Train Robots in a JIF: Joint Inverse and Forward Dynamics with Human and Robot Demonstrations

Gagan Khandate*†, Boxuan Wang*‡, Sarah Park*†, Weizhe Ni†‡,
Joaquin Palacios‡, Kathryn Lampo ‡, Philippe Wu‡, Rosh Ho‡, Eric Chang‡ and Matei Ciocarlie‡

†Dept. of Computer Science ‡Dept. of Mechanical Engineering *joint first authorship

Columbia University, New York, NY 10027, USA
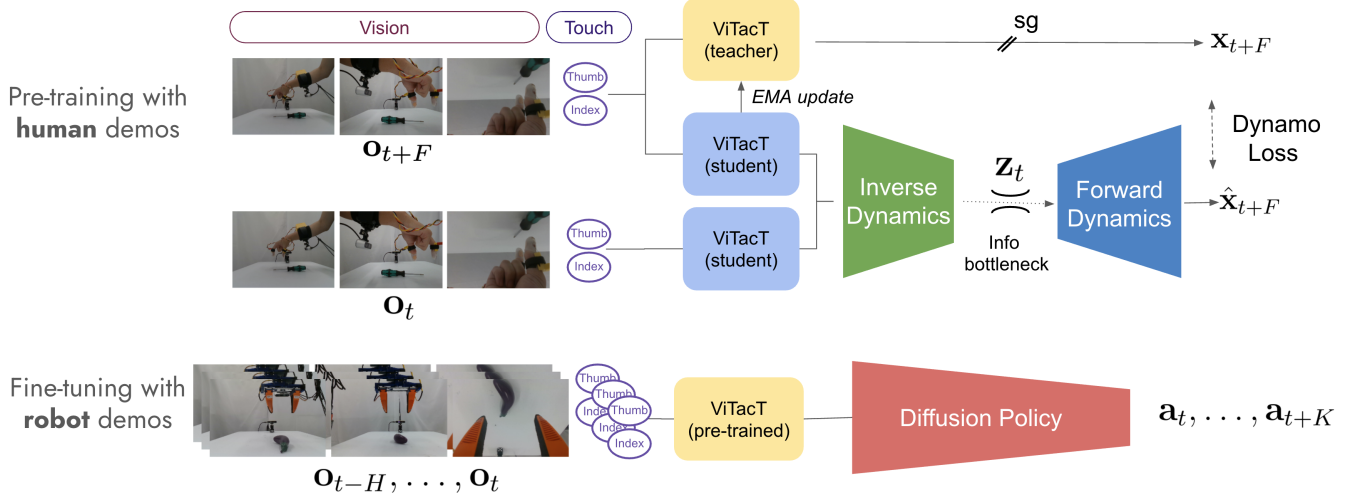
Corresponding email: gagank@cs.columbia.edu

Fig. 1: We introduce a framework for pre-training using multi-modal human demonstrations. We extract latent state representations by **J**ointly learning **I**nverse and **F**orward dynamics - JIF. Dynamics driven state representations maximize manipulation centric information in the human demonstrations allowing efficient and generalizable imitation learning by fine-tuning with a small number of robot demonstrations.

Abstract—**Pre-training on large datasets of robot demonstrations is a powerful technique for learning diverse manipulation skills but is often limited by the high cost and complexity of collecting robot-centric data, especially for tasks requiring tactile feedback. This work addresses these challenges by introducing a novel method for pre-training with multi-modal human demonstrations. Our approach jointly learns inverse and forward dynamics to extract latent state representations, towards learning manipulation specific representations. This enables efficient fine-tuning with only a small number of robot demonstrations, significantly improving data efficiency. Furthermore, our method allows for the use of multi-modal data, such as combination of vision and touch for manipulation. By leveraging latent dynamics modeling and tactile sensing, this approach paves the way for scalable robot manipulation learning based on human demonstrations.**

I. INTRODUCTION

Pre-training on large datasets of robot demonstrations [1, 2] using imitation learning [3–5] or offline reinforcement learning [6, 7] is rapidly becoming a standard technique for learning diverse robot manipulation skills, even pushing towards the development of foundation models [8–14]. However, these methods rely on extensive collections of robot demonstrations, typically gathered through teleoperation, where the demonstrator remotely operates the robot hardware that is performing the task. This process is expensive to scale; it is also impractical if complex hardware is needed, as is the case for tasks like dexterous manipulation which benefit from real-time tactile feedback for the teleoperator.

In contrast, human demonstrations, where the demonstrator directly performs the task of interest, are significantly less costly to acquire. Furthermore, flexible tactile sensors are making continuous strides [15], and instrumenting human fingertips with such sensors can preserve tactile sensation for the demonstrator while still recording tactile feedback throughout the demonstration, alongside commonly used visual observations. This suggests that pre-training with such multi-modal human demonstrations offers a promising avenue for overcoming the limitations of robot-centric data collection.

Prior visual representation learning work[16–23], such as masked visual pre-training (MVP) [20, 24–26], has explored leveraging visual human demonstrations such as videos for state representation learning. However, these approaches are inherently dynamics-agnostic, and often rely

on reconstruction-based objectives that do not capture the underlying system dynamics critical for manipulation tasks. Moreover, the reliance on reconstructing high-dimensional sensory inputs makes these methods computationally expensive and challenging to scale. In contrast, dynamics-driven representation learning provides a more promising alternative by encoding task-relevant information.

Dynamics-driven representation learning, which involves learning a forward dynamics model, presents challenges when applied to multi-modal human demonstrations. Obtaining action labels, such as hand joint motion, requires costly sensing systems, and existing methods often rely on reconstructing high-dimensional observations, making them computationally expensive. Approaches like LAPO [27] mitigate these issues by jointly learning inverse and forward dynamics to capture manipulation-specific information without action labels. Building on this, DynaMo [28] enhances efficiency through teacher-student distillation while preserving the benefits of dynamics-driven learning. However, these methods focus on using robot demonstrations alone.

