Masked Visual-Tactile Pre-training for Robot Manipulation

Qingtao Liu[1], Qi Ye[1†], Zhengnan Sun[1], Yu Cui[1], Gaofeng Li[1], Jiming Chen[1]

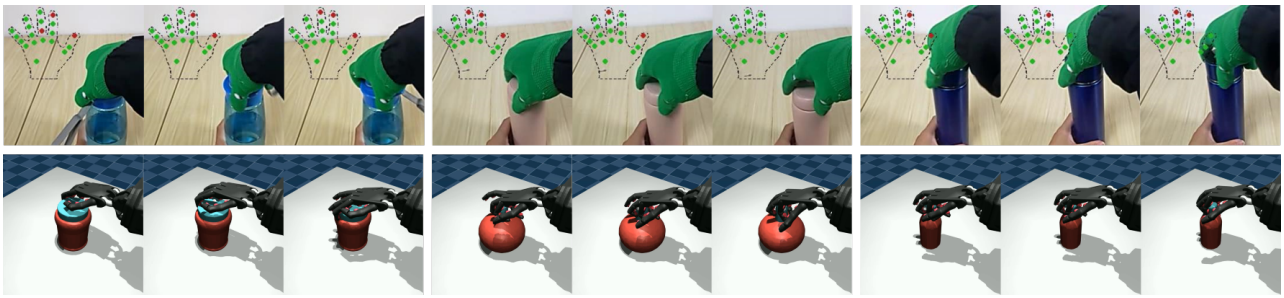

Fig. 1: The first row shows some examples of our bottle-cap turning dataset. The second row shows some simulation results of the bottle-cap task with Shadow Hand [1].

***Abstract*— Recent works on the pretraining for robot manipulation have demonstrated that representations learning from large human manipulation data can generalize well to new manipulation tasks and environments. However, these approaches mainly focus on human vision or natural language, neglecting tactile feedback. In this article, we make an attempt to explore how to pre-train a representation model for robotic manipulation using both human manipulation visual and tactile data. We develop a system for collecting visual and tactile data, featuring a cost-effective tactile glove to capture human tactile data and Hololens2 for capturing visual data. With this system, we collect a dataset of turning bottle caps. Furthermore, we introduce a novel visual-tactile fusion network and learning strategy M^{2}VTP, with one key module to tokenize 20 sparse binary tactile signals sensing touch states for the learning of tactile context and the other key module applying the attention and mask mechanism to the interaction of visual and tactile tokens for visual-tactile representation learning. We utilize our dataset to pre-train the fusion model and embed the pre-trained model into a reinforcement learning framework for downstream tasks. Experimental results demonstrate that our pre-trained model significantly aids in learning manipulation skills. Compared to methods without pre-training, our approach achieves a success rate increase of over 60%. Additionally, when compared to current visual pre-training methods, our success rate exceeds them by more than 50%.**

I. Introduction

For robot manipulation, learning representations from in-domain data [2], [3], [4] often faces challenges in terms of generalizing to unseen environments and tasks. Recent researches [5], [6], [7] have demonstrated the representations acquired through the pretraining of large-scale models on out-of-domain data exhibit strong generalization. Leveraging the approach of pretraining large models proves effective in addressing this challenge. However, collecting robotic data incurs high costs, making it exceedingly challenging to train a large model with a substantial amount of robotic data.

Alternatively, some prior efforts [8], [9], [10], [11], [12] have explored the use of extensive out-of-domain human manipulation dataset [13], [14], [15] for pretraining models endowed with significant human-operated knowledge, utilized to guide downstream robotic tasks, yielding noteworthy outcomes. They primarily focus on the visual and natural language modalities of human manipulation data, without integrating tactile information. It has been demonstrated in the community that tactile sensing can offer valuable information when vision is obstructed and can capture finer local geometrical details of manipulated objects that may elude visual perception.

Therefore, in this paper, we aim to bridge the gap and explore how to pre-train a representation model with visual-tactile data, which presents two key challenges: 1) how to gather tactile data during human manipulation and 2) how to fuse vision and tactile modalities. In robotics, tactile information can be acquired through the deployment of F/T sensors, visual-based tactile sensors, or array tactile sensors. However, most F/T sensors in literature are designed attached to robotic grippers and visual-based tactile sensors are too bulky, impacting the dexterity of manipulation when attached to human hands. On the other hand, array sensors designed for dexterous hands with accurate sensing capability and flexibility are usually expensive, hindering the possibility of collecting large-scale data with them [16]. For the second challenge, although there have been many explorations in leveraging both visual and tactile information [17], [18], [19], [20], [21], [22], most are designed for a specific task in a reinforcement learning framework, where learning to extract feature from inputs can be very inefficient. There

[1]College of Control Science and Engineering, Zhejiang University, Hangzhou, 310027, China

[†]Qi Ye (Corresponding author, qi.ye@zju.edu.cn) is with the College of Control Science and Engineering and the State Key Laboratory of Industrial Control Technology, Zhejiang University, and also with the Key Key Lab of CS&AUS of Zhejiang Province.

