ViTaS: Visual Tactile Soft Fusion Contrastive Learning for Reinforcement Learning*

Yufeng Tian[†1], Shuiqi Cheng[†2], Tianming Wei[3], Tianxing Zhou[3], Yuanhang Zhang[5], Zixian Liu[3], Qianwei Han[4], Zhecheng Yuan[3,4], and Huazhe Xu[3,4]

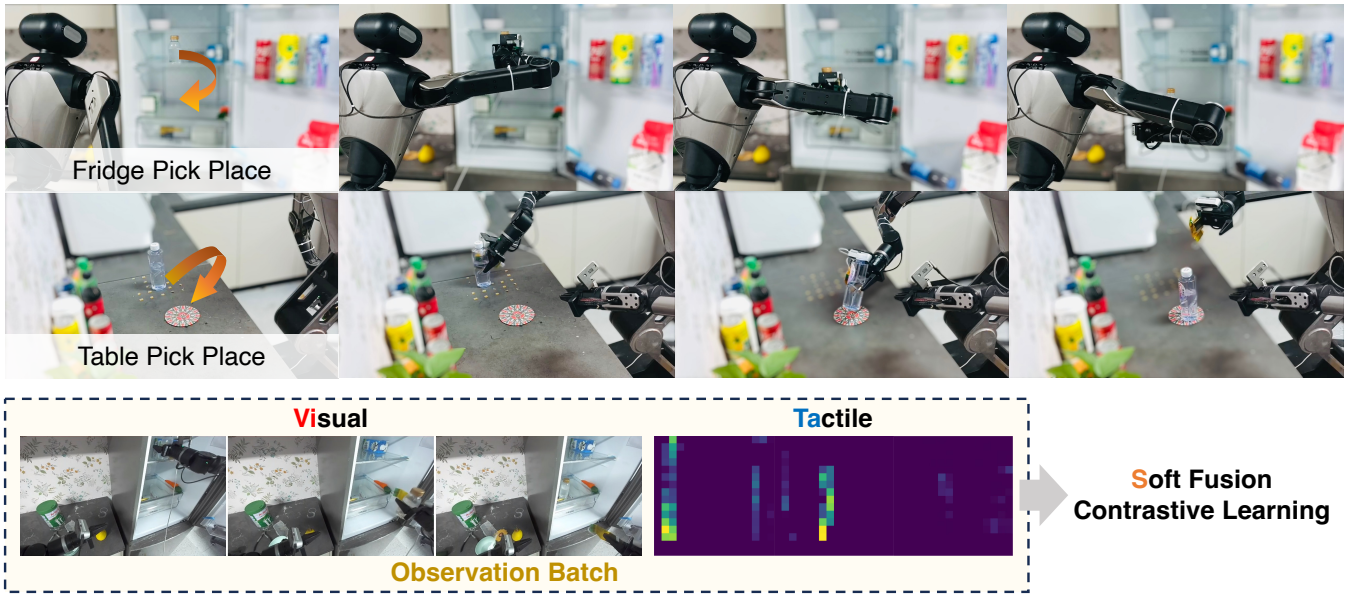

Fig. 1: **ViTaS** is capable of handling various real-world manipulation tasks, including transparent objects and self-occluded scenarios, by fusing visual and tactile observations into a unified representation for effective policy learning.

***Abstract*— Tactile information plays a crucial role in human manipulation tasks and has recently garnered increasing attention in robotic manipulation. However, existing approaches struggle to effectively integrate visual and tactile information, resulting in sub-optimal performance. In this paper, we present ViTaS, a simple yet effective framework that incorporates both visual and tactile information to guide the behavior of an agent. We introduce *Soft Fusion Contrastive Learning*, an advanced version of conventional contrastive learning method, to enhance the fusion of these two modalities, and adopt a CVAE module to utilize complementary information within visuo-tactile representations. We demonstrate the effectiveness of our method in 9 simulated and 3 real-world environments, and our experiments show that ViTaS significantly outperforms existing baselines. The code will be released upon acceptance.**

I. Introduction

Humans are adept at performing complex manipulation tasks, such as spinning an object or cleaning a table. While vision plays a critical role, other modalities, particularly touch, also provide rich information for these activities. Interestingly, visual and tactile information often exhibit significant relevance and complementarity [1]. For individuals with visual impairments, a clearer mental reconstruction of an original visual image can be achieved by combining a blurred visual perception with tactile information [2].

Most previous reinforcement learning (RL) algorithms have relied primarily on visual information to address manipulation tasks [3]–[12]. Recently, several efforts have aimed to incorporate tactile information to improve the performance of RL algorithms. However, these approaches generally exhibit limited fusion between the two modalities. For instance, [13] directly concatenates visual and tactile inputs and feeds them into MAE, while [14] segments visual and tactile data into patches and uses a transformer to extract representations. As a result, these methods often demonstrate limited performance in contact-rich manipulation tasks that rely heavily on both visual and tactile inputs, such as in-hand rotation. Moreover, many previous methods employ complex encoders like transformers and MAE, which involve intricate architectures with numerous parameters, rendering the process of hyperparameter tuning particularly tricky. Given these limitations, we pose the question: *how can we more effectively fuse visual and tactile information, to*

*The paper is completed during author's internship at Galaxea.
†These two authors contribute equally to this work.
[1]Harbin Institute of Technology. tianrainwind@gmail.com
[2]The University of Hong Kong. chengshuiqi28@gmail.com
[3]Tsinghua University, IIIS. huazhe xu@mail.tsinghua.edu.cn
[4]Shanghai Qi Zhi Institute.
[5]Carnegie Mellon University.

enhance the performance of RL algorithms for manipulation?