We extend this family of approaches to pre-training with multi-modal human demonstrations, enabling representation learning without action labels or costly reconstructions. By adapting the joint inverse-forward dynamics paradigm and incorporating teacher-student distillation, our method significantly improves computational efficiency. This makes representation learning scalable and practical for real-world robotics applications.

We demonstrate the effectiveness of our representation learning method through imitation learning. Our results show significant improvements in demonstration complexity and generalization, achieving strong performance with a small number of robot demonstrations—tasks that previously required larger datasets. Furthermore, our approach enables the introduction of a novel data collection paradigm: instrumented visuo-tactile human demonstrations. By equipping human demonstrators with tactile sensors, we capture rich tactile information, leading to improved robustness in the learned manipulation skills.

Overall, our method offers a promising path towards scaling robot learning via pre-training with human demonstrations to a wider range of complex tasks, such as multi-fingered manipulation, where traditional teleoperation is difficult. The key contributions of this work are:

1) We introduce a pre-training framework centered on multi-modal human demonstrations via learning inverse and forward dynamics through teacher-student distillation for computational efficiency.
2) We present a novel paradigm of robot learning from instrumented visuo-tactile human demonstrations, capturing rich tactile information alongside visual data.
3) To the best of our knowledge, we are the first to demonstrate the benefit of incorporating tactile observations in human demonstrations for improving the robustness of learned robot manipulation skills.

II. RELATED WORK

Using human demonstrations is a well-established approach in robot learning, often involving video demonstrations to learn reward models for reinforcement learning. For example, Shao et al. [29] classify human demonstrations into task categories and use the classification score as a reward signal to train a robot policy. Similarly, Chen et al. [30] train a discriminator to determine whether two videos depict the same task. By leveraging the similarity score as a reward and combining it with an action-conditioned video prediction model, these methods enable robots to execute new tasks based on human demonstrations. Additionally, ViPER [31] employs a video prediction transformer and uses the log-likelihood of predicted frames as a reward function to guide task execution.

Other works focus on imitation learning, developing frameworks to acquire semantic skills that can transfer across domains [32–35]. To leverage large-scale human video datasets, such as SomethingSomething-V2 [36] and Ego4D [37], self-supervised learning methods have been applied to develop visual representations, demonstrating benefits for downstream policy learning [16–23]. For example, extciteParisi2022-ea use MoCo representations, while R3M [38] applies masked autoencoders (MAE). Other works [20, 24–26] explore masked visual pre-training (MVP), though primarily with robot demonstrations. However, reconstruction-based methods like these are computationally expensive and face scalability challenges.

Multi-modal demonstrations are particularly valuable for complex tasks, such as multi-fingered dexterous manipulation, where precise, contact-rich interactions are required. Visual demonstrations alone are often insufficient to capture the detailed sensory feedback needed for these tasks. However, reconstruction based representation learning is too expensive and challenging. Recent methods have sought to overcome these challenges through knowledge self-distillation, a more efficient alternative to reconstruction. Techniques such as BYOL [39] and DINO [40] employ a teacher-student framework with identical networks, learning invariant representations by aligning predictions across augmented views of the same input. This approach reduces computational overhead, making it particularly suitable for multi-modal data.

Dynamics-driven representation learning has also been explored in various works. Early approaches, such as temporal contrastive learning [41], learn representations by contrasting temporally adjacent frames. Other methods, such as those proposed by Edwards et al. [42], simultaneously learn a world model and a latent policy using visual demonstrations. Similarly, Schmidt et al. [27] enhance representation learning by combining a forward dynamics model with an inverse dynamics model and an information bottleneck. Recently, dynamics-driven representations have been integrated with knowledge self-distillation (DynaMo [28]) , making these approaches particularly useful when explicit action information is unavailable, as with human demonstrations.

Building on this line of research, our work extends DynaMo

[28] by incorporating multi-modal human demonstrations, including tactile sensing, to improve downstream imitation learning with limited robot demonstrations. A key distinction of our approach is that we are the first to apply dynamics-driven representation learning for multi-modal data and the first to implement self-distillation for such data. Unlike prior efforts that rely on in-the-wild, uni-modal video datasets, we collect task-specific human demonstrations tailored to our downstream tasks. Additionally, we integrate low-cost tactile sensors for instrumented human demonstrations, learning visuo-tactile representations for contact rich tasks.

Although several other works have collected tactile data from human demonstrators [43–52], these methods have not been used to extract complex robot skills. In contrast, we use low-cost tactile sensors to capture rich multi-modal data and enable imitation learning with a small number of robot demonstrations. Furthermore, we are the first to demonstrate the benefit of incorporating tactile observations in human demonstrations for improving the robustness of learned robot manipulation skills.

Overall, this work explores multi-modal human demonstrations as a foundation for scalable and efficient pre-training, focusing on improving performance in complex, contact-rich tasks for future imitation learning scenarios.

III. METHODS

In this section, we present our framework for imitation learning, which leverages pre-training with multi-modal human demonstrations to improve the efficiency and effectiveness of policy learning with limited robot demonstrations. Our approach aims to address the challenges associated with learning complex manipulation skills by utilizing a two-stage process: (1) pre-training, where a multi-modal encoder is trained to learn latent state representations from human demonstrations providing multi-modal sensing data but no action labels, and (2) fine-tuning, where a diffusion-based policy is trained using a smaller set of robot demonstrations.

During pre-training, we extract a structured latent state-space by jointly learning forward and inverse dynamics models in the latent space. This allows us to capture task-relevant features from multi-modal human demonstrations without requiring explicit action labels, making the process more scalable and data-efficient. To enhance the quality and robustness of the learned representations, we employ knowledge self-distillation with a teacher-student setup and utilize DynaMo loss to ensure meaningful latent representations.