This work was supported in part by the National Natural Science Foundation of China (Grant Number: 62088101, 62103372, 62233013).

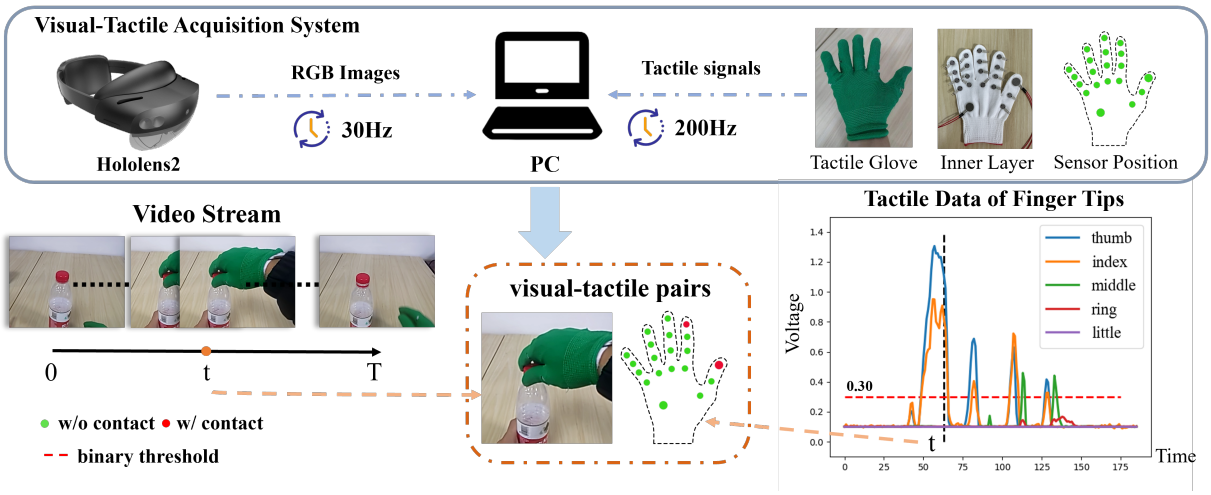

Fig. 2: **Visual-Tactile Acquisition System for Human Manipulation and Samples of the Collected Dataset. The top** displays the visual acquisition system. **The top right** illustrates the external and internal structure of our tactile glove, including the positions of tactile sensors; **the bottom** presents one set of collected data, consisting of aligned frames of images and corresponding tactile pairs.

is a lack of work for learning an effective visual-tactile representation model.

To tackle the first challenge, we design a visual-tactile data acquisition system to capture both visual and tactile information during human manipulation. We develop a cost-effective tactile glove with 20 small sensor patches attached to different hand parts, sensing whether these hand parts touch a surface or not. The scene is captured by a HoloLens2 [23] synchronized with the glove and worn by the performer from an ego-centric view. With the system, we collect a dataset comprising 120 manipulation sequences of unscrewing bottle caps to initiate the exploration for a visual-tactile pretrain model.

To acquire a joint representation of visual and tactile information, we propose a novel visual-tactile fusion framework M^{2}VTP built on the Masked Autoencoder (MAE) network [7]. Particularly, two modules are designed to embed the sparse and binary tactile information and fuse the vision and tactile modality. For the tactile embedding module, the 20 binary tactile signals are mapped to high dimensional codes via MLPs which are then added with positional encodings of the identities of these sensors to form tokens. The positional encoding for each hand part sensor is designed to enable the possibility of modeling the context information or the coordination between different hand parts during manipulation. Together with image tokens, these tactile tokens are fed into transformer blocks. The attention and mask mechanism enables each token to leverage information from other tokens to reconstruct itself when masked, which facilitates the effective learning of latent feature representations for visual-tactile information.

Further, we embed the pre-trained model into a reinforcement learning framework and feed the latent representations from the Transformer encoder into the policy network to verify the effectiveness of the petrain model. To the best of our knowledge, we are the first to utilize human visual and tactile data for pretraining representation models and demonstrate their utility in robotic manipulation tasks. In summary, our paper makes the following key contributions:

- We propose a novel Visual-Tactile Fusion framework M^{2}VTP, which enables representation learning from human visual-tactile data through pretraining and assists robot manipulation.
- We design a visual-tactile data collection system to capture both visual and tactile data during human manipulation. We collect a dataset to initiate the study of visual-tactile pretraining from human data.
- We utilize the collected data to pretrain our framework and apply the latent representations to a bottle cap turning task. Our approach achieves significant success compared to baseline methods.