Drawing on prior research in human physiology regarding the processing of visuo-tactile information, we propose **Vi**sual **Ta**ctile **S**oft Fusion Contrastive Learning (ViTaS), a novel visuo-tactile representation learning framework for reinforcement learning. Generally, ViTaS can be divided into two parts. Firstly, given the inherent relevance between visual and tactile modalities, we utilize contrastive learning to align the embeddings of visual data with their corresponding tactile information in the latent space. Notably, we employ *soft fusion contrastive learning* inspired by [15] to fuse features in alternated modalities. Specifically, we extend the original RGB single-modality framework to incorporate both visual and tactile modalities, enabling the agent to leverage samples of different timesteps with similar tactile information as positive samples. Secondly, inspired by the ability of humans to reconstruct clear images from blurred visual inputs combined with tactile information and complementarity of two modalities, we integrate conditional variational autoencoder (CVAE) introduced by [16] to reconstruct the original image with the embeddings of vision and touch, further improving the fusion of visual and tactile information.

To evaluate the performance of our algorithm, we conduct both simulated and real-world experiments. In simulation, 9 tasks across 5 environments are introduced: Insertion [13], Gymnasium [17], Robosuite [18], Mobile Catch [19] and Block Spin [20]. Additionally, to demonstrate the generalization capability of our system, we perform further experiments on 3 auxiliary tasks, as well as several ablation studies. The overall experiment results in Table I show that ViTaS achieves state-of-the-art performance compared to other visuo-tactile learning methods in all tasks, with an average success rate of 92% and average improvement of 51%. In real-world settings, we integrate ViTaS into imitation learning paradigm, with an increase of 16% compared to baseline in 3 hard manipulation tasks.

In summary, our contributions are as follows:

- We improve the traditional contrastive learning method and use it for the fusion of visual and tactile modalities.
- We propose ViTaS, a simple yet effective representation learning paradigm that can integrate visual and tactile inputs through soft fusion contrastive as well as CVAE, and utilize it to guide the training of reinforcement learning, imitation learning and visuo-motor agent.
- We evaluate our algorithm on various tasks in both simulation and real-world environment, demonstrating state-of-the-art performance against various baselines.

II. Related Work

a) Visuo-Tactile Representation Learning: In recent years, numerous cross-modal representation learning methods have emerged, particularly those focused on visuo-tactile integration, as demonstrated by [21], [9], [10], [21]–[28]. Among them, [29] utilizes an adversarial loss to learn representation in the latent space, while [14] leverages a transformer architecture to integrate multiple modalities, introducing alignment and contact loss to enhance performance. [13] proposes a jointly visuo-tactile training scheme using an MAE-based encoder trained through a reconstruction process, with the encoder co-trained for policy learning.

Despite the success of these approaches in specific tasks, they often fail to fully exploit the correspondence between visual and tactile modalities, leading to suboptimal encoder training and reduced success rates in tasks such as dexterous hand manipulation. In contrast, our method employs a simpler yet highly effective CNN-based encoder to improve the alignment and fusion of modalities, achieving superior performance across multiple benchmark tasks.

b) Contrastive Learning: Extended into computer vision by the MoCo series [30], [31] and SimCLR [32], contrastive learning has emerged as a prominent technique for representation learning. We intend to extend the contrastive learning paradigm to a visuo-tactile framework for reinforcement learning. Related examples include [33], [23], [34], [35], [36], [15], [24], [33], [20] and [37]. Among the works most closely related to ours, [23] proposes a visuo-tactile fusion approach based on contrastive pre-training. [37] employs contrastive loss within the visual modality to enhance policy learning. [24] incorporates tactile, vision, and text using contrastive learning to solve downstream tasks. [15] advances contrastive learning paradigm by using top K analogous samples in optical flow as positive samples in the RGB modality.

However, as [15] mentions, simply doing instance discrimination tends to neglect some key information since two resembling samples may be negatives for each other due to distinct timesteps. The phenomenon also pops up in the field of cross-modal contrastive learning. We refine the contrastive learning method to alleviate the issue and better integrate different modalities, which is elaborated in Section III-A.

III. Method

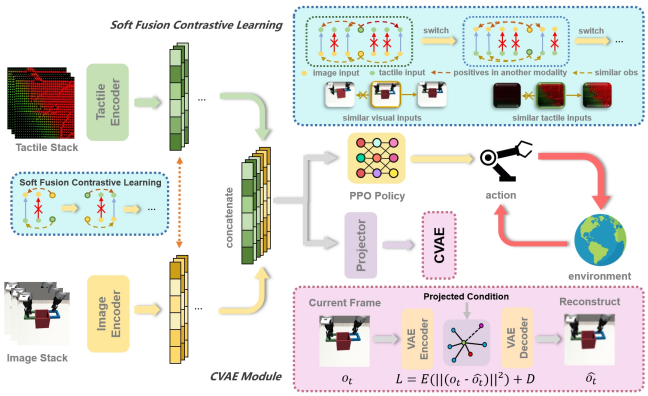

Fig. 2: **Method overview.** The agent takes information from two modalities, visual and tactile, as inputs, which are then processed through separate CNN encoders. Encoded embeddings are utilized by cross-modal soft fusion contrastive approach, yielding fused feature representation for policy network. A CVAE-based reconstruction framework is also applied for cross-modal integration.

