Robotics: Science and Systems 2024
Delft, Netherlands, July 15-July 19, 2024

Learning Manipulation by Predicting Interaction

Jia Zeng[1,*], Qingwen Bu[2,1,*], Bangjun Wang[2,1,*], Wenke Xia[3,1,*], Li Chen[1], Hao Dong[4], Haoming Song[1], Dong Wang[1,†], Di Hu[3], Ping Luo[1], Heming Cui[1], Bin Zhao[1,5,†], Xuelong Li[1,6], Yu Qiao[1] and Hongyang Li[1,2,†]

[1]Shanghai AI Lab [2]Shanghai Jiao Tong University [3]Renmin University of China
[4]Peking University [5]Northwestern Polytechnical University [6]TeleAI, China Telecom Corp Ltd
[*]Equal contribution. [†]Corresponding authors.

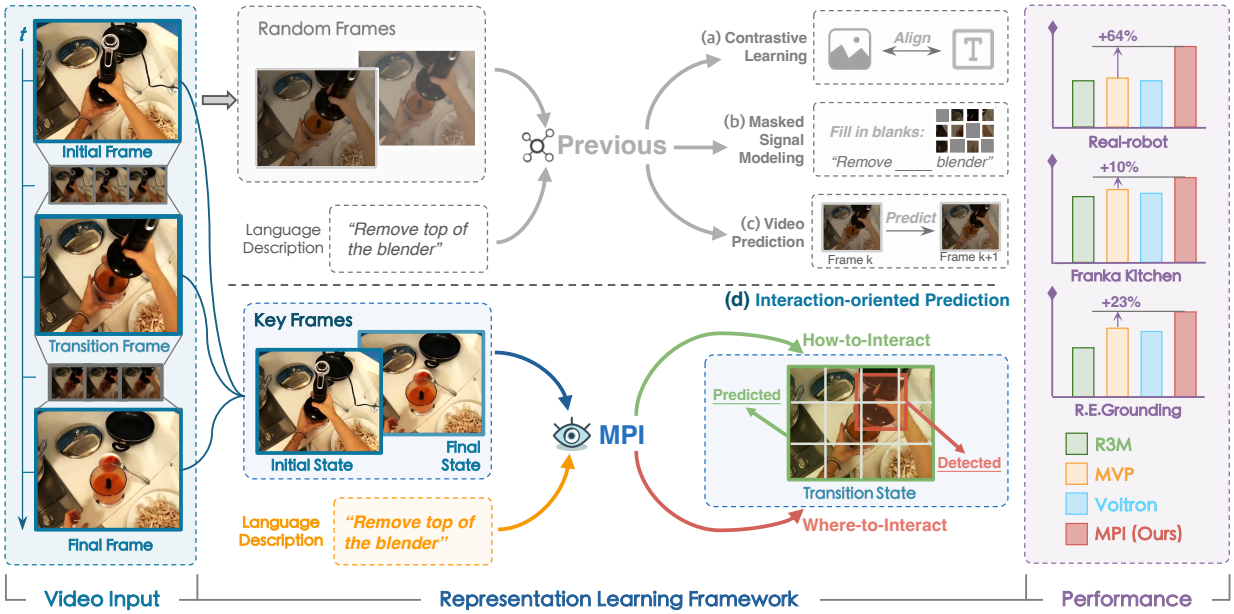

Fig. 1: **MPI** is an interaction-oriented representation learning pipeline for robotic manipulation. Diverging from prior arts grounded in (a) Contrastive Learning, (b) Masked Signal Modeling, or (c) Video Prediction using random frames, our proposed approach in (d) instructs the model towards predicting transition frames and detecting manipulated objects with keyframes as input. As such, the model fosters better comprehension of "how-to-interact" and "where-to-interact". MPI acquires more informative representations during pre-training and achieves evident improvement across downstream tasks.

***Abstract*—Representation learning approaches for robotic manipulation have boomed in recent years. Due to the scarcity of in-domain robot data, prevailing methodologies tend to leverage large-scale human video datasets to extract generalizable features for visuomotor policy learning. Despite the progress achieved, prior endeavors disregard the interactive dynamics that capture behavior patterns and physical interaction during the manipulation process, resulting in an inadequate understanding of the relationship between objects and the environment. To this end, we propose a general pre-training pipeline that learns Manipulation by Predicting the Interaction (MPI) and enhances the visual representation. Given a pair of keyframes representing the initial and final states, along with language instructions, our algorithm predicts the transition frame and detects the interaction object, respectively. These two learning objectives achieve superior comprehension towards "how-to-interact" and "where-to-interact". We conduct a comprehensive evaluation of several challenging robotic tasks. The experimental results demonstrate that MPI exhibits remarkable improvement by 10% to 64% compared with previous state-of-the-art in real-world robot platforms as well as simulation environments. Code and checkpoints are publicly shared at https://github.com/OpenDriveLab/MPI.**

I. Introduction

Visuomotor control of robotic systems entails perceiving and interpreting the surrounding environment from visual inputs, making informed decisions, and executing appropriate actions. This capability is of great significance to a wide range of robotic applications, encompassing object manipulation [4], grasping [60], navigation [50, 51, 55], *etc*. Motivated by the achievements of large-scale pre-training in fundamental vision [26, 27, 43] and natural language processing [14, 47], the field of robotics seeks to utilize large-scale data to build generalizable representations. However, for robot manipulation, collecting demonstrations is both laborious and costly. Consequently, the exploration of representation learning methodologies that circumvent dependence on limited in-domain robotics data has emerged as a prominent and trending research focus.

Notably, recent efforts have harnessed large-scale egocentric human video datasets, such as Ego4D [21], Something-Something V2 [20], and Epic Kitchens [10], to bootstrap

representation learning for robot manipulation. As illustrated in Fig. 1(a)-(b), previous approaches extensively employ contrastive learning [36, 40] and masked signal modeling [28, 40, 44]. While they offer valuable insights for enhancing the performance of robotic systems in downstream tasks, their primary focus tends to be either discerning high-level semantic cues or capturing fine-grained pixel information. This focus results in the overlook of crucial interactive dynamics [3, 48], which refers to the behavior patterns and physical interactions occurring between a robot and the environment.

In most contexts where robotic systems are deployed, their functions extend beyond passive perceptual capabilities [45] to active engagement with the environment [4, 64]. This prompts an exploration of how to connect interactions that shape the world and visual representations that speak the world. One viable approach in this regard is video prediction pre-training [54], as depicted in Fig. 1(c). Through learning to predict future frames, the model inherently acquires the ability to represent the temporal evolution of scenes [2]. However, in the specific context of robot manipulation scenes, objects typically do not exhibit autonomous movement, and scene changes primarily arise from interaction-driven movements. Conventional works [18, 23] for predicting future consecutive frames merely model the temporal relationships between frames. This task setting is less challenging and prone to introducing noise or redundant information, thus hindering the model's ability to discern interaction-relevant patterns or capture the dynamic interactions in manipulation scenarios effectively [16]. Therefore, it is imperative to propose an enhanced interaction-oriented video prediction method.