In the fine-tuning phase, we train a Diffusion Policy conditioned on a history of learned latent state representations alongside a conditioning variable such as a goal or task label to generate actions. By leveraging the pre-trained encoder, our method enables efficient imitation learning, reducing the number of robot demonstrations required to achieve high task success rates.

Our framework effectively combines visual and tactile modalities through our ViTacT multi-modal encoder, which processes multiple camera views and tactile data to produce rich latent representations. In this study, we use convolutional architectures due to their efficiency on smaller-scale datasets, but we anticipate that transformer-based approaches will improve scalability for future applications on larger datasets.

The following subsections provide a detailed explanation of the pre-training and fine-tuning stages, describing the model architectures, loss functions, and training procedures employed in our approach.

A. Problem Definition

Let $\mathcal{D}_h$ be the dataset of observation-only multi-modal human demonstrations consisting of observations $(\mathbf{o}_0^h, \ldots, \mathbf{o}_N^h)$, and let $\mathcal{D}_r$ be the data set of action-labeled demonstrations $(\mathbf{o}_0^r, \mathbf{a}_0^r \ldots, \mathbf{a}_{N-1}^r, \mathbf{o}_N^h)$ obtained by teleoperation. Note that the human demonstrations are action free. Given the ease of obtaining human demonstrations and the difficulty of robot demonstrations, we assume $\mathcal{D}_h >> \mathcal{D}_r$.

Our objective is to efficiently learn a policy for skills demonstrated in $\mathcal{D}_r$ by leveraging pre-training on human demonstrations in $\mathcal{D}_h$. We define the policy as $\pi(\mathbf{a}_t, \ldots, \mathbf{a}_{t+K} | \mathbf{o}_t, \ldots, \mathbf{o}_{t-H}, u)$, where $\mathbf{o}_t, \ldots, \mathbf{o}_{t-H}$ represent a history of observations, and u is a conditioning variable such as a goal, task label, or language instruction embedding. For simplicity, we omit the embodiment superscripts h and r throughout the discussion.

In the pre-training stage, we seek to pre-train multi-modal encoder ϕ by learning dynamics in the latent space. Particularly, a latent inverse dynamics model $h(\mathbf{z}_t | \mathbf{x}_t, \mathbf{x}_{t+F})$ and a forward dynamics model $f(\mathbf{x}_t | \mathbf{x}_t, \mathbf{z}_t)$ where $\mathbf{x}_t$ is the latent state of observation $\mathbf{o}_t$ and $\mathbf{z}_t$ is the latent action between the observation pair $\mathbf{x}_t$ and $\mathbf{x}_{t+F}$. Later in the imitation learning stage, we seek to learn a diffusion policy $\psi(\mathbf{a}_t, \ldots, \mathbf{a}_{t+K} | \mathbf{x}_t, \ldots, \mathbf{x}_{t-H}, u)$ that generates the actions given a history of observations.

B. Pre-training with Multi-modal Human Demonstrations

In the pre-training phase, we aim to model dynamics in the latent space to avoid a computationally expensive reconstruction. Let $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+F}$ denote the latent representations of observations $\boldsymbol{o}_t$ and $\boldsymbol{o}_{t+F}$, respectively, obtained through a multi-modal encoder ϕ.

Specifically, we learn a forward dynamics model $f(\hat{\boldsymbol{x}}_{t+F} | \boldsymbol{x}_t, \boldsymbol{z}_t)$ while simultaneously, we train an inverse dynamics model $h(\boldsymbol{z}_t | \boldsymbol{x}_t, \boldsymbol{x}_{t+F})$. The training objective minimizes the prediction error in the latent space:

$$d(\boldsymbol{x}_{t+F}, \hat{\boldsymbol{x}}_{t+F}),$$

where d is a similarity measure and and $F = 6$. The low-dimensional latent action $\boldsymbol{z}_t$, introduced by the inverse dynamics model, serves as an information bottleneck, which is essential for efficient representation learning [27]. However, naively minimizing the latent state prediction loss,

$$-\mathbb{E}[d(\boldsymbol{x}_{t+F}, \hat{\boldsymbol{x}}_{t+F})],$$

can lead to latent collapse.

To address this, we employ knowledge self-distillation, learning two identical encoders: a teacher network, $\phi_{teacher}$, and a student network, $\phi_{student}$. This approach has been shown to improve representation quality while reducing memory requirements and training time, especially with high-dimensional observations. The latent state $\boldsymbol{x}_{t+F}$ is computed using the teacher network:

$$\boldsymbol{x}_{t+F} = \phi_{teacher}(\boldsymbol{o}_{t+F}).$$

The predicted latent state $\hat{\boldsymbol{x}}_{t+F}$ is computed as follows:
1) Compute the student embedding: $\phi_{student}(\boldsymbol{o}_t)$.
2) Predict the forward dynamics:

$$\hat{\boldsymbol{x}}_{t+F} \sim f(\hat{\boldsymbol{x}}_{t+F}|\boldsymbol{x}_t, \boldsymbol{z}_t),$$

where $\boldsymbol{z}_t$ is sampled from the inverse dynamics model:

$$\boldsymbol{z}_t \sim h(\boldsymbol{z}_t|\boldsymbol{x}_t, \boldsymbol{x}_{t+F}).$$

To ensure stability and effective learning, we minimize the latent state prediction loss alongside regularization components:

$$-\mathbb{E}[d(\text{sg}[\boldsymbol{x}_{t+F}], \hat{\boldsymbol{x}}_{t+F})] + \beta d_{KL}(\mathbf{z}_t||\mathcal{N}(0,1))$$

where sg indicates a stop-gradient operation to prevent teacher updates from the gradient of this loss and d_{KL} is the KL-divergence. Finally, the teacher encoder $\phi_{teacher}$ is updated as an Exponential Moving Average (EMA) of the student encoder $\phi_{student}$. This enables knowledge self-distillation for learning state-representations.