In this section, we elaborate **Vi**sual **Ta**ctile **S**oft Fusion Contrastive Learning (ViTaS), an advanced visuo-tactile fusion framework tailored for reinforcement learning. We note that when two tactile maps are resembling, the key features derived from the corresponding visual data should likewise bear a strong resemblance and vice versa. This alignment makes the data particularly well-suited for contrastive learning. Meanwhile, given the complementary information tactile and image offer, our objective is to reconstruct the features extracted from two modalities, obtaining outputs with enriched and more detailed information. CVAE is a good candidate for this process. Therefore, ViTaS fuses visual and tactile modalities through the collaboration of Cross-modal Soft Fusion Contrastive Learning (Section III-A) and CVAE (Section III-B). We utilize the PPO [38], an on-policy reinforcement learning strategy, for the underlying algorithm framework of our method. Formally, our ultimate reinforcement learning objective can be defined as follows:

$$\mathscr{L} = \lambda \mathscr{L}_{\texttt{CON}} + \mu \mathscr{L}_{\texttt{VAE}} + \mathscr{L}_{\texttt{PPO}}, \tag{1}$$

where λ, μ be tunable parameters to bridge the gap between various components.

A. Soft Fusion Contrastive Learning

We denote a trajectory as $\Gamma = \{o_i, t_i\}_{i=1}^{N}$, where o_i stands for image observation at i-th timestep and t_i for tactile inputs, with total length N. For simplicity, we denote o_i and t_i are *dual* samples of each other. We use two convolutional neural networks separately to extract features from raw images and tactile maps. Formally, we denote $f_o(\cdot)$ and $f_t(\cdot)$ as the image and tactile extractors respectively.

Inspired by [15], we present *soft fusion contrastive learning*, a novel cross-modal contrastive learning paradigm to enhance the fusion of two modalities. We use *soft fusion contrastive* below for simplicity. Specifically, we accomplish this by identifying the K most analogous samples from one modality, say modality $\mathscr{A}$, leveraging their *dual* samples as positives for each other in alternated modality $\mathscr{B}$. During the process, the parameters of the encoder corresponding to modality $\mathscr{A}$ are frozen, with the counterpart in $\mathscr{B}$ updated actively. K is a hyper-parameter representing the number of positives needed to be utilized. Then, we reach the following formula in accordance to the description above as Section III.

In the formula, we denote $\mathscr{P}_1(i)$ as the set of positives of o_i, calculated by K most similar samples in corresponding inputs in tactile modality, while $\mathscr{N}_1(i)$ as negatives. S stands for universal samples of o_i in one trajectory. We use $topKmax_k(U)$ to obtain the top K similar samples in set U, which is obtained in replay buffer in implementation. $\mathrm{Sim}(x,y)$ calculate the similarity between key features x and y. We use *cosine similarity* to achieve this and we would like to emphasize that we discern positives by extracted features, where the encoders walk in.

Similarly, we periodically change the position of t_i and o_i like workflow presented in [26] to update both encoders equally, and the corresponding metrics are then denoted $\mathscr{L}_{\texttt{CON,2,i}}$ and $\mathscr{P}_2(i)$. Moreover, we replace all $o_i, f_o(\cdot)$ in the above formula with $t_i, f_t(\cdot)$ and vice versa.

To achieve a more balanced update of the target, we adopt alternating updates when calculating ultimate objective $\mathscr{L}_{\texttt{CON}}$ according to $\mathscr{L}_{\texttt{CON,1/2,i}}$. Specifically, $\mathscr{L}_{\texttt{CON}}$ is contributed by $\mathscr{L}_{\texttt{CON,1,i}}$ at start, and shifted to $\mathscr{L}_{\texttt{CON,2,i}}$ after exact $T_{\texttt{switch}}$ steps and so forth. Furthermore, we define the coefficient sequence as $u_i = 1/2 \times \left(1-(-1)^{\lceil i/T \rceil}\right) = [1,1,\cdots,1,0,0,\cdots,0,1,1,\cdots]$. Consequently, the target of the contrastive loss can be written as:

$$\mathscr{L}_{\texttt{CON}} = \sum_{i=1}^{N} u_i \cdot \mathscr{L}_{\texttt{CON,1,i}} + (1-u_i) \cdot \mathscr{L}_{\texttt{CON,2,i}} \tag{3}$$

B. Conditional VAE Visuo-Tactile Feature Integration

In the realm of visuo-tactile integration, VAE-based methods are commonly employed [39], [40]. Inspired by [41], we extend the CVAE framework for visuo-tactile fusion by incorporating the *condition* component, which is derived from the projection of image and tactile embedding. Consequently, the image and tactile encoders are optimized concurrently during the training process. A comprehensive depiction is presented in Figure 2.

We establish *condition* on the concatenated visuo-tactile feature c to reconstruct the current image frame o_{cur}. CVAE consists of an encoder $p_\theta(\cdot)$, decoder $q_\psi(\cdot)$, and visuo-tactile embedding projector $f_\phi(\cdot)$, which are parameterized by θ, ψ and ϕ separately. We use z to represent the latent variables, and the reconstructed frame $\hat{o}_{\text{cur}}$ conditioned on visuo-tactile feature c can be expressed as:

$$\hat{\mathbf{o}}_{\text{cur}} = q_\psi(p_\theta(o_{\text{cur}}, f_\phi(c)), f_\phi(c)) \tag{4}$$

In accordance with CVAE constraints, the target can be formulated as:

$$\mathscr{L}_{\texttt{VAE}} = \mathbb{E}\left[\|o_{\text{cur}} - \hat{o}_{\text{cur}}\|^2\right] + D_{\text{KL}}\left(p_\theta(z|p_\theta(o_{\text{cur}}), c)\| \ \mathscr{N}(0,1)\right) \tag{5}$$

Notably, the CVAE module is active only during training and does not impose any additional computational overhead during test time.

IV. Experiments

We evaluate our method on several contact-rich tasks in both simulation and real-world environments, in order to clarify the following questions:

(i) Does ViTaS have the capability to solve complicate manipulation tasks requiring compact tactile information (e.g. dexterous hand rotation)?
(ii) How does ViTaS demonstrate generalization and robustness in tasks involving objects of various shapes, significant noise or different physical parameters?
(iii) How does ViTaS perform in real-world settings?

