

Learning Generalizable Robot Policy with Human Demonstration Video as a Prompt

Xiang Zhu^{1,2*}, Yichen Liu^{1*}, Hezhong Li², Jianyu Chen^{1,2†}

Abstract—Recent robot learning methods commonly rely on imitation learning from massive robotic dataset collected with teleoperation. When facing a new task, such methods generally require collecting a set of new teleoperation data and finetuning the policy. Furthermore, the teleoperation data collection pipeline is also tedious and expensive. Instead, human is able to efficiently learn new tasks by just watching others do. In this paper, we introduce a novel two-stage framework that utilizes human demonstrations to learn a generalizable robot policy. Such policy can directly take human demonstration video as a prompt and perform new tasks without any new teleoperation data and model finetuning at all. In the first stage, we train video generation model that captures a joint representation for both the human and robot demonstration video data using cross-prediction. In the second stage, we fuse the learned representation with a shared action space between human and robot using a novel prototypical contrastive loss. Empirical evaluations on real-world dexterous manipulation tasks show the effectiveness and generalization capabilities of our proposed method.

I. INTRODUCTION

Traditional robot learning methods typically train language-conditioned policies on large datasets collected through teleoperation[1], [2]. While effective for known tasks, this paradigm faces two fundamental limitations when encountering novel tasks. First, language instructions, though intuitive for humans, provide only categorical information and lack the rich spatial and temporal details crucial for physical manipulation. Second, adapting to new tasks usually requires collecting additional robot demonstrations, a process that is both time-consuming and expensive due to the complexity of teleoperation systems.

Humans, in contrast, can efficiently acquire new skills simply by observing others perform tasks. This observation suggests that visual demonstrations may offer a more natural and information-rich medium for teaching robots. Videos inherently capture not just what task to perform, but also how to perform it, including critical aspects like object relationships, motion trajectories, and timing. Moreover, human demonstration videos are significantly more scalable to acquire compared to robot data, whether captured in laboratory settings or obtained from existing online resources.

Recent research has only begun to explore this direction, and existing approaches exhibit notable limitations. For instance, EgoMimic [3] primarily focuses on single-task scenarios. UniSkill [4] also incorporates human videos as prompts, but these are only utilized in the Stage 1 representation learning phase to learn action embeddings, while

the Stage 2 policy learning stage does not leverage them; its goal is to highlight the effect of human videos on simple generalization through representation learning. Similarly, Gen2Act [5] and Human2Robot [6] both integrate human videos into training, but the number of human demonstrations involved is relatively limited, leaving them insufficient for achieving robust generalization when combined with sparse robot demonstrations. Overall, the field still lacks a general framework that effectively leverages human videos together with sparse robot demonstrations to enable challenging forms of generalization.

We propose a two-stage human-prompted learning framework that combines robotic datasets with human demonstration data to address challenges in task learning. In the first stage, we use a video generation model that receives a prompt video of a human performing a task and an image of the robotic hand. The model generates a video of the robot performing the task, embedding embodiment transfer information through a cross-prediction strategy. This helps the model learn a representation that captures the task, context, and target modality effectively. In the second stage, we finetune the representation using a diffusion policy, leveraging abundant human demonstration data together with robot data. A unified action space bridges the gap between the two modalities, while a cluster-based loss enhances skill separation and multi-skill imitation performance. Experiments on real-world tasks demonstrate the framework’s effectiveness in improving human-robot interaction and manipulation.

Our core contribution is achieving elementary generalization ability to new tasks and novel skills from human prompt video by proposing a novel learning framework that effectively leverages extensive human demonstration videos. Concretely, our contributions are threefold: (1) a cross-embodiment learning strategy based on cross-prediction for representation learning, (2) joint training with human video–action pairs to enhance policy generalization, and (3) the ProtoDiffusion Contrastive Policy objective to improve representation quality. Throughout the paper, we focus on evaluating generalization across three progressive levels: basic variation, novel objects, and novel skills, under the setting of human video prompting.

II. RELATED WORKS

A. Learn from Human Videos

Recent advances in robot manipulation have increasingly utilized human video data to enhance both dexterous and gripper-based manipulation. In dexterous manipulation,

¹Tsinghua University, China. ²Shanghai Qi Zhi Institute, China. *Equal contribution. †Corresponding author.