To address the aforementioned challenges, we propose **MPI**, which stands for learning **M**anipulation by **P**redicting the **I**nteraction. As shown in Fig. 1(d), we formulate two primary objectives, which are also key elements that formulate an interaction, summarized as "how-to-interact" and "where-to-interact". Correspondingly, we propose two modules: Prediction Transformer and Detection Transformer. Given a pair of keyframes representing the initial and final states of an interaction process, along with language instruction, the Prediction Transformer predicts the unseen transition frame that represents the interaction between these two states. By employing this objective, the model comprehends "how-to-interact". Moreover, the Detection Transformer infers the location of the interaction object in the unseen frame, enabling the model to acquire knowledge of "where-to-interact". As illustrated in Fig. 2, within a transformer-based encoder-decoder architecture, we leverage a set of prediction queries to estimate the transition frame and a detection query to infer the position of the interaction object. The knowledge across the transition frame prediction and interaction object detection is woven through cross-attention among queries, fostering mutual support in the training process. As such, we construct a unified prediction-detection pre-training framework tailored for robot manipulation. In contrast to predicting future consecutive frames, our algorithm filters out the interaction-irrelevant information and emphasizes the key states, leading to a concentrated learning process.

To evaluate the effectiveness of MPI, we assemble a set of downstream tasks: 1) policy learning on a real-world Franka Emika Panda robot operating in complex scenarios; 2) policy learning on Franka Kitchen [22] within a complex kitchen environment; 3) manipulation of position-varying objects in Meta-World [59] simulation environment; and 4) a robotics-related recognition task (*i.e.*, referring expression grounding [53]), in which the model locates a specified object based on natural language descriptions. Extensive experimental results demonstrate the superior performance of our method compared to R3M [40], MVP [44], and Voltron [28]. As presented in Fig. 1, in comparison to MVP, MPI exhibits a relative improvement rate of 64% in real-world robot experiments, 10% on Franka Kitchen, and 23% (AP@0.75IoU) for Referring Expression Grounding (R.E.G.) task.

Contributions. Our contributions are three folds: **1)** We propose an interaction-oriented representation learning framework, MPI, towards robot manipulation. By learning interaction, MPI strengthens the model's comprehension of interactive dynamics in manipulation scenes. **2)** We introduce the Prediction and Detection Transformer to predict the transition frame and detect the interaction object. These two learning objectives facilitate the model to foster an understanding of "how-to-interact" and "where-to-interact". Moreover, these two learning objectives can be mutually reinforced. **3)** The experimental results reveal that MPI yields state-of-the-art performance on a broad spectrum of downstream tasks.

II. Related Work

A. Representation Learning for Robotics

Effective visual representation plays a crucial role in facilitating robots' interaction with the environment [6, 9, 24, 56, 58]. Given the limited availability of in-domain robot data, current methods attempt to leverage pre-trained visual models to enhance the generalizability of representations, thereby benefiting downstream manipulation tasks [40, 44]. Previous studies have demonstrated the utility of pre-trained models on general vision datasets such as CLIP [43] and ImageNet [12] for improving performance on robotic tasks [29, 42, 52]. To further enhance these representations, recent efforts [40, 44] have focused on pre-training visual models using large-scale egocentric datasets [20, 21], aiming to bridge the gap between pre-trained representations and downstream robotic scenarios. In terms of pre-training techniques, innovative learning approaches have been adopted to abstract representations for robotics. For instance, TCN [49] and VIP [37] propose time-contrastive learning to understand action sequences over time, while the recent work MVP [44] employs masked image modeling method MAE [27] for fine-grained object recognition. Drawing inspiration from advancements in multi-modal foundation models [33], recent studies such as R3M [40], Voltron [28], and LIV [36] bring language into the pre-training to improve semantic comprehension for robotic tasks.

Recently, there are interesting attempts on the practical effectiveness of representation learning using ego-centric hu-

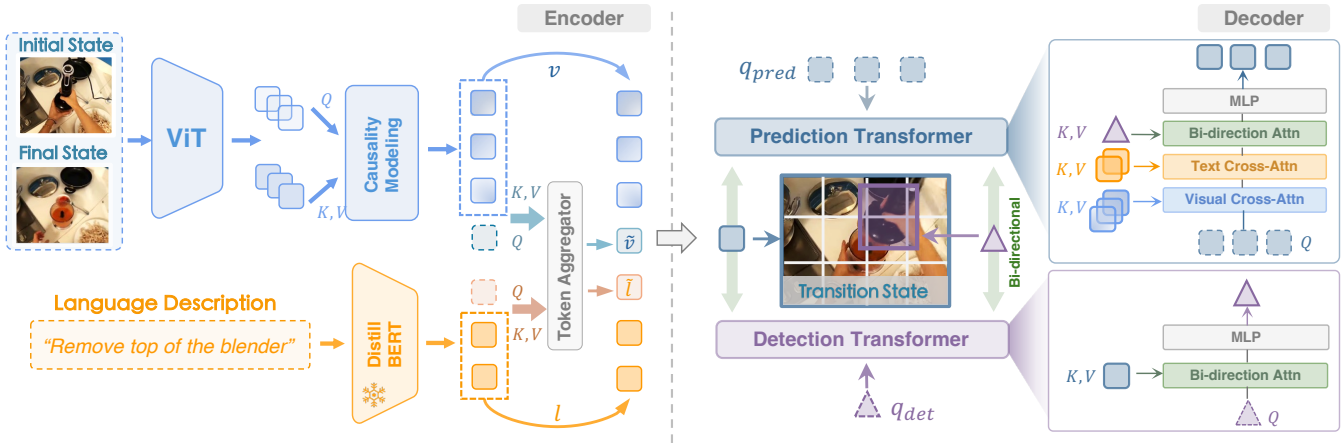

Fig. 2: **The pipeline for pre-training**. MPI comprises a *multi-modal transformer encoder* and a *transformer decoder* designed for predicting the image of the target interaction state and detecting interaction objects respectively. We achieve synergistic modeling and optimization of the two tasks through information transition between the prediction and detection transformers. The decoder is solely engaged during the pre-training phase while deprecated for downstream adaptations.

man data [5, 11]. Burns *et al.* [5] claim that models pre-trained on manipulation-relevant data do not generalize better than those trained on standard pre-training datasets. Based on the MAE pre-train paradigm, Dasari *et al.* [11] demonstrate that representations derived from traditional computer vision datasets such as ImageNet outperform those derived from Ego4D. However, we argue that the experimental phenomena arise because previous literature solely applies existing pre-training techniques to large-scale egocentric datasets, without explicit consideration of the interaction characteristics during manipulation processes. As such, they fail to fully utilize the advantages of modeling temporal relationships and interaction in manipulation-related datasets. In this work, we incorporate interaction-oriented learning to embed representations with manipulation task-specific knowledge, aiming to enhance performance in downstream robotic applications.

B. Learning Interaction in Robotics

Gaining a comprehensive understanding of interaction is essential for effectively manipulating objects. Past research on learning interaction can be broadly classified into two main categories: explicit representation and implicit encoding.

Explicit Representation. Within the field of robotics, affordance [19] refers to the properties of an object that enable specific interactions, *i.e.*, a perceptual cue for object manipulation. Previous works have explicitly expressed interaction information by estimating affordance, which can take various forms, including contact points [32], grasp point pairs [60], or heatmaps [8, 31, 38]. By leveraging internet videos of human behavior, Bahl et al. [1] train a visual affordance model capable of predicting contact point and trajectory waypoints, which explicitly represent the locations and manners in which humans interact within a scene. Building upon such affordances, SWIM [39] constructs affordance-space world models that enable robots to learn manipulation skills. In our proposed pipeline, interaction object detection can be viewed as an explicit representation utilized for supervision.