All three questions will be elaborated in the following parts. Moreover, ablation and qualitative study are also done for the understanding of components in ViTaS.

$$\begin{cases} \mathscr{L}_{\texttt{CON,1,i}} = -\mathbb{E}\left[\log \dfrac{\sum_{p\in\mathscr{P}_1(i)} \exp(f_o(o_p)\cdot f_o(o_i)\ /\ \tau)}{\sum_{p\in\mathscr{P}_1(i)} \exp(f_o(o_p)\cdot f_o(o_i)\ /\ \tau) + \sum_{n\in\mathscr{N}_1(i)} \exp(f_o(o_n)\cdot f_o(o_i)\ /\ \tau)}\right] \\ \text{s.t. } \mathscr{P}_1(i) = \{j | (\text{Sim}(f_t(t_j), f_t(t_i))) \in topKmax_k(\text{Sim}(f_t(t_i), f_t(t_k)))\}, \quad \mathscr{N}_1(i) = S \setminus \mathscr{P}_1(i) \end{cases} \tag{2}$$

A. Simulation Environment Setup

1) Tasks: We conduct experiments using 9 simulated tasks, categorized into 5 primary parts shown in Figure 3 (a) to (e): (a). shadow dexterous hand tasks [42], [43] based on Gymnasium [17] (pen rotation, block rotation, and egg rotation), (b). Robosuite [18]-based tasks (door opening, lift, and dual arm lift), (c). Insertion tasks originated by [13] simulated in mujoco, (d). Mobile-Catch environment implemented by [19] and (e). Block Spinning task created by [20]. Beyond these foundational experiments, we introduce a series of auxiliary tasks involving altering object shapes in Lift or modifying physical parameters in Pen Rotation. The outcomes of (a)-(d) environments are quantified in terms of success rate, and (e) is assessed based on training reward.

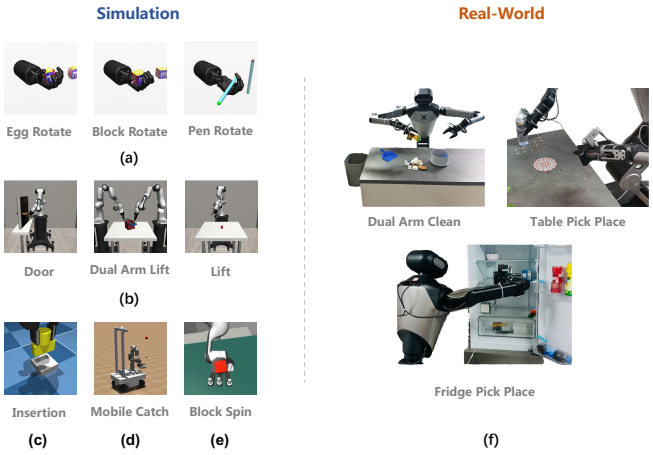

Fig. 3: **Tasks.** Our method is evaluated on 9 simulaion tasks and 3 real-world tasks, with various embodiment types.

2) Tactile sensors: It is crucial to integrate tactile sensors to obtain tactile data for ViTaS framework. For the 3 in-hand rotation tasks, we employ the built-in tactile modules. For Lift, Insertion, and Door Opening tasks, we employ a parallel gripper equipped with a $32 \times 32 \times 3$ tactile sensor at the contact surfaces between the gripper and the object. Among the 3 channels in tactile map, channel 1 and 2 represent the normal force and the value of channel 3 denotes shear force, following [44] and [10]. In the catch and block spin tasks, we enhanced the Allegro hand and Leap hand with tactile sensors by integrating four $3 \times 3 \times 3$ sensors on each finger (located at the proximal, middle, distal, and tip segments) and one $3 \times 3 \times 3$ sensor on the palm. These sensors are zero-padded to form a $32 \times 32 \times 3$ input, following [13].

3) Comparison methods: We compare ViTaS against 4 visuo-tactile representation learning baselines:

- M3L [13]: A visuo-tactile fusion training algorithm utilizing the MAE encoder for PPO policy learning.
- VTT [14]: A visuo-tactile fusion training method rooted in the transformer architecture with both image and tactile data segmented into patches.
- PoE [22]: A VAE-like framework to fuse two modality.
- Concatenation [21]: A multi-modal fusion method with contrastive method used to help training.

B. Simulation Experiment Results

Our algorithm is compared against 4 baseline methods across the aforementioned 9 primitive tasks. We evaluate each algorithm in each environment 5 times under different random seeds, and average the results when training 3×10^6 timesteps to obtain the performance metrics.

As the results shown in Table I, several baselines show excellent performance in simple tasks like Door Opening and Insertion. However, for tough tasks like Egg Rotation and Block Rotation, which are contact-rich and require methods to incorporate visual and tactile information jointly, few baselines can solve it within a limited horizon, while ViTaS maintains its performance. **This underscores its exceptional capability to extract features and solve complicate tasks, clarifying question (i).**

TABLE I: **Benchmark Performance.** Each experiment repeats 5 times. Green for optimal results while purple for suboptimal.