While various similarity loss can be used, we use Dynamo Loss as it has been successfully used for dynamics driven representation learning. We therefore minimize

$$\mathcal{L}_{Dynamo} + \beta d_{KL}(\mathbf{z}_t||\mathcal{N}(0,1)).$$

Dynamo Loss comprises two components: cosine similarity and covariance regularization terms. The cosine similarity is defined as

$$\mathcal{L}_{\text{cosine}}(\hat{\mathbf{x}}_{t+F}, \mathbf{x}_{t+F}) = 1 - \frac{\langle \hat{\mathbf{x}}_{t+F}, \text{sg}[\mathbf{x}_{t+F}]\rangle}{\|\hat{\mathbf{x}}_{t+F}\|_2 \cdot \|\text{sg}[\mathbf{x}_{t+F}]\|_2}.$$

This loss penalizes when the two vectors deviate from each other. The second loss component is the covariance regularization loss defined as

$$\mathcal{L}_{\text{cov}}(\hat{\mathbf{X}}_{t+F}) = \frac{1}{d}\sum_{i\neq j}\left[\text{Cov}(\hat{\mathbf{X}}_{t+F})\right]^2_{i,j},$$

where $\text{Cov}(\hat{\mathbf{X}}_{t+F})$ is the covariance matrix of predicted latent states $\hat{\mathbf{X}}_{t+F}$, and d is the dimensionality of the feature representations. This term minimizes the off-diagonal elements of the covariance matrix, thereby encouraging feature decorrelation and reducing redundancy among the learned features. Therefore, the total DynaMo loss is the weighted sum of the two components

$$\mathcal{L}_{Dynamo} = \mathcal{L}_{\text{cosine}} + \lambda\mathcal{L}_{\text{cov}},$$

where λ is a hyperparamter to control the tradeoff between the two components. The default value of λ is 0.04.

In our method, ϕ is a ViTacT multi-modal encoder, which processes multiple camera views and tactile modalities to generate latent representations. Each camera view is encoded into a token using a ResNet-based embedding, while tactile signals are processed through 1D convolution to produce tactile embeddings. These embeddings are then used as inputs to a transformer decoder, enabling the fusion of multi-modal information for downstream policy learning. However, in large-scale applications with more extensive training data, we anticipate that patch-based encoding, as implemented in Vision Transformers (ViTs), will offer superior performance due to the favorable scaling properties of transformer architectures compared to convolutional networks. In our approach, we utilize CNN-based feature extraction, as our work focuses on a relatively smaller scale of data, where convolutional architectures remain effective and computationally efficient.

C. Imitation Learning with Robot Demonstrations

In the second imitation learning phase, we train a diffusion policy conditioned on the latent state representations extracted by the pre-trained multi-modal encoder ϕ, facilitating efficient learning from a limited number of demonstrations. Specifically, the diffusion policy $\psi(\boldsymbol{a}_t, \ldots, \boldsymbol{a}_{t+K} \mid \boldsymbol{x}_t, \ldots, \boldsymbol{x}_{t-H}, u)$ generates actions based on a history of latent states and a conditioning variable u, such as a goal, task label, or language instruction. We use a history $H = 16$ and action chunk length $K = 16$ in our experiments.

By leveraging the pre-trained encoder ϕ, our approach effectively utilizes human demonstration data to provide informative latent representations, improving the efficiency of imitation learning. As shown in our results, this leads to faster convergence and better generalization with fewer robot demonstrations.

IV. EXPERIMENTAL RESULTS

In this section, we evaluate the effectiveness of our proposed pre-training method using human demonstrations, focusing on its data efficiency and robustness. Additionally, we assess the advantages of incorporating instrumented tactile feedback and its impact on policy performance and adaptability in complex manipulation tasks. Our results show that pre-training with human demonstrations improves the efficiency of downstream imitation learning, achieving better performance with fewer robot demonstrations and increased robustness across considered tasks.

A. Task and Demonstrations

We consider the task of grasping an arbitrarily oriented object on a table using a gripper equipped with tactile sensing. To achieve this, we collect demonstrations from both humans and the robot, as shown in Fig 3. The human dataset includes five objects, with 1,000 human demonstrations per object. Since robot demonstrations require teleoperation, which is more challenging and time-consuming, we collect them for only 100 demos on a single object. The objects used for pre-training with human demos, fine-tuning, and generalization are all shown in Fig 4.
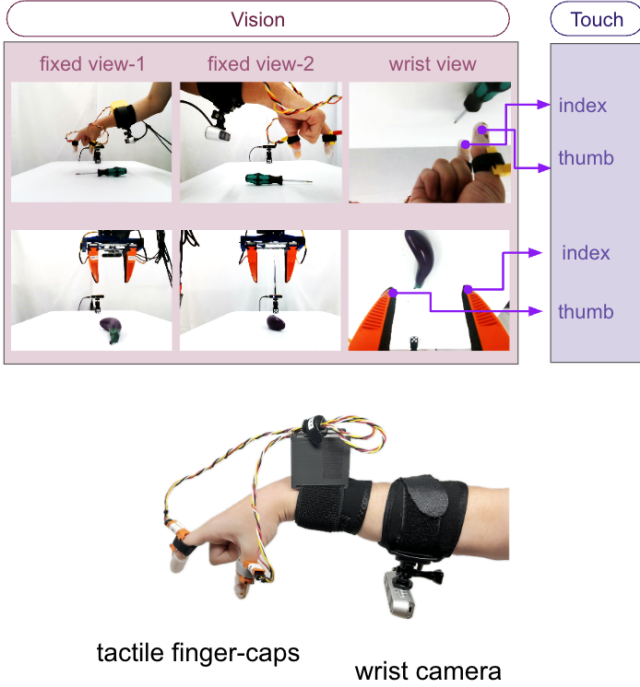
Fig. 3: Human and Robot Data Collection Setup. The robot setup includes three camera views—two side views and one wrist view—and a two-fingered gripper with embedded tactile sensors at the fingertips. The human data collection setup mirrors the robot's camera configuration, with tactile data collected using a fingertip cap device equipped with a Singletact capacitive sensor on index and the thumb fingers.
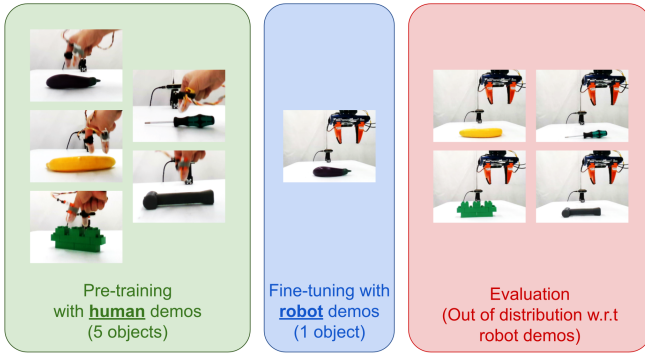