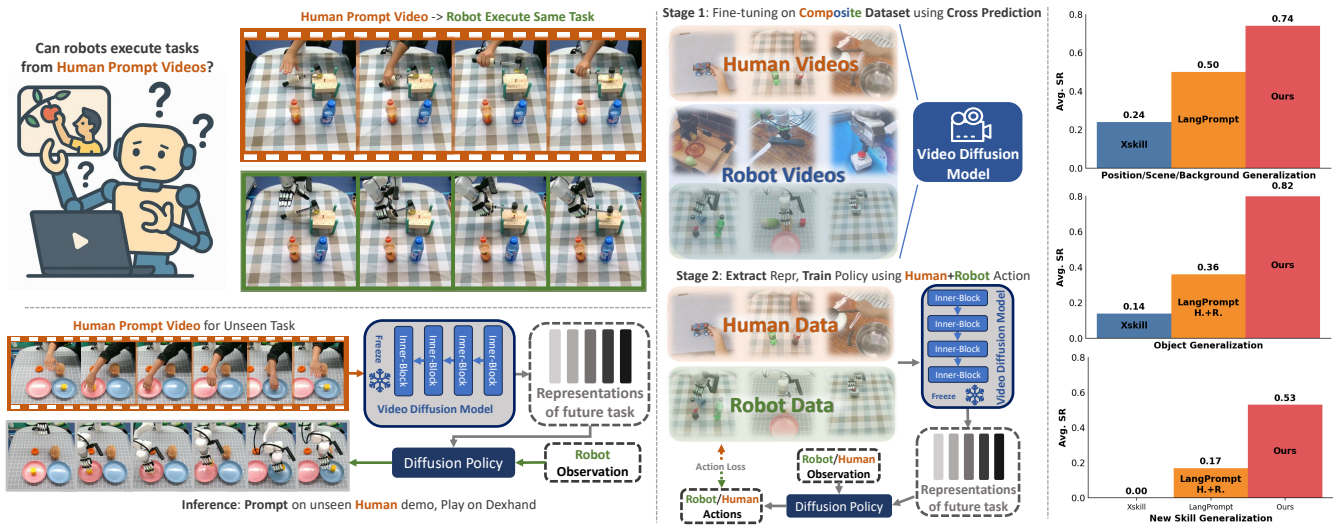


Fig. 1. First, fine-tune a video diffusion model on diverse datasets to obtain informative representations. In the second stage, use a video generation model to extract information from human prompt videos for skill learning with both human and robot data. Finally, during inference, the model prompts on unseen human demos to perform tasks based on the human input.

works like [7], [8], [9] focus on fine-grained control of multi-fingered systems, while [10] integrates affordance cues. For gripper manipulation, end-to-end video-conditioned policies such as [11], [12], [13], [14] translate visual cues into actionable policies. Approaches using paired human-robot demonstration data, including [15], [16], [17], [18], [19], [3], [6], and [20], address the domain gap by linking human actions with robot trajectories. More recently, generative video techniques like [21], [22], [5], and [23] leverage video synthesis and textual cues to generate visuomotor policies. These studies highlight the growing trend of using human video demonstrations, paired data, and generative methods to create more adaptable and robust robot manipulation policies.

B. Video as Prompt for Robotic Learning

Recent studies [24], [25] have increasingly leveraged human demonstration videos to guide robot learning. For instance, [17] mitigates the embodiment mismatch between humans and robots by converting human videos into robot-centric demonstrations via unsupervised domain adaptation and keypoint extraction. [26] enables zero-shot generalization by conditioning robot policies on pretrained video embeddings. Similarly, [11] employs cross-attention transformers to map human videos to robot actions, while [27] improves sample efficiency and generalization through contrastive learning, imitation, and limited adaptation. [14] focuses on cross-embodiment skill discovery to obtain transferable representations. Building on precisely aligned human-robot video pairs in the H&R dataset, [6] introduces a two-stage framework centered on video generation, achieving fine-grained temporal alignment and substantially enhancing both generalization and one-shot capabilities. In contrast, [4] does not rely on paired data but instead proposes a cross-embodiment representation learning framework that jointly leverages large-scale human videos and robot demonstrations, demonstrating stronger generalization in cross-embodiment tasks.

C. Diffusion Models for Robot Policy Learning

Diffusion models have shown great success in generative computer vision [28], [29], leading to their adaptation in robotic policy learning. Pioneering works [30], [31], [32] demonstrated their ability to generate denoised robot actions and capture multi-modal behavior distributions. Scaling efforts, such as [33], showcased generalization across diverse robotic platforms with transformer-based diffusion policies pretrained on the Open X-Embodiment dataset. Models like MDT [34] and RDT-1B [35] use transformer-based diffusion models, replacing the traditional U-Net. RDT-1B further unifies action representations across robots and incorporates multi-robot data for bimanual manipulation.

Diffusion-based policies can model multimodal action distributions in high-dimensional spaces [30], [36], [37]. [38] augments these models with unsupervised clustering and intrinsic rewards to maintain multiple behavior modes, while [39] adds an entropy regularizer to enhance robustness. Building on these approaches, we propose injecting SPCL’s [40], [41] prototypical contrastive loss into diffusion policy training, leveraging task category labels as supervised prototypes. Thereby sharpening policy representations along known task dimensions and enhancing multi-task imitation performance. Our method can be framed as “in-context learning”, where human demonstrations serve as the contextual basis for the policy to achieve zero-shot transfer of unseen robotic skills.

III. METHODS

In this section, we detail our framework by outlining its two-stage training process. Our goal is to enable the agent to learn and extract meaningful features and representations from human demonstration videos to perform specific tasks. The algorithm is divided into two stages. In the first stage, we develop a robust representation from a large collection of human data, robotic gripper data, and a small amount of dexterous hand data. This representation captures the