Tasks	Steps	ViTaS	M3L	PoE	VTT	Concat
Insertion	2M	**98**	72	11	78	19
Door	1M	**100**	**100**	98	99	**100**
Lift	1M	**97**	20	71	70	76
Pen Rotate	3M	**100**	73	0	0	2
Dual Arm Lift	1.2M	**100**	88	92	77	76
Mobile Catch	3M	**64**	15	0	53	0
Egg Rotate	3M	**85**	4	0	0	0
Block Rotate	3M	**93**	11	0	1	4
Block Spin	3M	**70**	30	20	0	15
Insertion w/ Noise	3M	**89**	47	20	63	26
Lift w/ Cap	3M	**99**	54	58	54	87
Lift w/ Can	3M	**97**	41	52	69	75
Average	-	**92.9**	47.7	36.5	50.3	42.3

In order to assess the generalization capability and robustness of our approach, we introduce auxiliary tasks derived from the Lift and Pen Rotate tasks mentioned earlier. For the Lift task, the object shape is modified from a cube to cylinder and capsule in both training and testing phases, allowing us to evaluate the method's resilience to changes in object geometry. As for the Pen Rotate task, we randomize the

TABLE II: **Generalization ability.**

Tasks / Method	ViTaS	M3L
Fixed	**99.2**	73.1
Random	**78.4**	42.7
Drop	**20.8**	30.4

target angle within a large range, enabling a thorough evaluation of the model's generalization across varying conditions. We also add Gaussian noise with 0.3 standard deviation in Insertion task (The intensity of Gaussian noise of different standard deviation could refer to Figure 6. The experimental parameters are kept in alignment with the preceding 9 tasks.

As illustrated in Table II, when the object shape is changed, every baseline model experiences a performance drop when training 3×10^6 timesteps, indicating sensitivity to these alterations. ViTaS, however, exhibits a negligible decrease, demonstrating its resilience to variations in object geometry. Pen rotation, among the most challenging tasks, is only successfully handled by ViTaS and M3L. When the target angle is randomized, the agent is required to extract the most significant information from the current observations via visual tactile representation learning to make appropriate moves. In this scenario, M3L struggles to maintain performance, while ViTaS continues to solve the task effectively with less drop than M3L as shown in Table II. Given the robust performance in 3 auxiliary tasks, **we provide a clear clarification of question (ii)**.

C. Real-World Experiment

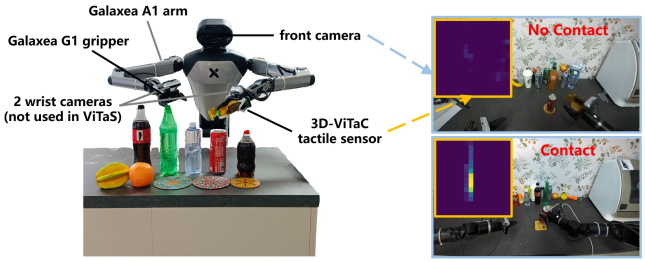

Fig. 4: **Real-World Robot Setting.**

1) Tasks: To better understand the overall performance of ViTaS, we develop 3 real-world experiments to show the effectiveness, shown in Figure 3 (f): (1). Dual Arm Clean (DAC). The robot has to sweep a small amount of rubbish (e.g. a piece of paper ball) to the trash can. (2). Table Pick Place (TPP). The robot has to move the bottles or cans to the coaster. This task has two settings: one (TPP-1) using a single type of bottle and the other (TPP-3) using three types. (3). Fridge Pick Place (FPP). The robot has to move the bottles from third level to the second of refrigerator.

2) Experiment setup: The overall working space is shown in Fig 4. We use Galaxea-R1 Humanoid Robot for manipulation, with tactile sensors attached to the end effector.

- Camera: we use camera for RGB visual information. **Only** the head camera of Galaxea-R1 (Zed 2) is used in all ViTaS experiments, while 2 wrist cameras (RealSense D435i) are needed for better performance in *baseline* experiments.
- Tactile sensors: We use sensors mentioned in [45] for tactile information, which produce real-time $16 \times 16 \times 1$ 1D haptic maps. The tactile sensors is attached to the gripper, obtaining tactile maps during data collecting and inference. A snapshot is shown in Figure 4.

Notably, the real-world setting is slightly different to that in simulation, so some minor changes have been made to offer better empirical results in real world. To exploit the 1D haptic map of our tactile sensor, the tactile encoder is adjusted for appropriate shape. Moreover, we adopt *imitation learning* paradigm [46] to further test the capability of extracting and integrating feature of ViTaS. Specifically, we replace the encoding process with that in ViTaS, and we collect 50 real-world expert trajectories for the training of each task. Both ViTaS and DP are trained for 10^3 epochs, and then transfer to real-world humanoid to calculate success rate.

3) Comparing method: We compare ViTaS with Diffusion Policy [46], a pioneering generative approach for robotic manipulation that formulates action sequence prediction as a conditional denoising process. This method employs a time-series diffusion model to progressively refine Gaussian noise into optimal actions conditioned on image observations, particularly effective in high-dimensional continuous control scenarios. We shall to emphasize that the input of ViTaS is sole head camera and tactile sensor, while *head, left, right* cameras for DP. Both methods adopt *learning-from-scratch* CNN encoders. The motivation is to show ViTaS owns better performance *even* possessing less information, further clarifying the merit of ViTaS.

4) Results: We have done Table Pick Place and Fridge Pick Place with 25 repetition, with the target position shifted slightly each time. We also complete Dual Arm Clean with 10 repetition, putting only *one* piece of litter during benchmarking. The results are shown in table III.