Fig. 4: Objects used for pre-training, fine-tuning, and generalization evaluation. As illustrated, we pre-train using human demonstrations collected for all five objects but fine-tune the imitation learning model using only one object. To evaluate generalization, we test on the remaining four objects, making the evaluation out-of-distribution relative to the robot fine-tuning phase but in-distribution with respect to the human demonstrations used during pre-training.

B. Task Performance and Generalization with Few Robot Demonstrations

In this section, we evaluate the improvement in demonstration complexity achieved through pre-training with our method. We compare our approach against baseline methods that use different pre-trained encoders, including alternative strategies for pre-training with human demonstrations. For these experiments, tactile input is not used, allowing us to focus solely on the impact of pre-training on performance. We evaluate the following methods:

1) Diffusion Policy (DP): This baseline is a diffusion policy using a ResNet-18 encoder to encode each view in history of observations. The ResNet is initialized with ImageNet pre-trained weights, serving as a baseline without task-specific pre-training. Note that weights from ImageNet pre-training are used exclusively for the this baseline and not in any other baselines. This baseline does not use any human demonstrations.

2) Diffusion Policy with ViTacT Encoder Pretrained on Human Demos Using Time-Contrastive Loss (DP + TCL): In this baseline, we consider the Time Constrastive Loss as proposed by R3M [38] baseline. Temporally relevant features are important for manipulation skills. Therefore, we consider a common temporal contrastive learning technique for learning state representations. We train our ViTacT encoder using our human demonstration dataset with a temporal modeling approach based on Time Contrastive Loss. Latents for temporally closer images from a video demonstration have higher similarity compared to images that are temporally distant or come from different videos using contrastive learning. Specifically, a batch of frame sequences is sampled, consisting of frames $\mathbf{o}_i$, $\mathbf{o}_j$ (where $j > i$), and $\mathbf{o}_k$ (where $k > j$) is sampled and the InfoNCE loss is minimized.

$$\mathcal{L}_{\text{TCL}} = -\sum_{b \in \mathcal{B}} \log \frac{e^{d(x_i^b, x_j^b)}}{e^{d(x_i^b, x_j^b)} + e^{d(x_i^b, x_k^b)} + e^{d(x_i^b, x_i^{\neq b})}},$$

where x is the latent state and $x_i^{\neq b}$ is a latent state from a $\mathbf{o}_i^{\neq b}$ - observation from a different demo in the batch. d denotes a measure of similarity, where we specify as the negative l_2 distance.

3) Diffusion Policy with JIF Pretraining (DP + JIF): this is our method.

The first metric we consider for all methods above is task success as a function of the number of robot demonstrations used in training. Higher success rate with fewer robot demonstrations implies that a method is better able to extract value from human pre-training. We evaluate all methods on a number of robot demonstrations varying between 1 and 200. The success rate of the resulting imitation policies for each method is measured over 15 trials. A detailed performance comparison is presented in Fig. 5.

We note that our method (DP + JIF) achieves more than twice the success rate compared to the DP baseline without human pre-training. Additionally, DP + JIF demonstrates superior sample efficiency, achieving a higher success rate compared to the DP + TCL baseline when using a similar
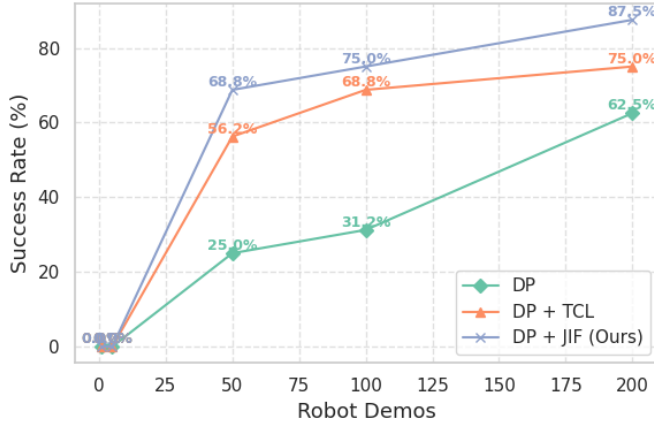
Fig. 5: Grasping success rate of our method against the baselines. While both pre-training approaches improve demonstration complexity, with JIF pre-training we achieve the highest success rate.

TABLE I: Generalization Performance. We compare the generalization performance of our method DP + JIF fine-tuned with 50 robot demonstrations on grasping objects out-of-distribution w.r.t fine-tuning robot demonstrations.

Object	Eggplant	Banana	Lego	Screwdriver	Hammer
Method	In-distribution	Out-of-distribution →			
DP	25.0%	37.5%	43.8%	12.5%	31.3%
DP + TCL	56.3%	62.5%	62.5%	50.0%	**56.3%**
DP + JIF (Ours)	**68.8%**	**68.8%**	**68.8%**	**50.0%**	43.8%

number of robot demonstrations. These results underscore the benefits of incorporating human pre-training via JIF in enhancing both success rates and data efficiency.