II. Related Works

A. *Pre-training for robotics.*

Pre-trained models, which can be trained through unsupervised learning methods [7], [24], [5], [25] have garnered significant attention, due to their low cost and strong generalization. To learn the generalizable manipulation policy, many prior works explore the utility of pretraining models using extensive human manipulation datasets [14], [15], [13] or in-the-wild videos to guide downstream robotic manipulation tasks. These methods [9], [26], [10], [27] employ the visual modality as the subject of representation learning. Additionally, some research [8], [12], [11] endeavors focus on learning representation models through the fusion of multiple modalities. Meanwhile, SGR [28] utilizes CLIP [5] for feature extraction from both images and text and integrates 3D point cloud data. However, for robotic manipulation, tactile perception is an indispensable sensory mode. Our work marks the first instance of incorporating human touch sensing into pre-trained models.

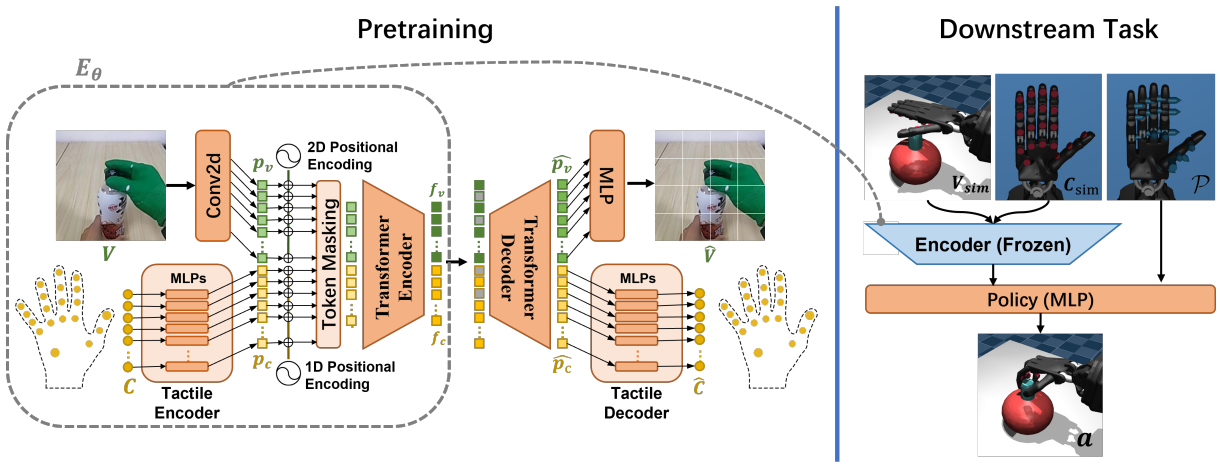

Fig. 3: **The M^{2}VTP Framework and Its Downstream Adaptation. The left** depicts an illustrative diagram of the M^{2}VTP pretraining framework. It takes a set of aligned images and tactile signals (V, C) as input. Inputs from different modalities are encoded into embeddings through their respective encoders and then fed into the MAE backbone. **The right** shows the pretrained M^{2}VTP encoder assisting in downstream RL training tasks.

B. Manipulation with tactile.

Due to its ability to provide interactive information, tactile sensing is extensively employed in robotic manipulation tasks. Lee et al. [17] and VTT [18] use tactile data collected by F/T sensors to learn the peg-intersection task. Visual-based tactile sensors like Gelsight [29], find extensive use in tasks like pouring liquids [19] and cable manipulation [30]. Array sensors are employed for learning tasks such as Baoding ball [31] and book opening [22]. Yin et al. [32] train an in-hand object rotation policy with reinforcement learning using only sparse binary tactile signals. Similarly, Li [33] uses hand contact information to generate a proper grasping hand pose.

C. Fusion model with tactile.

There have been numerous efforts to fuse visual and tactile information. [17], [18], for instance, integrated tactile signals collected by F/T sensors with other visual modalities to learn tasks such as inserting plugs. These works [19], [20], [21] combine tactile images from vision-based tactile sensors with visual images. Another line of research [22] fuses signals from densely arranged array tactile units with images. However, these fusion methods have been task-specific, and there has been no mention of a model suitable for large-scale pretraining on visual-tactile data.

III. Method

In this section, we introduce how to make a low-cost touch glove and build up a visual touch acquisition system to collect a visual-tactile dataset for human manipulation in Sec.III-A. We propose a novel visual-tactile fusion framework M^{2}VTP to fuse the vision and tactile modalities in Sec.III-B. We embed the pre-trained model into the RL structure, extracting visual-tactile latent representations to have the agent understand the environment in the downstream tasks in Sec.III-C.

A. Visual-tactile acquisition system for human manipulation

To collect human manipulation visual-tactile datasets, we fabricate a tactile glove and establish an integrated data acquisition system, comprising three main components: *1) the tactile glove* for capturing tactile data signals, *2) HoloLens2* for collecting visual data, and *3) a central computer* responsible for recording timestamps used to align the vision and tactile data. The overall architecture and physical setup of this system are illustrated in Fig2.