Implicit Encoding. Given the increasing prevalence of video generation and synthesis, recent research has investigated the potential of applying advanced diffusion techniques to robot manipulation tasks, aiming to model the interaction process implicitly. Approaches such as UniPi [17] leverage internet data and treat the sequential manipulation decision-making problem as a text-conditioned video generation problem. They train a separate inverse dynamics module to estimate actions. Additionally, UniSim [57] aims to develop a universal simulator for simulating the interaction process and training various policies through generative modeling. Instead, our method focuses solely on utilizing the image reconstruction task to guide the representation to incorporate interactive dynamics.

III. MPI: Manipulation by Predicting Interaction

The overall pipeline of our approach is depicted in Fig. 2. We begin with a triplet of keyframes that represent the initial, transition, and final states in an interaction process. Two of these frames are utilized as inputs for the visual encoder, while the remaining frame acts as the prediction target. To establish the state transition relationship between the input frames, we incorporate a causality modeling module that facilitates dynamic attention between the two states. Furthermore, a language description that corresponds to the entire interaction process is employed to provide high-level semantic cues, subsequently conditioning the decoding process for the target state. A shared token aggregator is devised to generate a concise embedding vector by combining the encoded vision and language representations.

In terms of the decoder, we configure the Prediction Transformer and Detection Transformer for interaction frame prediction and interaction object detection. The information exchange between these two modules allows the model to

establish potential relationships between "how-to-interact" and "where-to-interact", thereby optimizing both objectives.

A. Data: A Triplet of Keyframes

Based on the annotations provided in the Ego4D Hand-Object Interaction dataset [21], we define three keyframes denoted as $\{F_{\text{init}}, F_{\text{trans}}, F_{\text{final}}\} \in \mathbb{R}^{3\times H\times W}$. These frames represent the initial, transition, and final states of an interaction process. The initial state serves as the starting point and establishes the contextual foundation for the interaction process. Therefore, F_{init} is consistently used as the input during pre-training. The selection between F_{trans} and F_{final} is subject to a Bernoulli distribution parameterized by $p \in [0, 1]$:

$$F_{\text{input}} = \begin{cases} [F_{\text{init}}, F_{\text{trans}}], & \text{if } \alpha = 0, \\ [F_{\text{init}}, F_{\text{final}}], & \text{if } \alpha = 1, \end{cases} \tag{1}$$

where $\alpha \sim \text{Bernoulli}(p)$. Predicting the final state conditioned on the given motion endows our model with the ability to estimate consequences. On the other hand, predicting a transition state enhances the understanding of the causality involved in state transitions. By learning these two situations simultaneously during training, the model fosters a holistic comprehension of "how-to-interact" and "where-to-interact".

B. Encoder: Representation of Interactive Dynamics

Decoupled Multi-modal Encoder. The multi-modal encoder plays a central role in learning representations, as illustrated by the components highlighted in light blue and orange in Fig. 2. For our visual encoder design, we adhere to the established Vision Transformer (ViT) framework [15]. The input frame F_i is divided into non-overlapping patches with a window size s, and then transformed into a sequence of visual embeddings $v \in \mathbb{R}^{L_{\text{vis}}\times d}$ suitable for processing by ViT. Here, L_{vis} represents the length of the visual embedding, and d stands for feature dimension. The value of L_{vis} is equal to the number of divided windows HW/s^2, where H and W denote the height and width of the input frame accordingly.

In line with Voltron [28], we adopt DistillBERT [47] as the language encoder. The language description of the interaction process is initially tokenized into a sequence of numerical identifiers based on their position within a predefined vocabulary set V, and then padded to the maximum length L_{lang}. Consistent with previous multi-modal representation learning frameworks [28, 40], the language encoder remains frozen (without parameter updates). In addition, our visual encoder and language encoder are decoupled, enabling the omission of language input in downstream tasks.

Causality Modeling. The causality modeling module takes as input the visual embedding $\{v_0, v_1\} \in \mathbb{R}^{L_{\text{vis}}\times d}$ from two parallelly encoded frames. The embedding corresponding to the initial state is used as queries, while the subsequent state serves as both keys and values. These encoded frames undergo processing through a deformable attention layer [62] to capture patch-to-patch relations. This process can be elucidated as:

$$\begin{aligned} v_0' &= \texttt{DeformAttn}(Q = v_0, K = V = v_1), \\ v &= \texttt{Norm}(v_0 + v_0'). \end{aligned} \tag{2}$$

The merged visual representation $v \in \mathbb{R}^{L_{\text{vis}}\times d}$, which encodes causality relationships between the input states, is then leveraged in the decoder. In addition, this module reduces the computational burden in subsequent stages by combining the embeddings of the two frames into a unified representation. For downstream adaptations with single-image input, we simply replicate image features from the ViT to generate inputs for this module.

Token Aggregator. We introduce the token aggregator, which is motivated by two primary considerations: **1)** To address the challenge posed by disparate lengths between visual and language embeddings $\{v, l\}$, we aim to enable vision encoders to generate language-aligned, semantic-rich features by minimizing the distance between aggregated tokens $\{\tilde{v}, \tilde{l}\}$, and **2)** By employing aggregated tokens as concise representations for downstream tasks, we facilitate fair comparisons with prior methods [40, 44] that inherently rely on aggregated features.

As discussed in Voltron [28], multiheaded attention pooling (MAP) [30] proves to be significantly more effective compared to alternative prevalent approaches, such as the `[cls]` token [44] or global pooling [40], for deriving compact representations from a sequence of feature embeddings. Following this insight, we also employ a MAP block, shared between vision and language representations, to extract aggregated tokens. The aggregation process begins with employing zero-initialized latent embedding vectors, denoted as $\{\tilde{v}, \tilde{l}\} \in \mathbb{R}^d$ for vision and language, respectively. Taking the visual part as an example, this process can be formulated as follows:

$$\begin{aligned} \tilde{v} &= \texttt{Norm}(\tilde{v} + \texttt{Attn}(Q = \tilde{v}, K = V = v)), \\ \tilde{v} &= \texttt{Norm}(\tilde{v} + \texttt{MLP}(\tilde{v})). \end{aligned} \tag{3}$$

C. Decoder: Prediction and Detection

Prediction Transformer. The detailed structure of the Prediction Transformer is shown in deep blue on the right side of Fig. 2. Given visual and language embeddings from the encoder, the Prediction Transformer is responsible for predicting pixels of the frame that represent the unseen interaction state. In contrast to the original MAE [27] where queries are used to represent masked regions in the input image, our prediction process starts with a set of randomly initialized queries $q_{\text{pred}} \in \mathbb{R}^{L_{\text{vis}}\times d}$, which correspond to each patchified window of the target frame. To obtain the necessary information for accurate prediction, we perform cross-attention between the visual and language embeddings $\{v, l\}$ and use q_{pred} as queries. Additionally, we enhance the prediction queries by incorporating semantic information related to the interaction object, as encapsulated in the detection query q_{det}, using a bidirectional attention mechanism. After the iterative processing within the multi-layer decoder, the prediction queries are linearly projected (with $\mathcal{F}$) to generate pixel-level prediction results

denoted as $\hat{F}_{\text{pred}} = \mathcal{F}(q_{\text{pred}}) \in \mathbb{R}^{3\times H\times W}$.