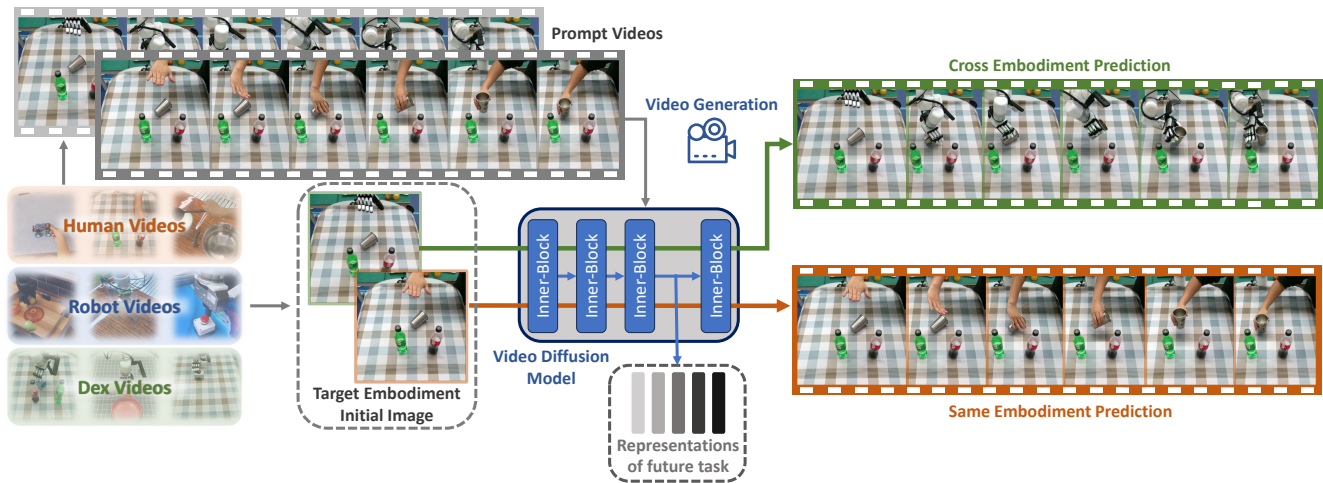


Fig. 2. **Stage 1:** Our video generation model takes as input a task demonstration from a source embodiment together with the initial scene of a target embodiment, and outputs a video in which the target embodiment performs the same task, thereby enabling embodiment transfer.

target modality, the intended task, and the scene context. In the second stage, using this solid representation, we learn human manipulation patterns from abundant human data and a limited set of high-quality dexterous hand operation data, transferring these patterns to the dexterous hand to achieve better generalization with only a small amount of valuable dexterous hand data.

A. Augment Video Generation by Cross Prediction (VGCP)

Recent advancements in video generation models have leveraged extensive online video datasets containing prior knowledge about physical world dynamics. However, these models lack data relevant to robotic manipulation. To address this, we fine-tune an existing model with a custom dataset focused on robotic and object manipulation. Our goal is to enable agents to perform tasks based on human prompt videos, particularly complex dexterous hand operations. Therefore, the video generation model is trained to include details on human manipulation, object motion, scene context, and affordances. Our dataset consists mainly of human manipulation videos, supplemented by robotic gripper and self-collected dexterous hand videos. While gripper videos differ from dexterous hand operations, they provide valuable information on object movement and scene understanding, enhancing the model’s performance.

Cross-Prediction: To further leverage our data and enhance the quality of the learned representation, we propose a method called cross-prediction. As shown in Fig.2, our video generation model receives a video prompt showing a task performed by a **source embodiment** and an initial scene for a **target embodiment**, then generates a video of the target embodiment performing the task, effectively transferring the embodiment. For example, given a prompt video of a human grasping a cup, the model produces a video of a dexterous hand performing the same action. During training, we choose cross-prediction with probability P (using different **source** and **target** embodiments) or normal-prediction with probability $1 - P$ (using the same embodiment). This approach embeds embodiment transfer information into the video

generation model, further improving the transferability of the learned representation. Our cross-prediction method randomly selects different source (s) and target (t) embodiments with probability P , and identical ones with probability $1 - P$. When the two embodiments are the same, the process mirrors typical video generation models that generate subsequent frames from an initial frame. Our goal is for the model to learn the modality transfer between human and robot, while preserving existing knowledge of their manipulation. We believe that this approach enables the model to capture a video prompt representation that encompasses the skills used by the source embodiment, the manipulated objects, and some environmental context. The fine-tuned model will then be frozen during the second stage of training.

B. Skill Learning by Human Video-Action Pair Boosting

Training manipulation policies with dexterous hand data has become a key approach in robotic manipulation. However, collecting such data through teleoperation is time-consuming and costly. To address this, we reduce reliance on robotic hand data by leveraging human hand demonstrations to enhance manipulation capabilities. The availability of human hand data is virtually unlimited, with videos easily sourced from the internet or self-collected, requiring minimal time and infrastructure. We propose incorporating human hand demonstrations into a compatible format alongside teleoperated robotic hand data, which are then jointly trained using Imitation Learning. This forms the core methodology of our Stage-2 algorithm.

Human Action Preprocessing: The objective of human data processing is to establish correspondence between human demonstration data and robot motion data. Given that we only have third-person view RGB videos of human demonstrations, while the robot data contains both third-person view RGB videos and joint state information. The robot’s end-effector state comprises two components: 6D Wrist pose (wrist position $\in \mathbb{R}^3$ and wrist orientation $\in \text{SO}(3)$) and finger joint θ^{12} . We employ the hand-tracking method WiLoR[42] for hand localization in video frames

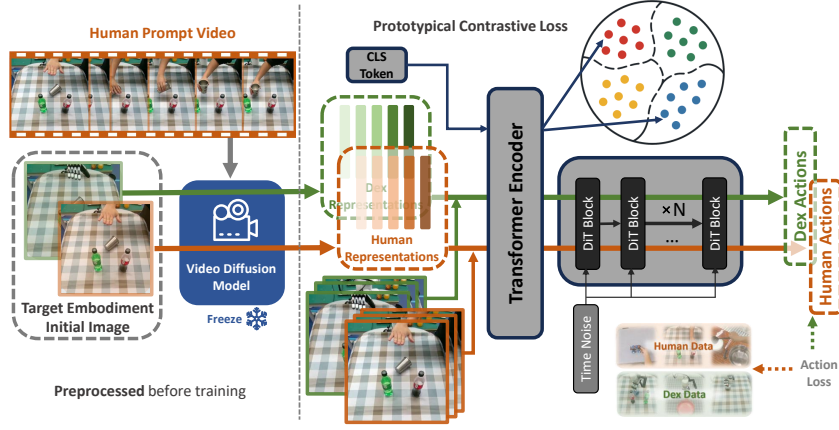


Fig. 3. **Stage 2:** We use the diffusion model trained in the first stage to combine information from human prompt videos with the target embodiment, generating an informative representation. This representation, along with a shared action between human and robot and a prototypical contrastive loss, enables the diffusion policy to learn task and skill information for both.

and 3D hand mesh reconstruction, and the model outputs 21 right-hand keypoints $J_{h_i} \in \mathbb{R}^3$ for each frame.