*1) **Fabrication and Calibration of Tactile Gloves**:* Our tactile glove uses low-cost commercial resistive pressure sensors. A total of 20 sensors are attached to the glove on all the hand parts to capture contact signals at these critical points during manipulation as shown in Fig.2 (top right). Each sensor unit has a diameter of 10 millimeters, with the exception of the thumb's distal joint and the palm, where sensors with a diameter of 18.3 millimeters are employed due to their larger surface area. These sensors are connected to resistance-voltage conversion modules to transform resistance signals into voltage signals. By adjusting potentiometers on the conversion modules, we ensure that the voltage values recorded when each sensor is in contact are approximately equal.

*2) **Visual-Tactile Data Collection**:* We employ the HoloLens2 to capture RGB images, with a resolution of 424x240 pixels and a frame rate of 30 frames per second (fps). Due to the substantial size of the image data, we record the timestamps corresponding to each image acquired by the HoloLens2 and transmit these timestamps to the central computer via UDP. The central computer records the received information and appends corresponding local timestamps.

For tactile signals, we connect the resistance-voltage conversion modules to the AD7606 module and utilize the STM32F205RBT6 microcontroller to sample tactile voltage signals at a rate of 200 Hz. The real-time tactile signals are stored in a FIFO buffer. The central computer periodically retrieves the tactile signals through a serial connection every

1 second and simultaneously records the beginning and end timestamps. Intermediate timestamps for tactile data are obtained using interpolation.

*3) **Alignment of vision and touch**:* The alignment of visual and tactile data relies on the timestamps recorded by the central computer when receiving signals from the two acquisition sources. We synchronize each image frame with a corresponding tactile frame to minimize the temporal disparity, forming a matched data pair.

B. Masked Visual-tactile Transformer for Pretraining

To leverage both visual and tactile data of human manipulation, we propose a novel pretraining framework M^{2}VTP, which extends the principles of the Masked Auto-Encoder (MAE) [7]. This framework is tailored for the integration and extraction of latent representations pretraining to visual and tactile signals encountered during human manipulation, subsequently amenable for utilization in downstream tasks. M^{2}VTP encompasses a visual-tactile encoder E_θ and a visual-tactile reconstructor D_θ, illustrated in Fig 3.

*1) **Visual-Tactile Encoder** $E_\theta : (V, C) \rightarrow (f_v, f_c)$:* The visual-tactile encoder E_θ is the core of M^{2}VTP. It takes in a set of aligned images and binary tactile signals ($V \in \mathbb{R}^{H\times W\times 3}$, $C \in \{0,1\}^{N_c}$) as input and produces extracted visual-tactile latent features ($f_v \in \mathbb{R}^{N_{uv}\times d_f}$, $f_c \in \mathbb{R}^{N_{uc}\times d_f}$) as output. Here, H and W represent the height and width of the input image, N_c signifies the count of tactile taxels on the tactile glove, N_{uv} and N_{uc} represent the number of visible visual and tactile embeddings, respectively, and d_f denotes the dimensionality of the encoded patches. E_θ consists of three components: the visual feature extractor F_{θ_v}, the tactile feature extractor F_{θ_c}, and the visual-tactile fusion block $B_{\theta_{vc}}$.

*a) **Visual Feature Extractor** $F_{\theta_v} : V \rightarrow b_v$:* The visual feature extractor F_{θ_v} is based on the MAE framework. It takes the input image V and yields visible visual embeddings $b_v \in \mathbb{R}^{N_{uv}\times d}$, where d signifies the dimensionality of the embeddings. The input image V undergoes convolutional operations to generate N_v visual patches $p_v \in \mathbb{R}^{N_v\times d}$, which are subsequently augmented with 2D sine-cosine positional encodings. Finally, a subset of these augmented patches is randomly discarded based on a mask ratio $r_v \in [0, 1)$, resulting in the generation of the visual embeddings b_v. Consequently, it can be inferred that $N_{uv} = N_v \times (1 - r_v)$.

*b) **Tactile Feature Extractor** $F_{\theta_c} : C \rightarrow b_c$:* The tactile feature extractor F_{θ_c} takes the tactile signal matrix C as input and produces visible tactile embeddings $b_c \in \mathbb{R}^{N_{uc}\times d}$. In order to facilitate the encoder E_θ in capturing the intrinsic relationships among tactile sensations across different regions of the hand, we generate embeddings for each tactile taxel on the tactile glove, corresponding to the contact values with objects. Each contact value $c \in \{0, 1\}$ in the input matrix $C \in \{0, 1\}^{N_c}$ is individually mapped through a dedicated Multi-Layer Perception (MLP), resulting in the generation of N_c tactile patches $p_c \in \mathbb{R}^{N_c\times d}$. We introduce 1D sine-cosine positional encodings for tactile patches p_c, distinguishing them from the positional encodings applied to the visual embeddings. To enhance the encoder E_θ's capacity to extract tactile features, we also randomly mask the tactile embeddings based on a mask ratio $r_c \in [0, 1)$. This leads to the ultimate formation of tactile embeddings b_c. It can be deduced that $N_{uc} = N_c \times (1 - r_c)$.