Detection Transformer. Taking into consideration the object-centric nature of an interaction process [13], we introduce a detection query $q_{\text{det}} \in \mathbb{R}^d$ to locate the interaction object, as highlighted in purple in Fig. 2. To expedite training convergence, we initialize the detection query with the aggregated visual embedding $\tilde{v}$. It is noteworthy that the detection query is designed to regress the object location mainly based on the information inferred from the prediction queries q_{pred}, which we expect to contain comprehensive information about the target state. The exchange of information occurs through bidirectional attention between the two sets of queries. A subsequent two-layer multi-layer perceptron (MLP) is applied to obtain the regression box: $\hat{b}_{\text{det}} = \text{MLP}(q_{\text{det}}) \in \mathbb{R}^4$.

Our empirical analysis reveals that, despite the slow convergence observed in the early training phases due to inadequately informative representations on both ends, the transmission of query information facilitates eventual convergence for both tasks. The integrated modeling and optimization of these two tasks synergistically enhance the model's capacity to acquire a more effective representation for robot manipulations.

D. Optimization Objectives

The primary objective of our proposed framework during pre-training is to predict the unseen state and detect interaction objects, which are framed as "how-to-interact" and "where-to-interact", respectively. For target frame prediction, we utilize mean-squared error as the loss function, while for detection, we supervise the model with L1 loss and GIoU loss [46]. Furthermore, our preliminary experiments demonstrate that incorporating contrastive loss $\mathcal{L}_{\text{con}}$ between aggregated visual and language embeddings (*i.e.*, $\tilde{v}$ and $\tilde{l}$) improves representation learning. In summary, our pre-training framework aims to minimize the following components of loss:

$$\begin{aligned} \mathcal{L}_{\text{pred}} &= \text{MSE}(\hat{F}_{\text{pred}}, F_{\text{gt}}), \\ \mathcal{L}_{\text{det}} &= \text{L1}(\hat{b}_{\text{det}}, b_{\text{gt}}) + \text{GIoU}(\hat{b}_{\text{det}}, b_{\text{gt}}), \\ \mathcal{L}_{\text{con}} &= \text{InfoNCE}(\tilde{v}, \tilde{l}), \\ \mathcal{L} &= \lambda_1\mathcal{L}_{\text{pred}} + \lambda_2\mathcal{L}_{\text{det}} + \lambda_3\mathcal{L}_{\text{con}}, \end{aligned} \tag{4}$$

where F_{gt} represents the target frame, b_{gt} denotes the bounding box annotation, and $\lambda_1, \lambda_2, \lambda_3$ are balancing factors for the different learning objectives. The InfoNCE loss [41] employs cosine similarity as the distance measure.

IV. Experiments

A. Evaluation Suites

We have devised a comprehensive evaluation suite consisting of four robot learning tasks. Specifically, the suite comprises the following components:

1) Visuomotor control on real-world robots (Sec. IV-C1): This task aims to validate the applicability of our proposed method in various tasks in real-world scenarios, and assess the generalization capabilities.
2) Visuomotor control on Franka Kitchen (Sec. IV-C2): In this task, we benchmark the performance of our approach in complex simulation environments, comparing it against the previous state-of-the-art.
3) Visuomotor control on Meta-World (Sec. IV-C3): The target is to verify the ability to identify interaction objects with unfixed locations and perform manipulation.
4) Referring expression grounding (Sec. IV-D): This task evaluates the representation of object-centric semantics and spatial relationships conditioned by language.

Following the established practice in previous literature [28, 40, 44], each evaluation includes adaptation data and evaluation metrics. We evaluate MPI and various baseline models by freezing the pre-trained vision encoders and adapting task-specific "heads" on top of the extracted representations. In the following sections, we provide the rationale for selecting each evaluation task and present the corresponding experimental results. Furthermore, we delve into ablations on pre-training paradigms and model designs in Sec. IV-E.

B. Pre-training Details

Due to the labor-intensive nature of collecting robot manipulation data, we leverage large-scale egocentric human video datasets as the source for pre-training. Instead of using the entire Ego4D dataset [21], we utilize the hand-object interaction subset only. This dataset contains recordings of dynamic interaction processes during manual manipulation, captured with head-mounted cameras. Each video clip is accompanied by a textual description of the ongoing action. Additionally, the dataset annotates keyframes that capture critical moments, including the pre-change, change onset, and post-change states, which we refer to as the initial, transition, and final states in our paper. The bounding box annotations for the interaction objects in these three frames are also provided. We select 93K video clips for pre-training.

All models are pre-trained on 8 NVIDIA A100 GPUs. The ViT-Small version is trained for 200 epochs, using an initial learning rate of 1e-4 and a batch size of 64 per GPU. We choose the AdamW [35] optimizer. Training ViT-Small versions take approximately 15 hours. The ViT-Base version follows most of the same settings as the small version but is trained for a total of 400 epochs. In our experiments, we assign uniform weights for the loss terms, specifically, λ_1, λ_2 and λ_3 equal to 1, without adjustments for performance improvement. Further implementation details can be found in the Appendix, and we release the codes and models publicly for reproduction purposes.

C. Evaluations on Visuomotor Control

1) Real-world Robots

Motivation. In order to authentically assess the effectiveness of various visual representation learning approaches for enabling efficient robotic learning in real-world environments, we have introduced a series of manipulation tasks.

Evaluation Details. In the deployment of a Franka Emika Panda robot for a series of tasks, we employ a 3D SpaceMouse

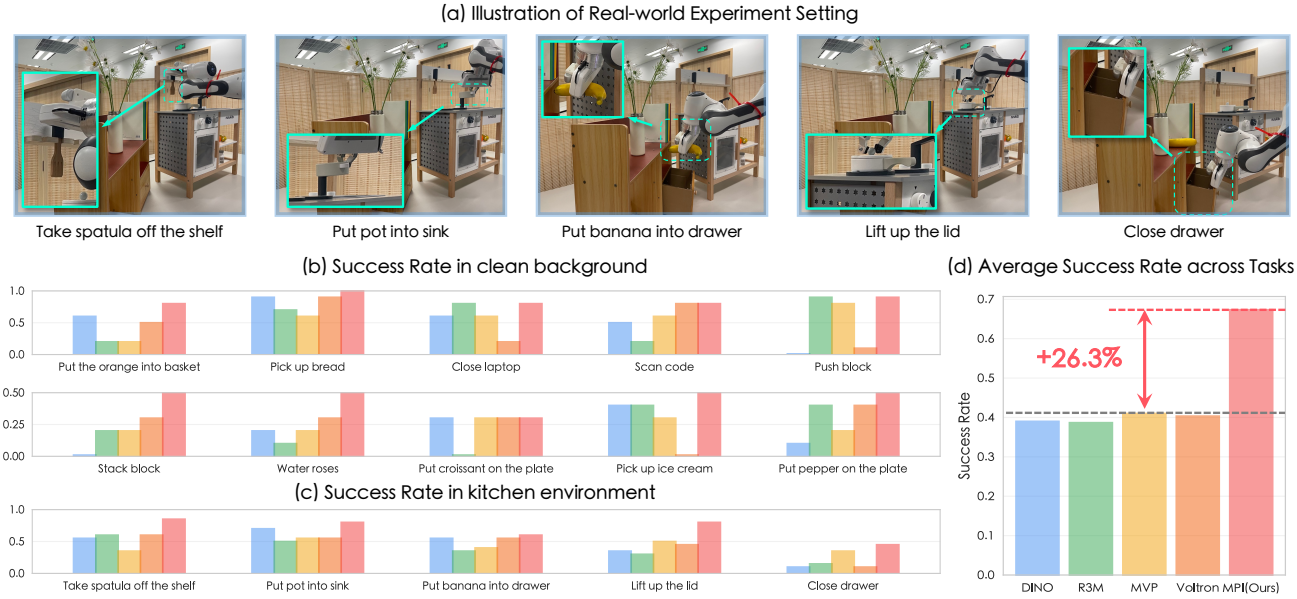

Fig. 3: **Real-world robot experiments.** (a) Illustrations of real-world experiments in the kitchen environment. (b) Detailed success rate of ten tasks within a clean background. (c) Results of five tasks in the complex kitchen environment. (d) MPI outperforms previous state-of-the-art with an average elevation of 26.3% success rate across 15 tasks.