We then assess the generalization performance of our approach. Specifically, we evaluate the success rate on objects that were not included in the robot demonstrations, as detailed in Table I. For each training method, we use the model trained on 50 demonstrations.

Our data shows that the model can generalize even without human data pre-training. However, the TCL and JIF models, both pre-trained with human data, achieve higher success rates, with our JIF model outperforming the others on most objects.

C. Robustness with Visuo-tactile Human Demonstrations

To evaluate the impact of instrumented human demonstrations, we incorporate tactile readings as additional inputs for state representation learning and evaluate with a peg-in-insertion task. As before we collect a 1,000 human demonstrations with 100 robot demonstrations, as shown in Fig 6. We compare the following tactile sensing conditions:

1) No Tactile Sensing: Both human demonstrations and the robot policy rely solely on visual data without incorporating tactile inputs.

2) Human and Robot Tactile Sensing: Both human demonstrations and the robot policy incorporate tactile sensing alongside visual data, providing a richer representation of the manipulation process.
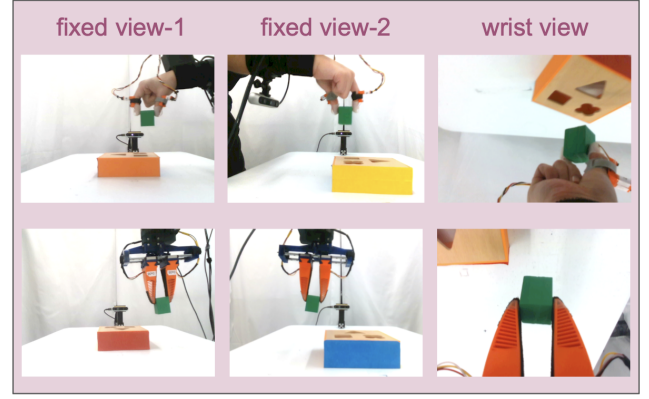

Fig. 6: The importance of instrumented human demonstrations with tactile sensing is evaluated using a challenging peg-in-hole insertion task. The figure illustrates both human and robot demonstrations during the placement of a cube into a square hole.
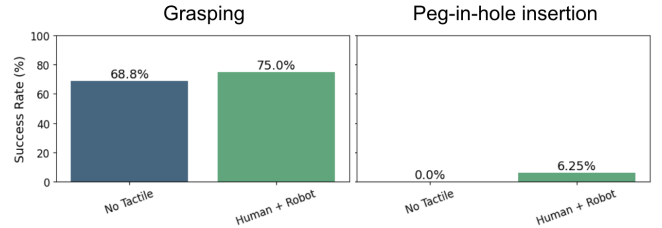

Fig. 7: The results underscore the importance of instrumented human demonstrations with tactile sensing, as evidenced by a success rate comparison across different tactile sensing configurations. The left plot shows notable improvements in grasping success rates, while the right plot highlights the successful execution of the challenging peg-in-hole insertion task, albeit with a lower success rate. Incorporating tactile feedback, particularly in both human and robot demonstrations, enhances performance and achieves higher success rates with fewer robot demonstrations.

The results of this comparison are summarized in Fig. 7. As shown, pre-training with tactile sensing leads to a notable improvement in success rates for grasping, demonstrating the value of incorporating tactile feedback in both human and robot demonstrations. More importantly, this approach enables the successful execution of the challenging task of peg-in-hole insertion, albeit with a low success rate, emphasizing the role of instrumented human demonstrations in learning complex, contact-rich tasks.

D. Visualizing Learned Latent States

We applied Uniform Manifold Approximation and Projection (UMAP) [53] to visualize the embedding space. We utilized two demonstrations per object to assess whether distinct object categories naturally emerge within the latent space. The UMAP visualization (Fig 8) revealed clear separability among the five different objects, indicating that the learned representations effectively capture object-specific features. Furthermore, the latent space exhibited meaningful structure with respect to task phases, as it distinctly separated the pre-grasp, grasp, and
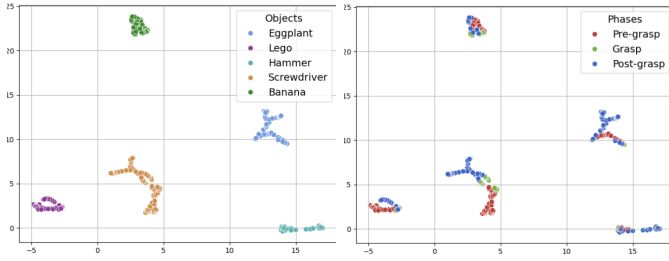
Fig. 8: UMAP visualization of the learned representations. Left: The representations exhibit clear separability among the five different objects, indicating that the model effectively captures object-specific features. Right: The latent space shows meaningful structure with respect to task phases, distinctly separating the pre-grasp, grasp, and post-grasp phases. This suggests that the learned representations capture manipulation-specific information.

post-grasp phases.

E. Limitations

While this study provides valuable insights into leveraging human demonstrations for pre-training in imitation learning, several limitations should be considered. First, the study's scale is limited to a reduced number of tasks. Furthermore, despite variations in inertial properties, the target objects used for grasping share similar aspect ratios, potentially limiting the diversity of learned representations. Additionally, the focus on grasping—a relatively simple task—may not fully reflect the complexities of real-world manipulation scenarios.

Methodologically, the reliance on human demonstrations does present its own challenges in terms scalability due to the need for instrumentation (cameras, wearable tactile sensors) which could limit its applicability "in the wild". While one possible direction for the field aims to completely eliminate the need for any robot demonstrations, our approach still depends on robot demonstrations for fine-tuning, which can be impractical in scenarios where robotic data collection is infeasible or costly. Nevertheless, we believe that the ability to extract value from human demonstrations, complementing a much smaller number of robot demonstrations, can offer a significant advantage towards scalability in deployment.