*c) **Visual-Tactile Fusion Block** $B_{\theta_{vc}} : (b_v, b_c) \rightarrow (f_v, f_c)$:* The visual-tactile fusion block $B_{\theta_{vc}}$ is the core component responsible for integrating and extracting features from both the visual and tactile modalities. Its input comprises the visible visual and tactile embeddings (b_v, b_c), while its output encompasses the visual-tactile features (f_v, f_c).

*2) **Visual-Tactile Reconstructor** $D_\theta : (f_v, f_c, m) \rightarrow (\widehat{V}, \widehat{C})$:* The visual-tactile reconstructor D_θ, takes input from two sources: a) the output (f_v, f_c) of E_θ and b) a set of mask tokens $m \in \mathbb{R}^{N_m\times d_f}$, where $N_m = r_v \times N_v + r_c \times N_c$. Likewise, it consists of three components: the visual-tactile reconstruction block $R_{\theta_{vc}}$, the visual reconstructor R_{θ_v}, and the tactile reconstructor R_{θ_c}.

*a) **Visual-Tactile Reconstruction Block** $R_{\theta_{vc}} : (f_v, f_c, m) \rightarrow (\widehat{p_v}, \widehat{p_c})$:* Before introducing the input set of tokens (f_v, f_c, m) into the transformer blocks, positional encodings, consistent with those utilized in the encoder section, are incorporated into (f_v, f_c, m) itself. This process yields reconstructed visual-tactile patches ($\widehat{p_v} \in \mathbb{R}^{N_v\times d}, \widehat{p_c} \in \mathbb{R}^{N_c\times d}$).

*b) **Visual Reconstructor** $R_{\theta_v} : \widehat{p_v} \rightarrow \widehat{V}$:* The visual reconstructor entails the utilization of an MLP that orchestrates the mapping of the input image reconstruction patches $\widehat{p_v}$, to reconstructed images $\widehat{V} \in \mathbb{R}^{H\times W\times 3}$.

*c) **Tactile Reconstructor** $R_{\theta_c} : \widehat{p_c} \rightarrow \widehat{C}$:* The tactile reconstructor R_{θ_c}, involves the individual mapping of each patch within the set $\widehat{p_c}$ through dedicated MLPs to derive tactile reconstruction values $\widehat{c} \in \mathbb{R}^1$ for each patch. Subsequently, these reconstructed values are concatenated to form the tactile reconstruction vector $\widehat{C} \in \mathbb{R}^{N_c}$.

*d) **Loss**:* To enable M^{2}VTP to simultaneously extract visual and tactile features, we formulate the following loss functions:

$$L(\theta) = W_{img} \times MSE(V, \widehat{V}) + W_{tac} \times MSE(C, \widehat{C}) \quad (1)$$

Here, $W_{img} = 1$ and $W_{tac} = 100$ are hyperparameters representing the loss weights for images and tactile data, respectively.

C. Visual-tactile RL for manipulations

Problem Formulation. We model the Visual-Tactile manipulation task as a Markov Decision Process (MDP), defined by a tuple: $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$. $\mathcal{S}$ and $\mathcal{A}$ represent the state and action space. The policy $\pi_\theta : \mathcal{S} \rightarrow \mathcal{A}$ maps the state space $\mathcal{S}$ to the action space $\mathcal{A}$. $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ is the transition dynamic. $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function and $\gamma \in (0, 1]$ is the discount factor. Our goal is to maximize the expected discounted reward $J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0} \gamma^t r\left(s_t, a_t\right)\right]$ to train a policy network π_θ. We use the PPO algorithm to make the agent learn manipulation skills. We illustrate our RL architecture in Fig.3 (Right Column).

Task. To better explore the application of vision and touch in robotic manipulation tasks, we set up Bottle-Cap Turning in MuJoCo [34], where the bottles stand in the center of the table. The manipulator (Shadow Hand with 24 Dofs) is fixed above the bottle. We select 15 different bottles from ShapeNet [35] (10 for training and 5 for evaluation). A joint is added to each bottle cap, allowing it to rotate along the z axis. We arrange 20 tactile sensors to get touch data and a camera to capture the ego-centric view.

State Space. As previous works [12], [9], [10], [11], [8] do, we freeze the pre-trained model and feed the latent features into the policy network, which makes full use of the perception ability of the pre-trained model and reduces the burden of policy network learning. Thus, the state space $\mathcal{S} = \{\mathrm{E}_{\theta_f}(V_{sim}, C_{sim}), \phi(\mathcal{P})\}$, where E_{θ_f} is the pre-trained Transformer Encoder with frozen parameter. V_{sim} is captured by the ego-centric camera. C_{sim} is binary tactile signals achieved by applying a threshold of $\lambda_{th} = 0.01N$. The proprioception $\mathcal{P}$ contains the positions and velocities of all the robotic hand joints and $\phi(\cdot)$ is a linear layer to project the proprioception input into a vector, whose dimension is the same as the output of $\mathrm{E}_{\theta_f}(\cdot)$.