to collect 10 teleoperated demonstrations for each task. An Operational Space Controller operating at 20Hz, as detailed in Deoxys [63], is utilized. In the evaluation phase, we use the frozen visual encoder to extract visual representations. These visual representations are then concatenated with the robot's proprioceptive state as inputs to a shallow MLP policy head. The objective is to predict continuous delta end-effector poses through action chunking [61]. To ensure fair comparisons with prior works [40, 44], we solely employ the aggregated visual embedding without any language interference. Each task is conducted 20 times, and we report the average success rate.

To provide a comprehensive evaluation of the effectiveness of different pre-trained encoders, we design two distinct scenarios with varying levels of complexity. The first scenario consists of ten diverse manipulation tasks in a *clean background*. These tasks include 1) putting the orange into the basket, 2) stacking the block, 3) picking up the ice cream, 4) closing the laptop, 5) scanning code, 6) watering roses, 7) putting the croissant on the plate, 8) picking up the bread, 9) pushing the block, and 10) putting the pepper on the plate. These tasks require fundamental manipulation skills such as Pick & Place, articulated object manipulation, *etc*.

In addition, we construct a more challenging *kitchen environment* that incorporates various interfering objects and backgrounds relevant to the target tasks. In this environment, we present five tasks: 1) taking the spatula off the shelf, 2) putting the pot into the sink, 3) putting the banana into the drawer, 4) lifting the lid, and 5) closing the drawer. As shown in Fig. 3(a), the complexity of these scenarios necessitates the visual encoder to possess both the "how-to-interact" and "where-to-interact" abilities to effectively handle these tasks.

Experimental Results. The results for tasks conducted in a *clean background* are presented in Fig. 3(b). While certain models exhibit superior performance in specific tasks (*e.g.*, the Voltron excels in the "scan code" task), they fail to consistently maintain high performance across all tasks. In contrast, MPI directs its focus towards the interactive objects and attains a more broadly applicable representation, consequently yielding consistently elevated performance across a spectrum of tasks.

Within the more challenging *kitchen environment*, as shown in Fig. 3(c), MPI exhibits outstanding performance across all manipulation tasks. Notably, it demonstrates particular proficiency in tasks that require precise object perception to interact with small objects, such as the "lift the lid" and the "take spatula off". Overall, as demonstrated in Fig. 3(d), MPI surpasses prior approaches, leading to an average success rate increase of **26.3%** across a broad spectrum of 15 tasks.

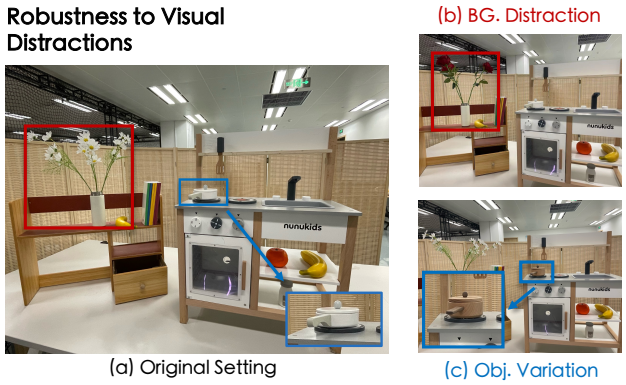

Fig. 4: **Illustration of real-world validation on generalization** to (b) background (BG.) distraction when we put a banana into the drawer, and (c) object variation when we lift the lid.

TABLE I: **Evaluation on generalization to background distraction and object variation.** MPI performs best in terms of success rate (%) under interferences nonetheless, benefiting from the robust representations it learns.

Method	Put banana into drawer		Lift up the lid	
	Original	Distraction	Original	Variation
DINO [7]	55	35 (↓36.4%)	35	20 (↓42.9%)
R3M [40]	35	25 (↓28.6%)	30	15 (↓50.0%)
MVP [44]	40	25 (↓37.5%)	30	15 (↓50.0%)
Voltron [28]	55	25 (↓54.5%)	45	20 (↓55.6%)
MPI (Ours)	**60**	**55** (↓8.3%)	**80**	**50** (↓37.5%)

TABLE II: **Evaluation on generalization to object position and lighting condition.** MPI performs best in terms of success rate (%) under various interferences. *We incorporate variations in lighting along with shifts in position.

Method	Object Position	*Lighting Condition
Voltron [28]	30	0
MVP [44]	40	20
MPI (Ours)	**60**	**30**

Generalization to background distraction and object variation. To assess the generalization capability and robustness of visual representations, we design two distinct scenarios within the complex kitchen environment. The first task involves placing a banana into a drawer. We introduce background distractions by replacing daisies with roses, as illustrated in Fig. 4(b). In Fig. 4(c), we conduct another task: lifting up the pot lid, where we substitute the white pot with a wooden one.

The results are in Table I. It is evident that all existing approaches suffer degraded performance when encountering out-of-distribution situations. However, MPI demonstrates remarkable robustness against such disturbances, experiencing merely a modest 8.3% decrease in the presence of background distraction and a minimum drop of 37.5% in object variation. These results validate the generalizability of our method to unseen environments and manipulation objects.

Generalization to object position and lighting condition. To further validate the generalization ability, we benchmark MPI with other literature. We select "put banana into basket" as the ablation task and collect 100 demonstration trajectories, where the banana tails are randomly positioned in a 16cm × 16cm rectangular region. Evaluations are conducted at ten marked locations to ensure consistency and fairness across different methods. For instance, we regulate the tails of bananas pointing at ten marked points during inference. Detailed settings are depicted in the Appendix. The success rates are reported in Table II. Our approach surpasses Voltron by a notable margin regardless of object position variation or lighting condition variation. Note that the prevailing alternative, MVP, still falls short against these variations in the real-robot setting.

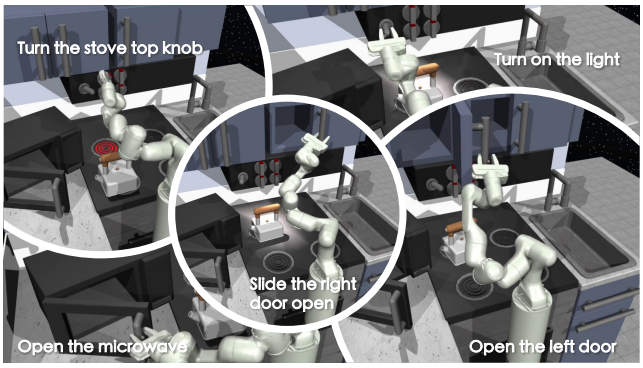

Fig. 5: **Franka Kitchen simulation environment** defined by Nair et al. [40]. In this environment, we include tasks of turning the stovetop knob, opening the microwave, sliding the right door open, turning on the light, and opening the left door. All tasks are trained with 25 demonstrations.