To extract the 6D wrist pose from 21 right-hand keypoints, we take the wrist keypoint \mathbf{W} as the position. For orientation, a local right-handed coordinate frame is constructed using \mathbf{W} , the index-finger joint \mathbf{I} , and the middle-finger joint \mathbf{M} . The x-axis is defined along the direction from \mathbf{W} to \mathbf{M} , the z-axis as the normalized normal of the plane formed by $(\mathbf{W}, \mathbf{I}, \mathbf{M})$ pointing toward the palm, and the y-axis is obtained via the right-hand rule. The resulting orthonormal basis forms the rotation matrix: $\mathbf{R} = [\mathbf{x} \ \mathbf{y} \ \mathbf{z}]$, which, together with \mathbf{W} , gives the complete 6D pose representation.

In addition, we compute the 3D position of the human hand in the camera coordinate system using the intrinsic parameters and transform it into the robot’s base coordinate system. To establish joint-level correspondence between the human and robot systems, motion retargeting is required, mapping the kinematic configuration of the human demonstrator’s hand to that of the robotic hand. Following AnyTeleop[43], we formulate this problem as an optimization, thereby obtaining the data that map human hand joints in demonstration videos to the corresponding dexterous hand joints.

Representation Conditioned Diffusion Policy: Our framework employs a diffusion policy that models the conditional action distribution $p(a_t|s_t, z)$, where: $s_t = f_{resnet34}(o_t) \in \mathbb{R}^{1000}$ is the visual observation feature extracted by a pretrained ResNet-34, $z \in \mathbb{R}^{4096}$ denotes the stage-one representation. Here we define the **human representation** as the generated representation with a human video as the prompt and a human image as the initial frame, while the **robot representation** as the representation with a human video as the prompt and a robot image as the initial frame. $a_t \in \mathbb{R}^{19}$ is the action vector at time t . The diffusion process operates in the action space through:

$$a_t^i = \sqrt{\alpha_i} a_0 + \sqrt{1 - \alpha_i} \epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I}) \quad (1)$$

while the reverse process learns to predict and remove the noise through a transformer-based denoising network ϵ_θ .

For dexterous hand control tasks, we decompose the action space into three distinct components: finger joint angles

$a_t^{finger} \in \mathbb{R}^{12}$, wrist orientation $a_t^{rot} \in \text{SO}(3)$, and wrist position $a_t^{pos} \in \mathbb{R}^3$. This decomposition enables specialized handling of each action modality through separate prediction heads in the denoising network. The training loss combines weighted component losses:

$$\mathcal{L}_{action} = \lambda_f \mathcal{L}_{finger} + \lambda_r \mathcal{L}_{rotation} + \lambda_p \mathcal{L}_{postion} \quad (2)$$

where:

$$\mathcal{L}_* = \mathbb{E}_{i, a_0, \epsilon} [|\epsilon_* - \epsilon_\theta(a_t^i, i, h_0)|^2] \quad (3)$$

$h_0 = [s_t, z]W_e + b_e$ is the concatenated input token, W_e is the embedding matrix, b_e is the bias. $\lambda_f, \lambda_r, \lambda_p$ are weighting coefficients determined through validation.

C. ProtoDiffusion Contrastive Policy (PDCP)

Siamese Prototypical Contrastive Learning (SPCL)[40], [41] first groups feature embeddings into prototypes via K-Means and then applies a Siamese-style metric loss to pull together embeddings within each prototype while pushing apart those from different prototypes, alongside a prototypical cross-entropy loss that treats prototype assignments as soft labels to sharpen each sample’s affinity to its cluster. By leveraging learned cluster structure rather than individual instances, SPCL mitigates false negatives and yields more stable positive sets, thereby enhancing semantic discriminability in self-supervised learning.

To adapt SPCL for diffusion policy training, we add a learnable clustering token to the input of our Diffusion Transformer (DiT) encoder and use the encoder’s output for that token—denoted h —as a prototype-aware latent. During training we jointly optimize three losses:

(1) the NT-Xent contrastive loss \mathcal{L}_{contra} , which treats tasks sharing the same skill as positives and tasks of different skills as negatives (analogous to data-augmented positives in vision); (2) a prototypical cross-entropy loss \mathcal{L}_{proto} , which encourages each the learned latent h to align with the cross-entropy distribution over its K-Means prototype label, promoting tighter intra-prototype clustering and clearer inter-prototype separation. And (3) a Siamese-style metric loss \mathcal{L}_{metric} at the prototype level, which further enforces proximity among same-prototype samples in metric space

while repelling different-prototype samples, thus reducing semantic confusion. Due to space constraints, we omit the detailed formulations of \mathcal{L}_{contra} , \mathcal{L}_{proto} , and \mathcal{L}_{metric} , which are similar in spirit to those in [40], [41], but adapted to our setting with modifications in sample construction.

The total loss for PDCP consist of the action loss and a weighted sum of the mentioned loss with coefficient w_p, w_p, w_m :

$$\mathcal{L} = \mathcal{L}_{action} + w_c \mathcal{L}_{contra} + w_p \mathcal{L}_{proto} + w_m \mathcal{L}_{metric} \quad (4)$$

These objectives encourage task-discriminative and modality-aligned representations, improving cross-modal generalization.

IV. EXPERIMENTS

In the following section, we present a comprehensive set of experiments to validate our proposed method. We deploy the algorithm on the dexterous-hand Xhand, assigning it 13 fundamentally distinct skills, such as pick-and-place, pressing, pouring, flipping cups, push operations, ball flicking and so on. These skills are further decomposed into more than 100 concrete tasks. We evaluate generalization across positions, scenes, objects, and, most critically, across skills. Our experiments are designed to address the following questions: (1) Can cross-prediction enhance the performance, effectiveness, and representational capacity of the learned representation? (2) Is it feasible to use different video prompts in combination with a video generation model for skill extraction, and what is the resulting impact? (3) Can the extracted representation and DPCP enhance the position/scene/background generalization performance of the diffusion policy? (4) Does incorporating human action data improve the object generalization ability of the diffusion policy? (5) Can the final learned policy demonstrate the ability to generalize to previously unseen robot skills as well as exhibit "in-context learning" capabilities?