Action Space. We adopt Shadow Hand as the manipulator, which is a five-fingered robotic hand with 24 Dofs. Except for the thumb, the distal joints are driven by tendons. Hence, the actions $a = \pi_\theta(s) \in \mathbb{R}^{20}$.

IV. Experiment Results

Our experiment aims to validate the effectiveness of our proposed visual-tactile fusion method M^{2}VTP in assisting downstream robotic manipulation tasks. To achieve this, we design several experiments to ask the following questions:

1) Is our method more effective than non-pretrained methods?
2) Does our method offer greater assistance for manipulation tasks compared to single-modal pretraining?
3) Why is our method incorporating tactile information more effective? How can this phenomenon be elucidated?

A. Experiment Settings

Dataset. We collected 20 different bottles and used our visual tactile acquisition system to collect 120 videos of opening and closing different bottle caps, varying from 4s to 20s. The length of all the videos is 30087 frames. Each frame of the image corresponds to tactile data, as shown in Fig.2. Before training M^{2}VTP, the images are cropped in the center to the size of 224*224. Every tactile data is transformed into a binary signal, using the threshold of 0.3V. The dataset supplied to the training stage is sequences of 224*224 images and 20-dimensional binary signals of touch status.

Metrics. To validate the quality of our method, we define a metric ***Success Rate***. When the robotic hand screws the bottle cap more than half a turn, we consider the manipulation successful. We evaluate this metric on 10 seen bottles 200 times and 5 unseen bottles 100 times.

Implementation Details. We train M^{2}VTP on the device with Intel Xeon Gold 6326 and NVIDIA 3090. During the pretraining phase, we use an AdamW optimizer with a learning rate of 2e-5 and weight decay of 0.05. The mask ratio r_v and r_c are predefined as 0.75 and 0.5, respectively. N_v is set to 14×14, implying that the input image is partitioned into a total of 196 patches. It takes about 1.5 hours to optimize all parameters for 400 epochs with a minibatch size of 8. For RL, we run 80 environments in parallel and train the policy network a total of 600 times with the PPO algorithm. It takes about 4 hours to get the final manipulation policy network.

B. Baselines

To validate the effectiveness of our approach, we designed several baselines for comparison as follows:

1) **VT-Scr-C.** This baseline utilizes a CNN [36] network to extract image features from scratch and subsequently concatenates these image features with tactile features before inputting them into the policy network.
2) **VT-Scr-R.** Similar with **VT-Scr-C**, the baseline uses a ResNet18 [37] network instead of CNN.
3) **V-Only.** In most previous work, vision is the sole modality perceivable by the agent. This baseline employs a pre-trained model in Voltron [12] and uses our dataset to fine-tune it without text.
4) **T-Only.** The agent relies solely on tactile perception as its sensory modality. It utilizes a network architecture identical to that of the **V-Only** model and is trained from scratch with our dataset.
5) **MVP [9].** We directly apply the pre-trained MVP model to our RL framework without finetuning.
6) **VT-Sep.** This baseline is a combination of **V-Only** and **T-Only**. We train the vision model and the tactile model with our dataset respectively and then feed the concatenated features into the policy network.

C. Effectiveness of pretraining

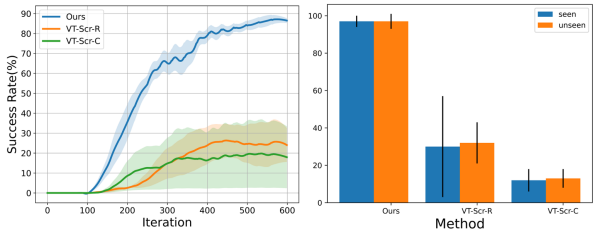

Fig. 4: Qualitative results of pretraining and non-pretraining methods. The left is the training process; the right is the evaluation results

To evaluate the effectiveness of M^{2}VTP, we compare our method with no-pretrained models **(VT-Scr-C and VT-Scr-R)**. Fig 4 (left) shows the training process of our method and baselines. As shown in the figure, it is difficult to train vision and tactile extractors from scratch. The model without pretraining exhibits significantly lower success rates compared to our method, highlighting the substantial assistance our

pretraining approach can provide for downstream reinforcement learning tasks. The evaluation results are shown in Fig. 4 (right).