2) Franka Kitchen

Motivation. Previous studies [28, 40, 44] have established imitation learning for visuomotor control in simulation as the standard evaluation method. This enables direct comparisons with prior works and focuses on assessing the sample-efficient generalization of visual representations and their impact on learning policies from limited demonstrations. We conduct this evaluation to compare the capabilities of different representations in acquiring both the knowledge of "where-to-interact" and "how-to-interact" in complex simulation environments.

Evaluation Details. For the complex simulation environment, we adopt the policy learning tasks depicted in Fig. 5 within the Franka Kitchen simulation environment as defined by Nair et al. [40]. This environment consists of five distinct tasks, each observed from two camera viewpoints. To predict 9-DoF joint velocities (7 joints and 2 grippers) based on visual representations and proprioceptive states (*i.e.*, joint velocities), we train shallow MLP policy networks for methods with a ResNet50 backbone and modified multiheaded attention pooling (MAP) [30] for those with a ViT backbone. Following the means of Nair et al. [40] and Karamcheti et al. [28], we calculate the average success rates for each setting across the 5 tasks, 2 viewpoints, and 3 random seeds. We train a separate policy head for each task to imitate RL experts and use Nair et al. [40]'s codebase. Further discussions about the structure of policy networks are provided in the Appendix.

Experimental Results. The results of single-task visuomotor control in the Franka kitchen are listed in Table III. Notably, representation learning frameworks tailored for robot manipulation exhibit a discernible superiority over two widely adopted visual pre-training methods in the realm of computer vision: ImageNet [12] classification and CLIP [43] contrastive pre-training. While prior methodologies demonstrate comparable performance, our implementations, built upon both smaller and larger variants of visual backbone, showcase a significant advantage over our counterparts. In particular, MPI with

TABLE III: **Results of single-task visuomotor control on Franka Kitchen.** We report the success rate (%) over 50 randomly sampled trajectories. We **bold** the best result for models with similar parameters and underline the second. "INSUP." represents classification-based supervised learning on ImageNet. MPI consistently exhibits superior performance across multiple tasks.

Method	Backbone	Param.	Turn knob	Open door	Flip switch	Open microwave	Slide door	Average
INSUP. [25]	ResNet50	25.6M	28.0	18.0	50.0	26.7	75.7	39.7
CLIP [43]	ResNet50	25.6M	26.3	13.0	41.7	24.7	86.3	38.4
R3M [40]	ResNet50	25.6M	53.3	**50.7**	86.3	59.3	97.7	69.5
Voltron [28]	ViT-Small	22M	71.7	45.3	**95.3**	40.3	99.7	70.5
MPI (Ours)	ViT-Small	22M	**83.3**	50.3	89.0	**59.7**	**100.0**	**76.5**
MVP [44]	ViT-Base	86M	79.0	48.0	90.7	41.0	**100.0**	71.7
Voltron [28]	ViT-Base	86M	76.0	45.3	91.0	41.0	99.3	70.5
MPI (Ours)	ViT-Base	86M	**89.0**	**57.7**	**93.7**	**54.0**	**100.0**	**78.9**

TABLE IV: **Robustness evaluation under Franka Kitchen environment**, with varying background distraction and lighting conditions [5]. MPI shows greater robustness against background distraction and lighting condition variation.

Method	Distractors	Lighting	Average
R3M [40]	20.1	0	10.1
Voltron [28]	49.3	27.6	38.5
MVP [44]	51.3	24.0	37.7
DINO [7]	49.7	26.4	38.1
MPI (Ours)	**54.3**	**45.2**	**49.7**

the ViT-Small architecture yields a noteworthy enhancement of **+6%** in average success rate compared to the leading precedent Voltron [28]. This improvement becomes even more pronounced with the utilization of the ViT-Base backbone, resulting in an escalation of **+7.2%**. Furthermore, our approach consistently shows optimal or near-optimal performance across all tasks, highlighting its sustained advantage in handling the intricate scenarios encountered in the Franka Kitchen.

Robustness Evaluation. We upgrade the Franka Kitchen simulation suite to the improved setting in Burns *et al.* [5], which involves robustness evaluation under varying background distraction and lighting conditions. The comparisons are shown in Table IV. R3M, learning with multimodal contrastive objectives, leans towards extracting high-level semantic features, which leads to a dramatic performance loss under distractions. MPI shows superior generalization ability, where it outperforms the previous leading option MVP by 3% on average and boosts the success rate approximately twice in challenging darker lighting conditions.

3) Meta-World

Motivation. While the scene comprises only a singular object, Meta-World [59] introduces random variations in the location of the target interaction object during each validation, providing a complementary aspect to the fixed scenes presented in Franka Kitchen. Understanding "where-to-interact" prior to taking action enables the incorporation of implicit action priors into subsequent policy learning processes. Evaluations conducted on Meta-World emphasize the essential role of visual representations in policy learning.

Evaluation Details. We adopt the tasks defined in Meta-World, which involves assembling a ring onto a peg, picking and placing a block between bins, pushing a button, opening a drawer, and hammering a nail. Following the setup introduced in Sec. IV-C2, we train shallow MLP policy heads with 25 demonstrations separately for each task using behavioral cloning. The average success rates for each setting, encompassing the 5 specified tasks, 2 viewpoints, and 3 random seeds, are calculated to obtain the results.

Experimental Results for visuomotor control on Meta-World are presented in Table V. Remarkably, the smaller variant of MPI achieves the highest success rate, surpassing MVP [44] with visual backbones approximately four times larger than ours. In comparison to Voltron [28] which also uses ViT-Small, our lead extends even further to **17%**. The superior performance in Meta-World highlights the effective generalization of the positional object movement achieved by MPI.

D. Evaluations on Referring Expression Grounding

Motivation. The precise understanding of "where-to-interact" serves as a prerequisite for visuomotor control. It necessitates models to capture object-centric priors and high-level semantics. This entails considering interrelated properties such as color, light, and spatial relationships within the context of language expressions [28]. Referring Expression Grounding (R.E.G.) aims to predict the bounding box of an object in a cluttered scene based on a given language expression. The performance on this task is language-dependent, allowing us to assess the impact of pre-training with language instructions.

Evaluation Details. We use the OCID-Ref Dataset [53], which consists of scenes representing typical robotics environments. As language plays a crucial role in this task, we enhance the visual embedding by employing the DistillBERT model [47] to incorporate language information into R3M and MVP that originally lack a language encoder. By combining the visual embedding with the dimensions specified in Table VI and the language embedding from DistillBERT, we ensure a standardized approach for a fair comparison. To extract features from the concatenation of visual and language embeddings, we introduce an MAP block [30] that is learned from the

TABLE V: **Results of single-task visuomotor control on Meta-World simulation environment.** We report the success rate (%) over 50 randomly sampled trajectories. The best results are **bolded** and the second highest are underlined. MPI showcases exemplary performance across three tasks, exhibiting a superior average success rate in comparison to prior methods.