TABLE III: **Real-world experiment results.**

Method / Tasks	DAC	TPP-1	TPP-3	FPP	Average
ViTaS	**30.0**	**42.0**	**36.0**	**76.0**	**46.0**
DP	20.0	36.0	24.0	40.0	30.0

It is evident that ViTaS outperforms DP in 3 real-world tasks even using less camera information, with an average success rate increase of about 16%. Moreover, The comparison between TPP-1 and TPP-3 reveals that ViTaS could offer a better generalization since the drop between two tasks of ViTaS is lower than DP. It is noteworthy that there may be occlusion of bottles by robotic arm from head camera view. The results indicate that tactile could help getting over such adverse factors, further showing the capability to integrate different modalities of ViTaS. Complete acting procedure will be provided in the supplementary video. Given the

better performance and generalization ability in real-world experiments, **we reach the answer to question (iii).**

D. Ablation Study

To verify the fidelity of each component in ViTaS, we conduct extensive ablation experiments, showing the necessity of tactile information, soft fusion contrastive, CVAE module and the choice of K. The overall ablation results are presented in Table IV, where we use abbreviations of experiments in the first row, corresponding to **V**iTaS, w/o. **Ta**ctile, w/ **U**nified Encoder, w/o. soft fusion **C**ontrastive module, w/ **T**ime **C**ontrastive, $K = 1$ and $K = 50$. Detailed analysis of each experiment are clarified in the following sections.

TABLE IV: **Ablation study.** Each experiment repeats 5 times. Green for optimal results while purple for suboptimal.

Tasks / Methods	V	w/o. TA	U	w/o. C	TC	K1	K50
Insertion	**99.2**	88.1	61.6	90.3	75.2	83.3	78.7
Block rotation	**92.7**	67.7	18.4	67.7	79.5	88.0	70.1
Egg rotation	**85.3**	24.3	3.3	6.5	57.7	65.2	3.6
Average	**92.5**	60.9	27.1	54.7	70.6	78.8	67.3

Is tactile information crucial? We conduct 2 main experiments in this part. Firstly we eliminate the tactile information, retaining only the visual data, and solely utilize the image encoder, while handling the corresponding tactile information through zero-padding. Additionally, the workflow outlined in [13] employs a unified MAE encoder across both modalities, disregarding the inherent distinctions between them. This oversight could potentially leads to less discriminative feature representations and a notable reduction in overall effectiveness. To prove that tactile maps can offer unique information beyond visual inputs can provide, we build another experiment that image and tactile are directly concatenated and subsequently fed into a shared encoder.

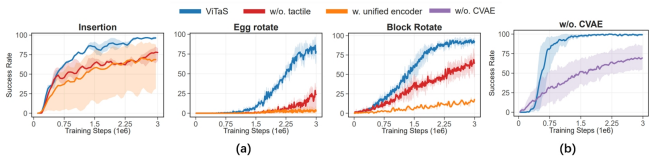

Fig. 5: **Learning curve for tactile and CVAE**. (a) shows results for ViTaS,ViTaS w/o. tactile information and w/ unified encoder, while (b) indicates pen rotation w/o. CVAE.

The results in Figure 5 (a) show that when ablating tactile information, the success rate in 3 benchmarks drops 34% on average. Thus, tactile information gets crucial in dexterous operation tasks like rotation, while it also makes difference in simpler tasks like Insertion.

Using a unified encoder, however, is not a good choice either, given the poor performance in the U-column in Table IV, especially for the 2 in-hand rotation tasks. We then clarify that tactile has some complementary information to image, which cannot be extracted via a unified encoder.

How much do CVAE and soft fusion contrastive contribute to ViTaS? In order to clarify the effectiveness of each component, we remove the CVAE and soft fusion contrastive components separately, conducting independent tests on the same benchmarks and comparing results.

Results of ViTaS without CVAE in pen rotation task are shown in Figure 5 (b). The learning curves show that the performance drops heavily (about 25%) without CVAE, and the training process is rather unstable. Moreover, as shown in Table IV, we remove soft fusion contrastive learning and the results drop for about 28.9%, with a surge in variance.

K in soft fusion contrastive learning. We explore the impact of varying K, for instance, setting it to 1, 10 (ours) and 50, to observe how the results are affected. It is noteworthy that image and tactile at the same timestep are the only positives for each other when $K = 1$, adopting the same process as conventional cross-modal contrastive learning. Therefore, by comparing results between ours and $K = 1$, we can also clarify whether soft fusion contrastive could outperform conventional contrastive learning method.

The last two columns of Table IV show the effectiveness of different K in ViTaS. The results when $K = 1$ show that though conventional contrastive learning can achieve relatively excellent performance, it still has performance gap with our method (i.e. $K = 10$), while too large K value as 50 also causes a drop in performance.

Soft fusion contrastive v.s. time contrastive. To verify the effectiveness of soft contrastive in another perspective, we carry out experiments utilizing an alternative contrastive approach, namely *time contrastive*, to highlight the indispensable role of cross-modal soft fusion contrastive learning. Neighboring frames (i.e., a fixed number of preceding and succeeding frames) are treated as positives in this method, while distant frames serve as negatives, echoing with [47]. The motivation behind this lies in emphasizing that, despite frames within close time intervals often appearing to be similar, it is crucial during the contrastive learning process to identify the K most analogous frames, which may not necessarily be temporally adjacent. This distinction underscores the importance of going beyond mere time contrastive. As shown in table IV, time contrastive learning cannot surpass soft fusion contrastive, which proves the necessity.

In conclusion, our ablation study delves deep into our algorithm to analyze the effectiveness of each component. The results prove that tactile information, soft fusion contrastive learning and CVAE are of high importance, while soft fusion contrastive performs better than other contrastive methods like conventional contrastive and time contrastive. We show the necessity of every design we use.

E. Qualitative Analysis

To effectively demonstrate the impact of CVAE module, we employ weights from ViTaS in the Egg Rotate task for image reconstruction from pure gaussian noise conditioned on visuo-tactile embedding. We compare the performance under varying levels of Gaussian noise added to the observation space (both visual and tactile) against the token-based

MAE method used in M3L.