D. Effectiveness of visual-tactile fusion

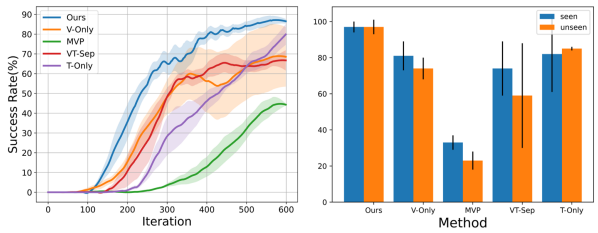

Fig. 5: Qualitative results of different modality methods. The left is the training process; the right is the evaluation results

To investigate which pretraining network extracts more effective information, we compared methods employing different modalities. The results are shown in Fig.5. Our approach facilitates the fusion of visual and tactile information, and the complementary nature of this information enhances the learning speed and effectiveness of downstream tasks. In comparison to methods based on other modalities, our approach exhibits a minimum improvement of 10%. Specifically, it surpasses the commonly used visual pretraining models by more than 50%. Given the presence of visual occlusion in our task, methods relying solely on the visual modality perform poorly. Furthermore, the method using only the tactile modality demonstrates strong performance, highlighting the substantial role tactile information plays in situations where visual occlusion is present.

E. Ablation Study

1) ***Effectiveness of # tactile patches****:* In our approach, we treat each tactile unit as a patch and input it into the Transformer encoder. In this part, we alter the way tactile segmentation is performed, dividing tactile data based on individual fingers **(*v2*)** and treating all tactile data as a single entity **(*v1*)**. The experimental results are presented in Table I. Setting the tactile patch size to 20 proves to be more beneficial to visual-tactile fusion, resulting in an improvement of over 15% in downstream tasks.

2) ***Effectiveness of different reconstructor****:* In our method, we reconstruct images and tactile data. We modify the reconstruction objective and compare our method with solely reconstructing visual data **(*v3*)** and solely reconstructing tactile data **(*v4*)**. The experimental results are presented in Table I. Our method has led to a success rate increase of over 25%. Particularly, the poorest performance was observed when reconstructing tactile information alone, which could be attributed to the sparsity of collected tactile data, not providing sufficient information to facilitate visual-tactile fusion.

3) ***Effectiveness of tactile position encoding****:* Our method adds 2D positional encoding to image patches and 1D positional encoding to tactile patches. In this part, we delete tactile position encoding **(*v5*)**. The result is shown in Table I. The experimental results indicate that the inclusion of position encoding can enhance the operation success rate by more than 20%. This suggests that incorporating position encoding enhances the utility of binarized tactile data, enabling patches to focus on more critical information within the data.

TABLE I: Success Rate (%) of different settings.

	ReconImg	ReconTac	PE	TacPatch	Seen	Unseen
ours	✓	✓	✓	20	**97±3**	**97±4**
v1	✓	✓	✓	1	81±3	80±9
v2	✓	✓	✓	5	78±19	86±8
v3	✓		✓	20	70±8	78±21
v4		✓	✓	20	56±9	62±16
v5	✓	✓		20	73±27	75±27

F. visualize the fusion between vision and tactile

We visualize the attention maps during the model's inference process to gain insights into the design of the fusion learning. Each attention map represents the contribution of N_v image patches when reconstructing a specific tactile patch. We compare models with and without the tactile modality input and visualize the attention maps for the thumb, index finger, and little fingertip tactile taxels. The final results are shown in Figure 6. It can be seen that the tactile input guides the network to attend to more features extracted from regions near the taxels to be reconstructed (more reddish color near the taxels), thus demonstrating the cross-modal fusion capability of our method.

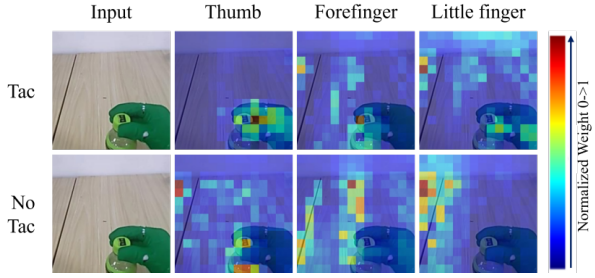

Fig. 6: The visualization of attention maps.

V. Discussion

In this paper, we explore the potential of using human visual and tactile data to pretrain representation models for robot manipulation. We establish a low-cost visual-tactile data acquisition system to collect human visual-tactile manipulation datasets and propose a representation model M^{2}VTP to learn latent representations. We integrate the pretrained representation model into an RL framework for robot manipulation. Experimental results demonstrate the effectiveness of our approach compared with baselines. The paper only makes an initial attempt to leverage human manipulation visual and tactile data for multi-modal robotic manipulation pretraining, more future work includes: 1) a more flexible and high-quality collection system with wireless [38], [39]; 2) a large-scale multimodality human manipulation dataset with a broader range of scenarios; 3) a multimodality fusion model, not limited to vision and tactile; 4) a platform with more visual-tactile tasks.

REFERENCES

[1] D. Sharma, K. Tokas, A. Puri, and K. Sharda, "Shadow hand," *Journal of Advance Research in Applied Science (ISSN: 2208-2352)*, vol. 1, no. 1, p. 04–07, Jan. 2014. [Online]. Available: https://nnpub.org/index.php/AS/article/view/692

[2] J. Pari, N. M. Shafiullah, S. P. Arunachalam, and L. Pinto, "The surprising effectiveness of representation learning for visual imitation," *arXiv preprint arXiv:2112.01511*, 2021.