In our real-robot experiments, we conducted a systematic evaluation of the proposed policy. For Stage-1 representation learning, we used over 250,000 videos of human, gripper, and dexterous-hand demonstrations from RH20T[44], Something-Something[45], HOI4D[46], Bridge[47] and our in-house dexterous manipulation dataset collected via Apple Vision Pro teleoperation.

For Stage-2 policy learning, we employed about 7,000 human and dexterous-hand videos, including paired demonstrations collected in-house and additional human videos from HOI4D[46]. The ratio of human to dexterous-hand videos was about 5:1, ensuring the dataset remained sparse in robot data but rich in human demonstrations.

At inference, our method extracts task representations from human prompt videos, fuses them with the robot's real-time observations, and then feeds them into the stage-2 policy to generate actions. We use the DiT from RDT-1B[35] as our policy network to generate continuous actions in a unified action space conditioned on multimodal inputs and embodiment embeddings. The actions are executed by

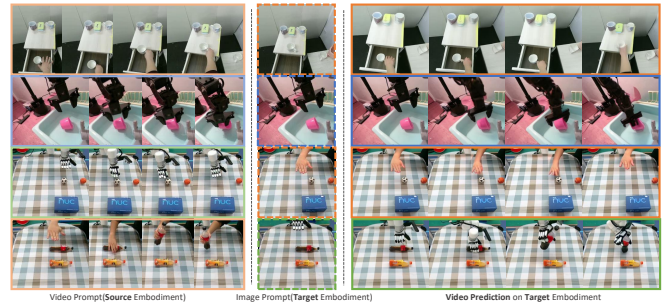


Fig. 4. Examples of Cross-Prediction Video Generation.

the robot's built-in position controller at 10 Hz. We compare our approach against the following baselines:

- **Representation+Robot.** Policies trained solely on robot data without incorporating any human demonstrations.
- **Language+Robot+Human.** Policies conditioned on CLIP-encoded task labels (e.g., "pour water") as inputs, rather than on learned skill representations.
- **Language-conditioned Diffusion Policy.** The language input is processed with a CLIP text encoder and subsequently incorporated as a conditional input to the Diffusion Policy[30], which then generates the corresponding action.
- **Representation+Robot+Human.** Policies jointly trained on both robot and human data, optimized using learned skill representations.
- **XSkill[14].** In the first-stage representation learning, XSkill employed the same human and robot datasets as ours. However, due to methodological constraints, its second-stage training relied solely on robot data.
- **UniSkill[4].** In UniSkill[4], the first stage trains Inverse and Forward Skill Dynamics using the same human and robot data as our method. The second stage learns the policy with robot data, and inference is conditioned on human videos as prompts.

For each algorithm and each generalization level, we conducted 60-80 trials to ensure statistical robustness. To guarantee a fair comparison, all methods were evaluated under identical experimental conditions: For each evaluation task, the generalization level (position/scene/background&object&skill) and the prompt (either human video prompts or corresponding textual descriptions) are the same. Performance was measured by two metrics: Success Rate (SR), defined as the fraction of successful trials, and Task Score, which quantifies task progress by decomposing long-horizon tasks into subtasks, assigning partial credit, and normalizing the cumulative score to 100.

A. Evaluation for Cross Prediction Video Generation

Our cross-prediction mechanism extracts the task performed by the source embodiment from the prompt video and, using the scene context and target embodiment information from another given image, generates a video of the target embodiment executing that task. As shown in Fig.4, we show two examples of same-embodiment prediction alongside with two examples of cross-embodiment prediction. The generated videos faithfully capture the task details from the

prompt, and in the cross-embodiment cases, the generated target embodiment motion and scene context align precisely with the original task. This demonstrates that our model has learned not only the physical dynamics of the world but also learned how to transfer task and skill across different embodiments.

B. Evaluation for the Learned Representations

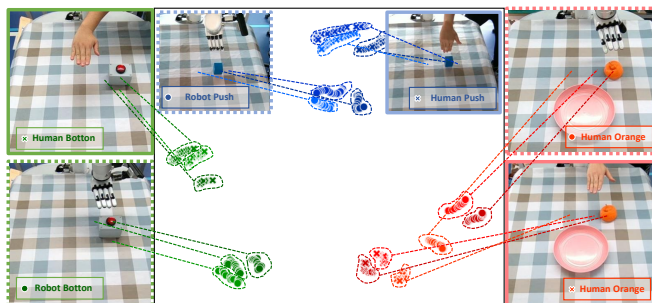
As shown in previous experiments, applying cross-prediction enables the video generation model to convert between human and robot embodiments while retaining its original capabilities. However, our goal is to extract the task information performed by the source embodiment from the prompt video using the cross-prediction fine-tuned SVD, and then transfer that information to the target embodiment. So, how do we evaluate and verify the performance of the representations extracted by the video generation model? We conducted the following experiments:

1) t-SNE Visualization of the Learned Representation:

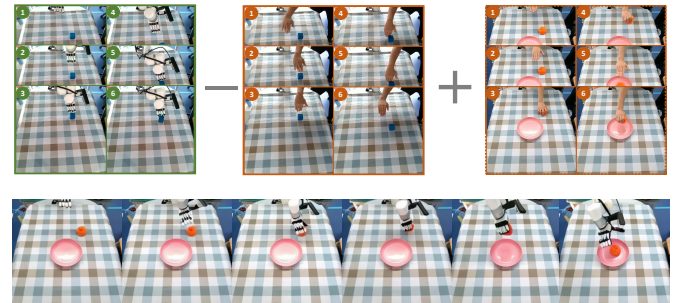
We conducted a t-SNE visualization experiment. We applied t-SNE to reduce the dimensionality of the skills extracted from the **human representation** (with a human video as the prompt and a human image as the initial frame) and the **robot representation** (with a human video as the prompt and a robot image as the initial frame), and then visualized them on a two-dimensional plane. The results, shown in Fig.5a, reveal that tasks of the same category are grouped by similar color schemes, with identical hues indicating representations at different time points. Dashed lines highlight the positions of target objects in the two-dimensional space. The plot demonstrates that representations can effectively distinguish tasks, with the same task’s representations clustering closely based on embodiment, while still distinguishing target object positions. Additionally, representations from different time steps within the same video are tightly grouped yet distinguishable, reflecting the capture of continuous temporal information. These results show that the learned representations not only encode embodiment and task information but also capture object positioning and motion details, which is challenging to obtain with language prompts, showcasing the advantage of video prompts.

2) Vector Properties of the Learned Representations:

In word embeddings, a fascinating phenomenon is observed: these embeddings exhibit an intrinsic vector structure. For



(a) t-SNE visualization of learned representations.



(b) Vector properties of learned representations.

Fig. 5. Evaluation for the Learned Representations

TABLE I

SUCCESS RATE (SR) AND SCORE METRICS ON POSITION, SCENE, AND BACKGROUND GENERALIZATION

Method	Position		Scene		Background	
	SR	Score	SR	Score	SR	Score
Xskill[14]	0.31	60.1	0.27	46.4	0.18	47.3
Language+R.+H.	0.58	69.5	0.56	55.6	0.36	50.9
Repr.+R.	0.68	77.9	0.67	81.1	0.55	69.1
Repr.+R.+H.	0.74	81.6	0.67	84.4	0.64	74.5
Repr.+R.+H.+DPCP (Ours)	0.79	85.3	0.69	88.9	0.73	80.9

example, taking the representation of “king,” subtracting that of “queen,” and then adding the representation of “nurse” yields that of “doctor.” This arithmetic property is celebrated as both elegant and intriguing. Thus, we aim to explore whether our learned representations possess a similar vectorial nature. We subtracted the human representation from the corresponding robot representation for the “push blue box” task and then added the human representation for the “grasp orange” task. Using this composite representation, we generated a video that, as shown in the Fig.5b, ultimately depicted a robot executing the “grasp orange” task. Moreover, we conducted tests on other tasks—such as “push button” and “pour water” and observed similarly intriguing results. This demonstrates that our learned representations indeed exhibit notable vector properties.

C. Position, Scene, Background Generalization

To assess the generalization capabilities of our policy, we conducted comprehensive testing across three critical dimensions. **Positional Generalization** means randomly relocating target objects within the workspace. **Scene Generalization** means randomly changing or replacing inoperable objects within the workspace. **Background Generalization** means testing under different background (tablecloth, ...).

As shown by the comparative results in Table I, our method consistently outperforms the four baseline approaches across all three generalization scenarios. A further comparison between our Stage-1 representation-guided policy and representation learning methods conditioned on language or based on XSkill[14] reveals the fundamental advantages of our learned representations in encoding richer task-relevant information. This advantage arises from the

integration of end-effector positional information and partial operational knowledge within our representation learning framework. Our analysis also demonstrates that the proposed PDCP algorithm effectively mitigates mis-grasping issues when handling closely spaced objects. We argue that while the Stage-1 representation primarily captures spatial variations, the PDCP module provides complementary functionality by enabling task-aware feature clustering.

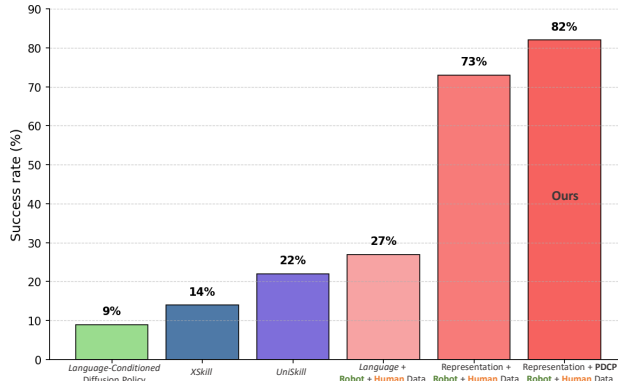


Fig. 6. Evaluation on Obj Generalization

D. Object Generalization

To investigate the complementary benefits of our representation learning framework and human demonstration data, we design object-level generalization experiments to evaluate their individual and combined effects. While our representations perform well across all generalization scenarios, these experiments isolate and quantify the contribution of human data with novel object instances.

We define a **New Object** as one that does not appear in robot data during the second-stage training but is present in human demonstrations. Accordingly, for our method and the baselines that incorporate human videos in stage-2 training (Representation+Robot+Human Data and Language+Robot+Human Data), we use all available robot data excluding demonstrations involving the new object, together with the full set of human data, which includes demonstrations of that object. In contrast, for the baselines that train solely on robot videos in stage-2 (Language-Conditioned Diffusion Policy, XSkill[14], and UniSkill[4]), we use only the robot data while excluding demonstrations involving that new object.

As shown in Figure 6, we evaluate four distinct tasks involving unseen objects and compare our approach against baseline methods. At the new object generalization level, compared to training without human demonstrations, the success rate on new objects improves significantly, underscoring the role of human data in transferring knowledge of object affordances and manipulation strategies. Our representation further complements this by supplying spatial cues and partial operational knowledge for tasks absent from the robot training set, while human demonstrations offer complementary insights into novel objects and strategies. In contrast, the XSkill and UniSkill baselines cannot leverage human actions during the second-stage policy learning process, leaving them

unable to transfer knowledge from human demonstrations and resulting in poorer performance on unseen-object tasks.