V. DISCUSSION AND FUTURE WORK

In this work, we introduced a novel pre-training framework that leverages multi-modal human demonstrations to acquire manipulation-centric representations for imitation learning. Our approach provides several key advantages, including improved fine-tuning efficiency, enhanced generalization, and greater computational efficiency by avoiding costly action annotations and high-dimensional reconstructions. These findings position human demonstration-based pre-training as a promising solution to critical challenges in imitation learning, particularly in addressing the pervasive issue of data scarcity.

A key contribution of our work is the use of instrumented multi-modal human demonstrations, particularly with tactile sensing, which significantly improves performance. This highlights the potential of richer sensory inputs, such as touch, in

advancing robotic manipulation capabilities and closing the gap between human and robot skill acquisition.

We aim to extend our framework by incorporating latent actions alongside latent states during fine-tuning to further enhance policy learning efficiency. Additionally, we plan to explore the integration of instrumented human demonstration pre-training within offline reinforcement learning to better leverage human priors for data-efficient policy optimization.

We believe that multi-modal human demonstrations, especially those enriched with tactile and proprioceptive feedback, will continue to play a crucial role in developing robust robotic systems. As many valuable representations of the physical world are inherently grounded in sensory interaction, learning through touch and other modalities will be pivotal for advancing robotic dexterity. Future research focus on scaling our approach to diverse, real-world tasks and addressing the challenges posed by complex, high-dimensional manipulation scenarios seems promising.

REFERENCES

[1] Homer Walke et al. "BridgeData V2: A Dataset for Robot Learning at Scale". In: *CoRL* 229 (Aug. 2023). Ed. by Jie Tan, Marc Toussaint, and Kourosh Darvish, pp. 1723–1736.

[2] Alexander Khazatsky et al. "DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset". In: *arXiv [cs.RO]* (Mar. 2024).

[3] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. "Diffusion Policy: Visuomotor policy learning via action diffusion". In: *arXiv [cs.RO]* (Mar. 2023).

[4] Huy Ha, Peter R Florence, and Shuran Song. "Scaling up and distilling down: Language-guided robot skill acquisition". In: *CoRL* abs/2307.14535 (July 2023).

[5] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. "Mobile ALOHA: Learning bimanual mobile manipulation with low-cost whole-body teleoperation". In: *arXiv [cs.RO]* (Jan. 2024).

[6] Aviral Kumar, Anikait Singh, Frederik Ebert, Mitsuhiko Nakamoto, Yanlai Yang, Chelsea Finn, and Sergey Levine. "Pre-training for robots: Offline RL enables learning new tasks from a handful of trials". In: *arXiv [cs.RO]* (Oct. 2022).

[7] Aviral Kumar, Rishabh Agarwal, Xinyang Geng, George Tucker, and Sergey Levine. "Offline Q-learning on diverse multi-task data both scales and generalizes". In: *arXiv [cs.LG]* (Nov. 2022).

[8] Kevin Black et al. "π_0: A vision-language-action flow model for general robot control". In: *arXiv [cs.LG]* (Oct. 2024).

[9] Open X-Embodiment Collaboration et al. "Open X-Embodiment: Robotic Learning Datasets and RT-X Models". In: *arXiv [cs.RO]* (Oct. 2023).

[10] Moo Jin Kim et al. "OpenVLA: An Open-Source Vision-Language-Action Model". In: *arXiv [cs.RO]* (June 2024).

[11] Octo Model Team et al. "Octo: An Open-Source Generalist Robot Policy". In: *arXiv [cs.RO]* (May 2024).

[12] Anthony Brohan et al. "RT-2: Vision-language-action models transfer web knowledge to robotic control". In: *arXiv [cs.RO]* (July 2023).

[13] Jinghuan Shang, Karl Schmeckpeper, Brandon B May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. "Theia: Distilling diverse vision foundation models for robot learning". In: *arXiv [cs.RO]* (July 2024).

[14] Zhao Mandi, Homanga Bharadhwaj, Vincent Moens, Shuran Song, Aravind Rajeswaran, and Vikash Kumar. "CACTI: A framework for scalable multi-task multi-scene visual imitation learning". In: *arXiv [cs.RO]* (Dec. 2022).

[15] Yiyue Luo et al. "Learning human–environment interactions using conformal tactile textiles". en. In: *Nat. Electron.* 4.3 (Mar. 2021), pp. 193–201.

[16] Sudeep Dasari, Mohan Kumar Srirama, Unnat Jain, and Abhinav Gupta. "An Unbiased Look at Datasets for Visuo-Motor Pre-Training". en. In: *Conference on Robot Learning*. PMLR, Dec. 2023, pp. 1183–1198.

[17] Jyothish Pari, Nur Muhammad Shafiullah, Sridhar Pandian Arunachalam, and Lerrel Pinto. "The surprising effectiveness of representation learning for visual imitation". In: *arXiv [cs.RO]* (Dec. 2021).

[18] Arjun Majumdar et al. "Where are we in the search for an Artificial Visual Cortex for Embodied Intelligence?" In: *Neural Inf Process Syst* abs/2303.18240 (Mar. 2023). Ed. by A Oh, T Naumann, A Globerson, K Saenko, M Hardt, and S Levine, pp. 655–677.

[19] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. "Language-driven representation learning for robotics". In: *arXiv [cs.RO]* (Feb. 2023).

[20] Fangchen Liu, Hao Liu, Aditya Grover, and Pieter Abbeel. "Masked Autoencoding for Scalable and Generalizable Decision Making". In: *arXiv [cs.LG]* (Nov. 2022).

[21] Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. "Scaling proprioceptive-visual learning with Heterogeneous Pre-trained Transformers". In: *arXiv [cs.RO]* (Sept. 2024).

[22] Shizhe Chen, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. "SUGAR: Pre-training 3D Visual Representations for Robotics". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 18049–18060.

[23] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. "Simple but Effective: CLIP Embeddings for Embodied AI". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 14829–14838.