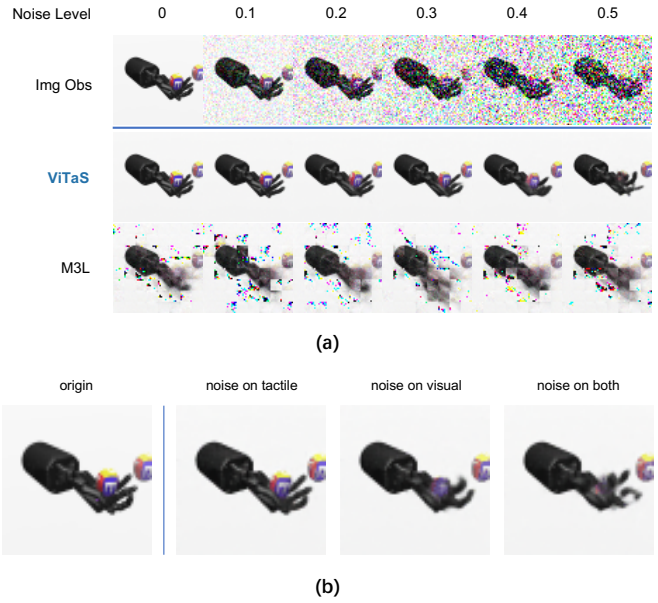

Fig. 6: **Reconstruction visualization** (a) compares the reconstruction quality of ViTaS and M3L under different level of observation noise. (b) shows reconstruction results of ViTaS under heavy noise applied to different modalities

As illustrated in Figure 6 (a), the results indicate that our approach surpasses the token-based MAE in reconstructing critical interaction details, such as finger joint positions and the egg's location, which are vital for the task. Our method also maintain robust under higher level of noise, underscoring the high quality of the visuo-tactile embeddings used as conditions.

Furthermore, we conduct experiments where heavy noise (noise level 0.5) is introduced to either the visual or tactile inputs while keeping the other noise-free. As shown in Figure 6 (b), experiments yield superior generation performance compared to scenarios with heavy noise in both inputs, showing complementary nature of two modalities.

V. Conclusion and Limitations

In general, we introduce ViTaS, a succinct yet effective visuo-tactile fusion framework. Drawing an analogy to human physiology, we extend the application of visual and tactile perception to the domain of reinforcement learning, yielding remarkable results in both simulated and real-world experiments. More specifically, *soft fusion contrastive learning* is proposed to extract key features from one modality according to another, and a CVAE module is developed to utilize complementary information from different modalities. Real-world experiments verify the effectiveness of ViTaS. Ablation and qualitative analysis are meticulously conducted, exhibiting the necessity of each comonent in ViTaS.

Despite its success, ViTaS faces 2 main limitations. The first is that due to the physical capability and the bottleneck of RL in high-dimensional observation, some high dynamic accurate manipulation like pen spinning in real-world remains challenging. The other observation is that the potential of visuo-tactile sensing in deformable object manipulation [48] warrants additional exploration. In the feature, we will further explore the ability of fused visuo-tactile information in more complex scenarios, through synergistic integration with advanced simulation platforms.

References

[1] P. Apkarian-Stielau and J. M. Loomis, "A comparison of tactile and blurred visual form perception," *Perception & Psychophysics*, vol. 18, pp. 362–368, 1975.

[2] A. M. Kappers, "Human perception of shape from touch," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 366, no. 1581, pp. 3106–3114, 2011.

[3] Z. Yuan, Z. Xue, B. Yuan, X. Wang, Y. Wu, Y. Gao, and H. Xu, "Pre-trained image encoder for generalizable visual reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13 022–13 037, 2022.

[4] Z. Yuan, G. Ma, Y. Mu, B. Xia, B. Yuan, X. Wang, P. Luo, and H. Xu, "Don't touch what matters: Task-aware lipschitz data augmentation for visual reinforcement learning," *arXiv preprint arXiv:2202.09982*, 2022.

[5] S. Li, X. Wang, R. Zuo, K. Sun, L. Cui, J. Ding, P. Liu, and Z. Ma, "Robust visual imitation learning with inverse dynamics representations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 12, 2024, pp. 13 609–13 618.

[6] T. Haarnoja, B. Moran, G. Lever, S. H. Huang, D. Tirumala, J. Humplik, M. Wulfmeier, S. Tunyasuvunakool, N. Y. Siegel, R. Hafner, *et al.*, "Learning agile soccer skills for a bipedal robot with deep reinforcement learning," *Science Robotics*, vol. 9, no. 89, p. eadi8022, 2024.

[7] Z. Yu, W. Xu, S. Yao, J. Ren, T. Tang, Y. Li, G. Gu, and C. Lu, "Precise robotic needle-threading with tactile perception and reinforcement learning," in *Conference on Robot Learning*. PMLR, 2023, pp. 3266–3276.

[8] J. Pitz, L. Röstel, L. Sievers, and B. Bäuml, "Dextrous tactile in-hand manipulation using a modular reinforcement learning architecture," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1852–1858.

[9] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, "Learning visuotactile skills with two multifingered hands," *arXiv preprint arXiv:2404.16823*, 2024.

[10] J. Xu, S. Kim, T. Chen, A. R. Garcia, P. Agrawal, W. Matusik, and S. Sueda, "Efficient tactile simulation with differentiability for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 1488–1498.

[11] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik, "General in-hand object rotation with vision and touch," in *Conference on Robot Learning*. PMLR, 2023, pp. 2549–2564.

[12] H. Qi, A. Kumar, R. Calandra, Y. Ma, and J. Malik, "In-hand object rotation via rapid motor adaptation," in *Conference on Robot Learning*. PMLR, 2023, pp. 1722–1732.