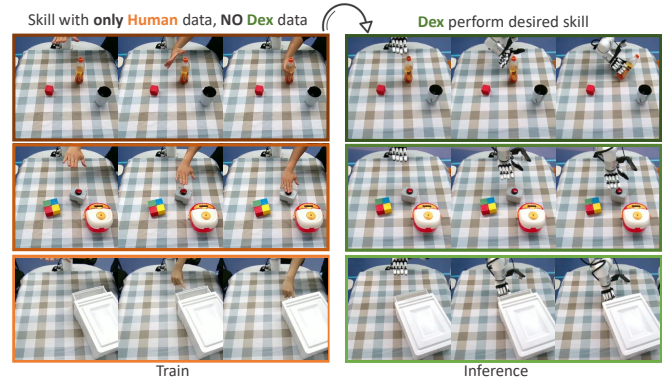


Fig. 7. Unseen Skill transfers

E. Skill Generalization

Having shown that our policy can generalize to objects absent from the robot training dataset, we next examine whether it can also acquire entirely new skills missing from the original robot data. To this end, we evaluate four novel tasks: grasping an orange juice bottle, pressing a button, closing a drawer, and opening a carton. For each skill, the policy was trained on all available robot trajectories **excluding those involving the target skill**, while jointly incorporating the full set of human demonstrations that included the skill. For instance, in the case of pressing skill, all robot trajectories involving pressing action were excluded, whereas the human dataset retained demonstrations for all pressing tasks. To improve visual alignment between human and robot demonstrations, we further applied a pre-processing trick: in human videos, a solid circle was overlaid on the detected wrist position with its size inversely proportional to camera distance; in robot videos, the wrist was annotated in the same way. As illustrated in Figure 7, our approach enables the policy to “watch” human demonstrations (left) and successfully execute previously unseen skills on the robot (right), showcasing a degree of video-based “in-context learning”.

For baselines that incorporate human videos in stage-2 training (Language+R.+H. and Repr.+R.+H.+DPCP (w/o Circle)), we use all available robot data excluding the target skill, along with the full set of human demonstrations. For baselines relying solely on robot videos (XSkill[14], UniSkill[4], and Language-Conditioned Diffusion Policy), only robot data excluding the target skill is used.

As shown in Table II, our method consistently outperforms these baselines. Unlike XSkill and UniSkill, which cannot exploit human data in stage-2, our approach effectively leverages human demonstrations to acquire novel skills and improve execution on unseen skills. Moreover, adding circle tags reduces the human–robot visual gap and provides position-aware cues, yielding more accurate actions. These results underscore the importance of human demonstrations and image-level alignment, and validate the effectiveness of our method in enabling robots to generalize to entirely unseen skills.

TABLE II
PERFORMANCE ON NOVEL SKILL GENERATION TASKS IN TERMS OF
SUCCESS RATE (SR) AND SCORE.

Method	Grasp Bottle		Press Button		Close Drawer		Open Carton	
	SR	Score	SR	Score	SR	Score	SR	Score
XSkill [14]	0.00	0.0	0.00	0.0	0.00	0.0	0.00	0.0
UniSkill [4]	0.00	0.0	0.00	0.0	0.00	0.0	0.00	0.0
Language-conditioned Diffusion Policy	0.00	0.0	0.00	0.0	0.00	0.0	0.00	0.0
Language+R.+H.	0.18	27.3	0.22	28.9	0.25	41.3	0.08	17.7
Repr.+R.+H.+DPCP(w/o Circle)	0.27	43.3	0.25	36.3	0.35	49.4	0.23	36.9
Repr.+R.+H.+DPCP(Ours)	0.47	65.3	0.56	63.8	0.65	66.5	0.46	50.8

V. CONCLUSIONS

This work fully exploits human data across both the representation learning and diffusion-based policy stages. By using human videos as prompts, which provide substantially richer information than language inputs, our approach achieves superior performance on dexterous hand manipulation tasks.