[24] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. "Real-World Robot Learning with Masked Visual Pre-training". en. In: *Conference on Robot Learning*. PMLR, Mar. 2023, pp. 416–426.

[25] Ilija Radosavovic, Baifeng Shi, Letian Fu, Ken Goldberg, Trevor Darrell, and Jitendra Malik. "Robot Learning with Sensorimotor Pre-training". en. In: *Conference on Robot Learning*. PMLR, Dec. 2023, pp. 683–693.

[26] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. "The Unsurprising Effectiveness of Pre-Trained Vision Models for Control". In: *arXiv [cs.CV]* (Mar. 2022).

[27] Dominik Schmidt and Minqi Jiang. "Learning to Act without Actions". In: (Oct. 2023).

[28] Zichen Jeff Cui, Hengkai Pan, Aadhithya Iyer, Siddhant Haldar, and Lerrel Pinto. "DynaMo: In-domain dynamics pretraining for visuo-motor control". In: *arXiv [cs.RO]* (Sept. 2024).

[29] Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. "Concept2Robot: Learning manipulation concepts from instructions and human demonstrations". en. In: *Int. J. Rob. Res.* 40.12-14 (Dec. 2021), pp. 1419–1434.

[30] Annie S Chen, Suraj Nair, and Chelsea Finn. "Learning Generalizable Robotic Reward Functions from "In-The-Wild" Human Videos". In: *arXiv [cs.RO]* (Mar. 2021).

[31] Alejandro Escontrela et al. "Video Prediction Models as Rewards for Reinforcement Learning". In: *arXiv [cs.LG]* (May 2023).

[32] Karl Pertsch, Ruta Desai, Vikash Kumar, Franziska Meier, Joseph J Lim, Dhruv Batra, and Akshara Rai. "Cross-Domain Transfer via Semantic Skill Imitation". In: *arXiv [cs.LG]* (Dec. 2022).

[33] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. "XSkill: Cross Embodiment Skill Discovery". In: *arXiv [cs.RO]* (July 2023).

[34] Kevin Zakka, Andy Zeng, Pete Florence, Jonathan Tompson, Jeannette Bohg, and Debidatta Dwibedi. "XIRL: Cross-embodiment Inverse Reinforcement Learning". In: *arXiv [cs.RO]* (June 2021).

[35] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. "XSkill: Cross embodiment skill discovery". In: *arXiv [cs.RO]* (July 2023).

[36] Raghav Goyal et al. "The "something something" video database for learning and evaluating visual common sense". In: *arXiv [cs.CV]* (June 2017).

[37] Kristen Grauman et al. "Ego4D: Around the World in 3,000 Hours of Egocentric Video". In: *arXiv [cs.CV]* (Oct. 2021).

[38] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. "R3M: A universal visual representation for robot manipulation". In: *arXiv [cs.RO]* (Mar. 2022).

[39] Jean-Bastien Grill et al. "Bootstrap Your Own Latent: A new approach to self-supervised learning". In: *Adv. Neural Inf. Process. Syst.* abs/2006.07733 (June 2020).

[40] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand

Joulin. "Emerging Properties in Self-Supervised Vision Transformers". In: *arXiv [cs.CV]* (Apr. 2021).

[41] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. "Time-Contrastive Networks: Self-Supervised Learning from Video". In: *arXiv [cs.CV]* (Apr. 2017).

[42] Ashley D Edwards, Himanshu Sahni, Yannick Schroecker, and Charles L Isbell. "Imitating Latent Policies from Observation". In.

[43] Philipp Ruppel and Jianwei Zhang. "Elastic tactile sensor glove for dexterous teaching by demonstration". en. In: *Sensors (Basel)* 24.6 (Mar. 2024), p. 1912.

[44] John C S McCaw, Michelle C Yuen, and Rebecca Kramer-Bottiglio. "Sensory glove for dynamic hand proprioception and tactile sensing". en. In: *Volume 2B: 44th Design Automation Conference*. American Society of Mechanical Engineers, Aug. 2018, V02BT03A025.

[45] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. "Learning the signatures of the human grasp using a scalable tactile glove". en. In: *Nature* 569.7758 (May 2019), pp. 698–702.

[46] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. "Learning the signatures of the human grasp using a scalable tactile glove". en. In: *Nature* 569.7758 (May 2019), pp. 698–702.

[47] Dehao Wei and Huazhe Xu. "A wearable RObotic Hand for hand-over-hand imitation learning". In: *arXiv [cs.RO]* (Sept. 2023).

[48] Takashi Sagisaka, Yoshiyuki Ohmura, Akihiko Nagakubo, Kazuyuki Ozaki, and Yasuo Kuniyoshi. "Development and applications of high-density tactile sensing glove". en. In: *Haptics: Perception, Devices, Mobility, and Communication*. Lecture notes in computer science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 445–456.

[49] T B Martin, R O Ambrose, M A Diftler, R Platt, and M J Butzer. "Tactile gloves for autonomous grasping with the NASA/DARPA Robonaut". en. In: *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*. Vol. 2. IEEE, 2004, 1713–1718 Vol.2.

[50] Takashi Sagisaka, Yoshiyuki Ohmura, Yasuo Kuniyoshi, Akihiko Nagakubo, and Kazuyuki Ozaki. "High-density conformable tactile sensing glove". en. In: *2011 11th IEEE-RAS International Conference on Humanoid Robots*. IEEE, Oct. 2011, pp. 537–542.

[51] Joo Chuan Yeo, Cassidy Lee, Zhiping Wang, and Chwee Teck Lim. "Tactile sensorized glove for force and motion sensing". en. In: *2016 IEEE SENSORS*. IEEE, Oct. 2016, pp. 1–3.

[52] Yeongmi Kim, Jongeun Cha, Jeha Ryu, and Ian Oakley. "A tactile glove design and authoring system for immersive multimedia". en. In: *IEEE Multimed.* 17.3 (2010), pp. 34–45.

[53] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv [stat.ML]* (Feb. 2018).