location but approaches the object from an unsuitable direction or fails to align with its graspable side. Richer conditioning signals, such as wrist orientation, additional finger cues, or contact states, could further improve the representation; we leave their incorporation to future work.

Surrogate error. Directly measuring the surrogate error on DexYCB is not possible because the dataset contains human interaction videos but no spatially aligned and temporally synchronized robot end-effector (EEF) trajectories. To provide a quantitative assessment, we collected 10 synchronized human–robot pick-and-place demonstrations in lab. After calibrating the camera, we compared the estimated 2D human-hand surrogate with the image-space projection of the corresponding robot EEF trajectory. The resulting position and orientation errors are summarized in Table 5.



Figure 8: **Typical failure cases of the 2D hand surrogate on DexYCB.** Although the generated gripper is positioned near the target object, limited orientation and contact information in the 2D surrogate can lead to an incorrect approach direction or misalignment with the object's graspable side.

Table 5: **Error between the estimated 2D hand surrogate and projected robot EEF pose.** Results are computed from 10 synchronized human–robot pick-and-place demonstrations collected with a calibrated camera.

Metric	Mean	Max	Min	Std.
Position error (px)	18.80	26.34	7.58	4.58
Orientation error (°)	5.73	11.29	0.04	3.49

Overall, the surrogate exhibits moderate error, which includes both pose-estimation and synchronization noise. We further observe that the position error remains relatively stable across pre-contact and post-contact phases. In contrast, the orientation error is small before contact ($\sim 1.2°$) but increases during contact ($\sim 9°$), as the thumb direction can change with force application and object geometry and therefore becomes a less accurate proxy for the robot gripper orientation.

Importantly, the generator is robust to moderate perturbations of the surrogate. On the DROID validation set, adding Gaussian noise with standard deviations of 5 pixels to position and $4°$ to orientation causes no measurable degradation in SSIM or LPIPS. We additionally provide supplementary videos with severely corrupted pose estimates. Although such cases can lead to temporal misalignment with the reference motion, the generator can still produce visually plausible robot videos, suggesting that it does not simply copy the conditioning trajectory frame by frame.

Table 6: **Decomposed VLM evaluation of physical plausibility on DexYCB.** We decompose the overall physical plausibility metric used in the main paper into four criteria and ask the VLM to evaluate each independently. Scores range from 1 to 5, with higher values indicating better performance.

Method	Gripper Structure	Contact Realism	Motion–Object Causality	Object Stability
RoboMaster	0.8	1.4	2.2	2.2
Runway Aleph	1.8	2.0	3.0	3.3
Kling	3.4	3.2	4.0	3.8
Ours	**4.0**	**3.6**	**4.4**	**4.6**

G Detailed Analysis of Physical Plausibility

To better understand the overall physical plausibility score reported in the main paper, we further decompose it into four interpretable aspects: *gripper structure*, *contact realism*, *motion–object causality*, and *object stability*. The last criterion measures whether the manipulated object remains stable when it should, rather than exhibiting unintended sliding or motion.

As shown in Table 6, H2R-Grounder achieves the highest VLM score across all four criteria on DexYCB. Specifically, it obtains scores of 4.0 for gripper structure, 3.6 for contact realism, 4.4 for motion–object causality, and 4.6 for object stability. This decomposition suggests that the improvement in overall physical plausibility is not driven by a single factor: H2R-Grounder better preserves robot structure, produces more plausible robot–object contacts, maintains more consistent cause–

effect relationships between robot motion and object response, and reduces visually implausible object motion.

H VIDEO EXAMPLES

We provide multiple translated video examples in the supplementary material, which can be viewed using the accompanying HTML viewer. If the HTML viewer encounters any issues, the videos can also be played directly from their respective folders.