REFERENCES

- [1] A. Brohan *et al.*, “Rt-1: Robotics transformer for real-world control at scale,” *ArXiv*, vol. abs/2212.06817, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:254591260>
- [2] A. Brohan *et al.*, “Rt-2: Vision-language-action models transfer web knowledge to robotic control,” *ArXiv*, vol. abs/2307.15818, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:260293142>
- [3] S. Kareer *et al.*, “Egomimic: Scaling imitation learning via egocentric video,” *arXiv preprint arXiv:2410.24221*, 2024.
- [4] H. Kim *et al.*, “Uniskill: Imitating human videos via cross-embodiment skill representations,” *arXiv preprint arXiv:2505.08787*, 2025.
- [5] H. Bharadhwaj *et al.*, “Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation,” *arXiv preprint arXiv:2409.16283*, 2024.
- [6] S. Xie *et al.*, “Human2robot: Learning robot actions from paired human-robot videos,” *ArXiv*, vol. abs/2502.16587, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:276575296>
- [7] Z. Chen *et al.*, “Vividex: Learning vision-based dexterous manipulation from human videos,” *arXiv preprint arXiv:2404.15709*, 2024.
- [8] Y. Qin *et al.*, “Dexmv: Imitation learning for dexterous manipulation from human videos,” in *European Conference on Computer Vision*. Springer, 2022, pp. 570–587.
- [9] C. Wang *et al.*, “Dexcap: Scalable and portable mocap data collection system for dexterous manipulation,” *arXiv preprint arXiv:2403.07788*, 2024.
- [10] A. Kannan *et al.*, “Defit: Dexterous fine-tuning for real-world hand policies,” *arXiv preprint arXiv:2310.19797*, 2023.
- [11] V. Jain *et al.*, “Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers,” *arXiv preprint arXiv:2403.12943*, 2024.
- [12] C. Yuan *et al.*, “General flow as foundation affordance for scalable robot learning,” *arXiv preprint arXiv:2401.11439*, 2024.
- [13] Y. Ju *et al.*, “Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation,” in *European Conference on Computer Vision*. Springer, 2024, pp. 222–239.
- [14] M. Xu *et al.*, “Xskill: Cross embodiment skill discovery,” in *Conference on robot learning*. PMLR, 2023, pp. 3536–3555.
- [15] C. Wang *et al.*, “Mimicplay: Long-horizon imitation learning by watching human play,” *arXiv preprint arXiv:2302.12422*, 2023.
- [16] L. Smith *et al.*, “Avid: Learning multi-stage tasks via pixel-level translation of human videos,” *arXiv preprint arXiv:1912.04443*, 2019.
- [17] H. Xiong *et al.*, “Learning by watching: Physical imitation of manipulation skills from human videos,” in *2021 IEEE/RSSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 7827–7834.
- [18] C. Wen *et al.*, “Any-point trajectory modeling for policy learning,” *arXiv preprint arXiv:2401.00025*, 2023.
- [19] J. Gu *et al.*, “Rt-trajectory: Robotic task generalization via hindsight trajectory sketches,” *arXiv preprint arXiv:2311.01977*, 2023.
- [20] S. Ye *et al.*, “Latent action pretraining from videos,” *arXiv preprint arXiv:2410.11758*, 2024.
- [21] J. Liang *et al.*, “Dreamitate: Real-world visuomotor policy learning via video generation,” *arXiv preprint arXiv:2406.16862*, 2024.
- [22] Y. Du *et al.*, “Learning universal policies via text-guided video generation,” *Advances in neural information processing systems*, vol. 36, pp. 9156–9172, 2023.
- [23] M. Shridhar, Y. L. Lo, and S. James, “Generative image as action models,” *arXiv preprint arXiv:2407.07875*, 2024.
- [24] M. Chang and S. Gupta, “One-shot visual imitation via attributed waypoints and demonstration augmentation,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5055–5062.
- [25] H. Xiong *et al.*, “Robotube: Learning household manipulation from human videos with simulated twin environments,” in *Conference on Robot Learning*. PMLR, 2023, pp. 1–10.
- [26] E. Chane-Sane, C. Schmid, and I. Laptev, “Learning video-conditioned policies for unseen manipulation tasks,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 909–916.
- [27] Z. Qian *et al.*, “Contrast, imitate, adapt: Learning robotic skills from raw human videos,” *IEEE Transactions on Automation Science and Engineering*, 2024.
- [28] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *ArXiv*, vol. abs/2006.11239, 2020.
- [29] Y. Song *et al.*, “Score-based generative modeling through stochastic differential equations,” *ArXiv*, vol. abs/2011.13456, 2020.
- [30] C. Chi *et al.*, “Diffusion policy: Visuomotor policy learning via action diffusion,” *ArXiv*, vol. abs/2303.04137, 2023.
- [31] T. Pearce *et al.*, “Imitating human behaviour with diffusion models,” *ArXiv*, vol. abs/2301.10677, 2023.
- [32] M. Reuss *et al.*, “Goal-conditioned imitation learning using score-based diffusion policies,” *ArXiv*, vol. abs/2304.02532, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:257952177>
- [33] O. M. Team *et al.*, “Octo: An open-source generalist robot policy,” *ArXiv*, vol. abs/2405.12213, 2024.
- [34] M. Reuss *et al.*, “Multimodal diffusion transformer: Learning versatile behavior from multimodal goals,” *ArXiv*, vol. abs/2407.05996, 2024.
- [35] S. Liu *et al.*, “Rdt-1b: a diffusion foundation model for bimanual manipulation,” *ArXiv*, vol. abs/2410.07864, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:273233386>
- [36] Y. Wang *et al.*, “Diffusion actor-critic with entropy regulator,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 54 183–54 204, 2024.
- [37] V. Jain, T. Akhound-Sadegh, and S. Ravanbakhsh, “Sampling from energy-based policies using diffusion,” *arXiv preprint arXiv:2410.01312*, 2024.
- [38] S. Li *et al.*, “Learning multimodal behaviors from scratch with diffusion policy gradient,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 38 456–38 479, 2024.
- [39] R. Zhang *et al.*, “Entropy-regularized diffusion policy with q-ensembles for offline reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 98 871–98 897, 2024.
- [40] J. Li *et al.*, “Prototypical contrastive learning of unsupervised representations,” *arXiv preprint arXiv:2005.04966*, 2020.
- [41] S. Mo, Z. Sun, and C. Li, “Siamese prototypical contrastive learning,” *arXiv preprint arXiv:2208.08819*, 2022.
- [42] R. A. Potamias *et al.*, “Wilor: End-to-end 3d hand localization and reconstruction in-the-wild,” *ArXiv*, vol. abs/2409.12259, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:272753318>
- [43] Y. Qin *et al.*, “Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system,” *ArXiv*, vol. abs/2307.04577, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:259367735>
- [44] H.-S. Fang *et al.*, “Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot,” *arXiv preprint arXiv:2307.00595*, 2023.
- [45] R. Goyal *et al.*, “The” something something” video database for learning and evaluating visual common sense,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850.
- [46] Y. Liu *et al.*, “Hoi4d: A 4d egocentric dataset for category-level human-object interaction,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20 981–20 990, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:247222786>
- [47] H. R. Walke *et al.*, “Bridgedata v2: A dataset for robot learning at scale,” in *Conference on Robot Learning*. PMLR, 2023, pp. 1723–1736.