Method	Backbone	Param.	Assemble	Pick & Place	Press Button	Open Drawer	Hammer	Average
R3M [40]	ResNet50	25.6M	**94.0**	60.3	66.3	100	93.7	82.9
MVP [44]	ViT-Base	86M	82.7	**82.0**	62.7	100	95.7	84.6
Voltron [28]	ViT-Small	22M	72.3	57.3	30.7	100	83.0	68.7
MPI (Ours)	ViT-Small	22M	69.0	64.0	**98.7**	**100**	**96.0**	**85.7**

TABLE VI: **Results of Referring Expression Grounding.** We report results of smaller variants (ResNet50 for R3M, and ViT-Small for MVP, Voltron and ours), measured by average precision (%) under three IoU thresholds. *We leverage the aggregated visual embedding $\tilde{v}$ from the encoder. MPI yields the best detection results regardless of employing full-length visual embeddings or adopting aggregated embedding.

Method	Embedding	Average Precision (AP) @0.25	@0.5	@0.75
R3M [40]	$\mathbb{R}^{2048}$	85.27	71.79	42.66
MVP [44]	$\mathbb{R}^{384}$	93.07	85.32	60.37
Voltron [28]	$\mathbb{R}^{196\times384}$	92.93	84.70	57.61
MPI (Ours)*	$\mathbb{R}^{384}$	**96.29**	**92.10**	71.87
MPI (Ours)	$\mathbb{R}^{196\times384}$	96.04	92.05	**74.40**

task data. Further details regarding adaptation procedures and additional analysis are provided in the Appendix.

Experimental Results are in Table VI. Despite its effectiveness for visuomotor control, R3M [40] struggles to accurately locate objects based on language descriptions, as evident from its modest 42.66% AP under 0.75 IoU. This limitation may stem from its pre-training objective, which exclusively relies on contrastive loss to acquire high-level semantic features, with the loss of low-level localization features after global average pooling. In contrast, our method, with embeddings of the same dimension, achieves superior AP compared to the previous leading method MVP [44]. We observe improvements of **+3.22%**, **+6.78%**, and **+11.5%** under three different IoUs. In addition, the token aggregator learned from pre-training enables the model to achieve comparable or even better performance with fewer embeddings, demonstrating the effectiveness of our interaction-oriented learning target.

E. Ablations & Further Analysis

Input Frames. MPI uses three keyframes from a video clip, distinguishing it from conventional pre-training approaches based on video prediction. To validate this design, we evaluate representation learning with three consecutive frames, as presented in the top row of Table VIIa. Specifically, the selection of data frames centers around the transition frame F_{trans} or the final frame F_{final} with corresponding box annotations. This is to support the learning of interaction object detection and maintain consistency with our designated learning objectives.

As outlined in Table VIIa, the performance with sequential frames is inferior to our pre-training paradigm, which is grounded upon key interaction states. Furthermore, the alternating use of transition and final frames as prediction targets (with p set to 0.5 in Eq. (1)) enhances the model's comprehensive understanding of the interaction process, leading to further advancements in performance.

Decoder Design. As depicted in the right section of Fig. 2, we employ a pair of transformer decoder blocks designated as the prediction transformer and detection transformer. Each decoder is assigned a specific role: the Prediction Transformer predicts the unseen interaction state, while the Detection Transformer infers the interaction object in the unseen frame. In Table VIIb, we investigate the functions performed by each decoder component. In contrast to employing transformer-based decoders, the first row of Table VIIb showcases an alternative approach where we utilize a 3-layer MLP to directly decode the target frame from encoder embeddings. The utilization of either of the introduced decoders leads to noteworthy improvements in both tasks, evidenced by an increase of over +12% average success rate on Franka Kitchen and a +26% AP improvement in the R.E.G. task. The synergistic integration of the prediction and detection transformer contributes to enhanced representation learning, ultimately achieving optimal performance and demonstrating the necessity of these two interaction-oriented designs in our framework.

Causality Modeling. For the multi-modal encoder, we design a causality modeling mechanism built upon the deformable attention [62] to capture the causality relationships between two input states during pre-training. As shown in Table VIIc, excluding this module results in a decline in performance. The comparable performance in R.E.G. can be attributed to the task's focus on static scene detection, which does not strictly require an understanding of interaction dynamics. It is important to note that the image features from the ViT are simply replicated to constitute inputs for this module in all downstream tasks. The additional parameters (*e.g.*, 0.3M for ViT-Small) and computational load introduced are negligible.

Involvement of Language. Our pre-training pipeline is language-conditioned. Nevertheless, we have decoupled the visual and language branches in our encoder. Therefore, the language modality is not requisite in downstream tasks. As shown in Table VIId, the inclusion of language leads to a +2% increase in success rate in Franka Kitchen. This improvement

TABLE VII: **Ablation studies and further comparisons.** For ablation analysis on real-world robot experiments, we select picking up the banana and placing it into the basket as our ablation tasks. We evaluate the task ten times and report the success rate (%). Moreover, we report the average success rate (%) on Franka Kitchen and the Average Precision (%) @ 0.5 IoU for the Refering Expression Grounding (R.E.G.) task. We report the results on Franka Kitchen with a fixed random seed.

(a) **Ablation of input frames.** The input selection is subjected to the Bernoulli distribution characterized by p (in Eq. (1)). Our pre-training methodology, employing the keyframes that signify states of interaction, exhibits heightened efficacy in comparison to the utilization of randomly selected frames.

Input Frames		Real-world Robot	Franka Kitchen	R.E.G. @0.5IoU
Sequential Frames		40	74.0	90.57
Key Frames (Ours)	$p = 0$	40	76.4	91.73
	$p = 0.5$	**60**	**79.2**	**92.04**
	$p = 1$	**60**	78.8	91.77

(b) **Ablation of decoder design.** *Prediction* and *Detection* represent the decoder blocks as introduced in Fig. 2. Both decoder blocks, targeting the holistic understanding of "how-to-interact" and "where-to-interact" within the interaction process, contribute indispensably to the improved representation learning and downstream performance.

Decoder		Real-world Robot	Franka Kitchen	R.E.G. @0.5IoU
Prediction	Detection			
		30	62.4	63.94
✓		50	74.8	90.58
	✓	50	75.2	90.72
✓	✓	**60**	**79.2**	**92.04**

(c) **Ablation of causality modeling mechanism.** *Dual-frame* indicates that the visual embedding of the two frames is fed to the decoder in parallel during pre-training.

Causality Modeling	Franka Kitchen	Robustness Evaluation	R.E.G. @0.5IoU
w/o (Dual-frame)	77.6	47.1	91.93
Deformable Attn.	**79.2**	**49.7**	**92.04**

(d) **Ablation of input modality.** For *Vision-only*, we discard language prior in policy learning. Language modality is beneficial while competitive performance is still achieved without language.

Input Modality	Franka Kitchen	Robustness Evaluation
Vision-only	77.2	44.3
Vision + Language	**79.2**	**49.7**

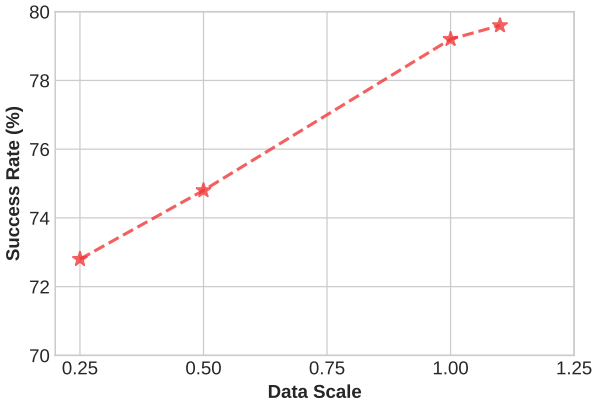

Fig. 6: **Influence of pre-training data scale.** We test on the single-task visuomotor control task. MPI demonstrates improvement with scaled pre-training data, holding the potential for further performance enhancement with large-scale data.

primarily originates from the ability of textual features to highlight the most relevant visual features for the task. Note that notwithstanding the absence of language, MPI still achieves higher performance compared to other representation learning frameworks evaluated in Table III.