[3] S. Nair, S. Savarese, and C. Finn, "Goal-aware prediction: Learning to model what matters," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7207–7219.

[4] M. Hong, K. Lee, M. Kang, W. Jung, and S. Oh, "Dynamics-aware metric embedding: Metric learning in a latent space for visual planning," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3388–3395, 2022.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[8] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=tGbpgz6yOrI

[9] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *Conference on Robot Learning*. PMLR, 2023, pp. 416–426.

[10] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "VIP: Towards universal visual reward and representation via value-implicit pre-training," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=YJ7o2wetJ2

[11] Y. J. Ma, W. Liang, V. Som, V. Kumar, A. Zhang, O. Bastani, and D. Jayaraman, "Liv: Language-image representations and rewards for robotic control," *International Conference on Machine Learning (ICML)*, 2023.

[12] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang, "Language-driven representation learning for robotics," *Robotics: Science and Systems (RSS)*, 2023.

[13] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 995–19 012.

[14] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 720–736.

[15] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850.

[16] S. Sundaram, P. Kellnhofer, Y. Li, J.-Y. Zhu, A. Torralba, and W. Matusik, "Learning the signatures of the human grasp using a scalable tactile glove," *Nature*, vol. 569, no. 7758, pp. 698–702, 2019.

[17] M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Learning multimodal representations for contact-rich tasks," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 582–596, 2020.

[18] Y. Chen, M. V. der Merwe, A. Sipos, and N. Fazeli, "Visuo-tactile transformers for manipulation," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=JqqSTgdQ85F

[19] H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. Adelson, L. Fei-Fei, R. Gao, and J. Wu, "See, hear, and feel: Smart sensory fusion for robotic manipulation," in *Proceedings of The 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. PMLR, 14–18 Dec 2023, pp. 1368–1378.

[20] R. Gao, Y.-Y. Chang, S. Mall, L. Fei-Fei, and J. Wu, "Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations," in *Proceedings of the 5th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 08–11 Nov 2022, pp. 466–476.

[21] R. Gao, Y. Dou, H. Li, T. Agarwal, J. Bohg, Y. Li, L. Fei-Fei, and J. Wu, "The objectfolder benchmark: Multisensory learning with neural and real objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 276–17 286.

[22] I. Guzey, B. Evans, S. Chintala, and L. Pinto, "Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play," *arXiv preprint arXiv:2303.12076*, 2023.

[23] D. Ungureanu, F. Bogo, S. Galliani, P. Sama, X. Duan, C. Meekhof, J. Stühmer, T. J. Cashman, B. Tekin, J. L. Schönberger *et al.*, "Hololens 2 research mode as a tool for computer vision research," *arXiv preprint arXiv:2008.11239*, 2020.

[24] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[25] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[26] X. Lin, J. So, S. Mahalingam, F. Liu, and P. Abbeel, "Spawnnet: Learning generalizable visuomotor skills from pre-trained networks," *arXiv preprint arXiv:2307.03567*, 2023.

[27] Y. Seo, K. Lee, S. L. James, and P. Abbeel, "Reinforcement learning with action-free pre-training from videos," in *International Conference on Machine Learning*. PMLR, 2022, pp. 19 561–19 579.

[28] T. Zhang, Y. Hu, H. Cui, H. Zhao, and Y. Gao, "A universal semantic-geometric representation for robotic manipulation," *arXiv preprint arXiv:2306.10474*, 2023.

[29] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[30] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1385–1401, 2021.

[31] L. Yang, B. Huang, Q. Li, Y.-Y. Tsai, W. W. Lee, C. Song, and J. Pan, "Tacgnn: Learning tactile-based in-hand manipulation with a blind robot using hierarchical graph neural network," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3605–3612, 2023.

[32] Z.-H. Yin, B. Huang, Y. Qin, Q. Chen, and X. Wang, "Rotating without seeing: Towards in-hand dexterity through touch," *arXiv preprint arXiv:2303.10880*, 2023.

[33] H. Li, X. Lin, Y. Zhou, X. Li, Y. Huo, J. Chen, and Q. Ye, "Contact2grasp: 3d grasp synthesis via hand-object contact constraint," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 1053–1061.

[34] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.

[35] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[38] Y. Shu, C. Bo, G. Shen, C. Zhao, L. Li, and F. Zhao, "Magicol: Indoor Localization Using Pervasive Magnetic Field and Opportunistic WiFi Sensing," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 7, pp. 1443–1457, July 2015.

[39] Y. Shu, Y. Huang, J. Zhang, P. Coué, P. Cheng, J. Chen, and K. G. Shin, "Gradient-based Fingerprinting for Indoor Localization and Tracking," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 4, pp. 2424–2433, 2016.