[13] C. Sferrazza, Y. Seo, H. Liu, Y. Lee, and P. Abbeel, "The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning," 2023.

[14] Y. Chen, M. Van der Merwe, A. Sipos, and N. Fazeli, "Visuo-tactile transformers for manipulation," in *6th Annual Conference on Robot Learning*, 2022.

[15] T. Han, W. Xie, and A. Zisserman, "Self-supervised co-training for video representation learning," *Advances in neural information processing systems*, vol. 33, pp. 5679–5690, 2020.

[16] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[17] M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. De Cola, T. Deleu, M. Goulão, A. Kallinteris, M. Krimmel, A. KG, *et al.*, "Gymnasium: A standard interface for reinforcement learning environments," *arXiv preprint arXiv:2407.17032*, 2024.

[18] Y. Zhu, J. Wong, A. Mandlekar, R. Martín-Martín, A. Joshi, S. Nasiriany, and Y. Zhu, "robosuite: A modular simulation framework and benchmark for robot learning," *arXiv preprint arXiv:2009.12293*, 2020.

[19] Y. Zhang, T. Liang, Z. Chen, Y. Ze, and H. Xu, "Catch it! learning to catch in flight with mobile dexterous hands," *arXiv preprint arXiv:2409.10319*, 2024.

[20] Z. Yuan, T. Wei, S. Cheng, G. Zhang, Y. Chen, and H. Xu, "Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning," *arXiv preprint arXiv:2407.15815*, 2024.

[21] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *2019 International conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 8943–8950.

[22] 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.

[23] V. Dave, F. Lygerakis, and E. Rueckert, "Multimodal visual-tactile representation learning through self-supervised contrastive pre-training," *arXiv preprint arXiv:2401.12024*, 2024.

[24] F. Yang, C. Feng, Z. Chen, H. Park, D. Wang, Y. Dou, Z. Zeng, X. Chen, R. Gangopadhyay, A. Owens, *et al.*, "Binding touch to everything: Learning unified multimodal tactile representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 340–26 353.

[25] 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," *arXiv preprint arXiv:2212.03858*, 2022.

[26] M. Yang, Y. Lin, A. Church, J. Lloyd, D. Zhang, D. A. Barton, and N. F. Lepora, "Sim-to-real model-based and model-free deep reinforcement learning for tactile pushing," *IEEE Robotics and Automation Letters*, 2023.

[27] S. Li, Z. Wang, C. Wu, X. Li, S. Luo, B. Fang, F. Sun, X.-P. Zhang, and W. Ding, "When vision meets touch: A contemporary review for visuotactile sensors from the signal processing perspective," *arXiv preprint arXiv:2406.12226*, 2024.

[28] S. Li, X. Yin, C. Xia, L. Ye, X. Wang, and B. Liang, "Tata: A universal jamming gripper with high-quality tactile perception and its application to underwater manipulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6151–6157.

[29] Y. Li, J.-Y. Zhu, R. Tedrake, and A. Torralba, "Connecting touch and vision via cross-modal prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 609–10 618.

[30] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[31] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9640–9649.

[32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[33] D. Wang and M. Hu, "Contrastive learning methods for deep reinforcement learning," *IEEE Access*, vol. 11, pp. 97 107–97 117, 2023.

[34] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal contrastive training for visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6995–7004.

[35] M. Laskin, A. Srinivas, and P. Abbeel, "Curl: Contrastive unsupervised representations for reinforcement learning," in *International conference on machine learning*. PMLR, 2020, pp. 5639–5650.

[36] D. Wang and M. Hu, "Contrastive learning methods for deep reinforcement learning," *IEEE Access*, 2023.

[37] A. Zhan, R. Zhao, L. Pinto, P. Abbeel, and M. Laskin, "Learning visual robotic control efficiently with contrastive pre-training and data augmentation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 4040–4047.

[38] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[39] A. V. Nair, V. Pong, M. Dalal, S. Bahl, S. Lin, and S. Levine, "Visual reinforcement learning with imagined goals," *Advances in neural information processing systems*, vol. 31, 2018.

[40] C. Bai, P. Liu, K. Liu, L. Wang, Y. Zhao, L. Han, and Z. Wang, "Variational dynamic for self-supervised exploration in deep reinforcement learning," *IEEE Transactions on neural networks and learning systems*, vol. 34, no. 8, pp. 4776–4790, 2021.

[41] P. Bachhav, M. Todisco, and N. Evans, "Latent representation learning for artificial bandwidth extension using a conditional variational auto-encoder," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7010–7014.

[42] M. Plappert, M. Andrychowicz, A. Ray, B. McGrew, B. Baker, G. Powell, J. Schneider, J. Tobin, M. Chociej, P. Welinder, V. Kumar, and W. Zaremba, "Multi-goal reinforcement learning: Challenging robotics environments and request for research," 2018.

[43] A. Melnik, L. Lach, M. Plappert, T. Korthals, R. Haschke, and H. Ritter, "Using tactile sensing to improve the sample efficiency and performance of deep deterministic policy gradients for simulated in-hand manipulation tasks," *Frontiers in Robotics and AI*, p. 57, 2021.

[44] I. H. Taylor, S. Dong, and A. Rodriguez, "Gelslim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 10 781–10 787.

[45] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, "3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing," *arXiv preprint arXiv:2410.24091*, 2024.

[46] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[47] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1134–1141.

[48] C. Jiang, W. Xu, Y. Li, Z. Yu, L. Wang, X. Hu, Z. Xie, Q. Liu, B. Yang, X. Wang, *et al.*, "Capturing forceful interaction with deformable objects using a deep learning-powered stretchable tactile array," *Nature Communications*, vol. 15, no. 1, p. 9513, 2024.