Data Regime. We conduct ablation experiments on data scale, where we pre-train MPI using data volumes of 25%, 50%, 100%, and 110%, respectively. Subsequently, we evaluate them in the Franka Kitchen environment. Results in Fig. 6 demonstrate that increasing the scale of pre-training data correlates with improved performance in downstream tasks.

V. Conclusions

In this work, we present MPI, an interaction-oriented representation learning framework to empower manipulation tasks. By predicting transition states based on initial and final states, MPI enhances the model's understanding of interactive dynamics. Through extensive experiments, we demonstrate that our method achieves state-of-the-art performance across a diverse range of downstream tasks.

Limitations and Future Work. Our framework by far utilizes explicit annotations (*i.e.,* keyframes, languages, and bounding boxes for interaction object) provided in the Ego4D-HoI [21] dataset. This could limit the applicability of our methods to broader datasets. Nonetheless, it is feasible to leverage some well-established vision tools [8, 34] to collect annotations cheaply and scale up the pre-training dataset with web-crawled images. Additionally, the policy model employed in this study, which relies on behavior cloning from a limited amount of expert demonstrations, restricts the generalization of the model to different scenarios. Besides, the lack of a large language model constrains the model's ability to perform long-horizon planning and causal reasoning. Exploring the incorporation of the pre-trained model from MPI into vision-language models to tackle long-horizon interaction-related tasks represents an interesting direction for future research.

Acknowledgments

This work was supported by National Key R&D Program of China (2022ZD0160104), NSFC (62206172), Shanghai Committee of Science and Technology (23YF1462000), and China Postdoctoral Science Foundation (2023M741848).

References

[1] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *CVPR*, 2023. 3

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 2

[3] Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science*, 2019. 2

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S Ryoo, Grecia Salazar, Pannag R Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics transformer for real-world control at scale. In *RSS*, 2023. 1, 2

[5] Kaylee Burns, Zach Witzel, Jubayer Ibn Hamid, Tianhe Yu, Chelsea Finn, and Karol Hausman. What makes pre-trained visual representations successful for robust manipulation? *arXiv preprint arXiv:2312.12444*, 2023. 3, 8

[6] Kaylee Burns, Zach Witzel, Jubayer Ibn Hamid, Tianhe Yu, Chelsea Finn, and Karol Hausman. What makes pre-trained visual representations successful for robust manipulation? *arXiv preprint arXiv:2312.12444*, 2023. 2

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 7, 8

[8] Joya Chen, Difei Gao, Kevin Qinghong Lin, and Mike Zheng Shou. Affordance grounding from demonstration video to target image. In *CVPR*, 2023. 3, 10

[9] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *arXiv preprint arXiv:2306.16927*, 2023. 2

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 1

[11] Sudeep Dasari, Mohan Kumar Srirama, Unnat Jain, and Abhinav Gupta. An unbiased look at datasets for visuo-motor pre-training. In *CoRL*, 2023. 3

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 7

[13] Coline Devin, Pieter Abbeel, Trevor Darrell, and Sergey Levine. Deep object-centric representations for generalizable robot learning. In *ICRA*, 2018. 5

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 1

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4

[16] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Kaelbling, Andy Zeng, and Jonathan Tompson. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023. 2

[17] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *NeurIPS*, 2023. 3

[18] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *ICRA*, 2017. 2

[19] James J Gibson. The ecological approach to the visual perception of pictures. *Leonardo*, 1978. 3

[20] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 1, 2

[21] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi,

Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 1, 2, 4, 5, 10

[22] Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. In *CoRL*, 2019. 2

[23] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. MaskViT: Masked visual pre-training for video prediction. In *ICLR*, 2023. 2

[24] Nicklas Hansen, Zhecheng Yuan, Yanjie Ze, Tongzhou Mu, Aravind Rajeswaran, Hao Su, Huazhe Xu, and Xiaolong Wang. On pre-training for visuo-motor control: Revisiting a learning-from-scratch baseline. *arXiv preprint arXiv:2212.05749*, 2022. 2

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 8

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 2, 4

[28] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. In *RSS*, 2023. 2, 4, 5, 7, 8, 9

[29] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *CVPR*, 2022. 2

[30] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, 2019. 4, 7, 8

[31] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. LOCATE: Localize and transfer object parts for weakly supervised affordance grounding. In *CVPR*, 2023. 3

[32] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. ManipLLM: Embodied multimodal large language model for object-centric robotic manipulation. *arXiv preprint arXiv:2312.16217*, 2023. 3

[33] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023. 2

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 10

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[36] Yecheng Jason Ma, William Liang, Vaidehi Som, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. LIV: Language-image representations and rewards for robotic control. In *ICML*, 2023. 2

[37] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *ICLR*, 2023. 2

[38] Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. Grounding language with visual affordances over unstructured data. In *ICRA*, 2023. 3

[39] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. In *RSS*, 2023. 3

[40] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *CoRL*, 2022. 2, 4, 5, 6, 7, 8, 9

[41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[42] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. In *ICML*, 2022. 2

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 7, 8

[44] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *CoRL*, 2022. 2, 4, 5, 6, 7, 8, 9

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2

[46] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 5

[47] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 1, 4, 8

[48] Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked world models for visual control. In *CoRL*, 2022. 2

[49] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine,

and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *ICRA*, 2018. 2

[50] Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action. In *CoRL*, 2022. 1

[51] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. ViNT: A foundation model for visual navigation. In *CoRL*, 2023. 1

[52] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. CLIPort: What and where pathways for robotic manipulation. In *CoRL*, 2021. 2

[53] Ke-Jyun Wang, Yun-Hsuan Liu, Hung-Ting Su, Jen-Wei Wang, Yu-Siang Wang, Winston Hsu, and Wen-Chin Chen. OCID-Ref: A 3d robotic dataset with embodied language for clutter scene grounding. In *NAACL*, 2021. 2, 8

[54] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *ICLR*, 2024. 2

[55] Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In *NeurIPS*, 2022. 1

[56] Penghao Wu, Li Chen, Hongyang Li, Xiaosong Jia, Junchi Yan, and Yu Qiao. Policy pre-training for autonomous driving via self-supervised geometric modeling. In *ICLR*, 2023. 2

[57] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *ICLR*, 2024. 3

[58] Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *CVPR*, 2024. 2

[59] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *CoRL*, 2020. 2, 8

[60] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R. Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, Nima Fazeli, Ferran Alet, Nikhil Chavan Dafle, Rachel Holladay, Isabella Morena, Prem Qu Nair, Druck Green, Ian Taylor, Weber Liu, Thomas Funkhouser, and Alberto Rodriguez. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In *ICRA*, 2018. 1, 3

[61] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *RSS*, 2023. 6

[62] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 4, 9

[63] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. VIOLA: Imitation learning for vision-based manipulation with object proposal priors. In *CoRL*, 2022. 6

[64] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R Sanketi, Grecia Salazar, Michael S Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J Joshi, Alex Irpan, brian ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023. 2