Learning Robotic Tasks with Object-Centric Value-Implicit Pre-Training

Bo-Ruei Huang
Department of Electrical Engineering
National Taiwan University
b09901171@ntu.edu.tw

Tung-Yu Wu
Department of Electrical Engineering
National Taiwan University
b08901133@ntu.edu.tw

Po-Jung Chou
Graduate Institute of Electrical Engineering
National Taiwan University
r12942088@ntu.edu.tw

Yun-Rong Du
Department of Electrical Engineering
National Taiwan University
b09901011@ntu.edu.tw

Abstract

Offline RL aims to leverage datasets of prior experiences to boost the training process of new tasks, while collecting high-quality data for generalization remains a critical issue. Recent works like VIP and V-PTR show that value function pre-training on Internet videos can achieve generalization on downstream robotic RL tasks. However, they both require a huge amount of robotics videos to overcome the domain gap. In this work, we propose a method with only a few robotic videos needed to fine-tune a value function pre-trained from the human dataset (Ego4D). Specifically, we add object awareness to the value function by fine-tuning it with data augmentation, temporal-cycle consistency loss, and task-specific sampling. Experimental results show that our framework is able to achieve imitation learning on downstream RL tasks with merely a few amounts of data, with the frame feature extractor being more aware of the target object, not only the robot itself. Moreover, it also demonstrates great potential in the cross-embodiment settings.

1 Introduction

In robot learning, the conventional approach demands extensive real-time interactions between the robot and its environment to acquire actions and rewards, incurring significant costs. Recognizing the challenges associated with this method, the paradigm of data-driven robot learning has emerged. This approach involves gathering valuable insights from past experiences before the formal training process begins. However, acquiring training trajectories comprising actions and rewards remains a complex and resource-intensive task.

In recent years, learning from videos has emerged as a promising potential solution to address this bottleneck. Videos stand out as the most accessible and abundant form of data, yet they present a notable drawback: the absence of explicit actions and rewards crucial for training.

Previous works have developed theories and techniques to utilize Internet videos without the need for actions and rewards. Value-Implicit Pre-Training (VIP)[9] transforms the original optimization problem into an actionless dual problem. V-PTR[1] leans the intent-conditioned value function (ICVF)[3] that needs no actions to train. However, they do not tackle the domain gap issues well. VIP, though claimed to be capable of generalizing to robotic scenes, failed to succeed in a large portion of

Preprint. Under review.

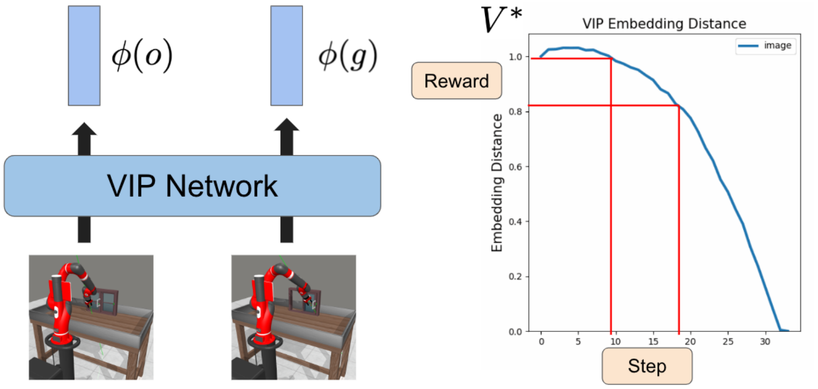


Figure 1: VIP embedding distance as implicit value for reward calculation. The Euclidean distance of the current frame's feature to the goal frame's feature is served to calculate this stage's value for the training agent. A smaller distance to the goal state indicates higher value and thus reward.

downstream tasks, as shown in our preliminary analysis. V-PTR requires large-scale bridge data for phase-2 pre-training, which requires frame-level actions.

In light of this, our research endeavors to bridge this domain gap with a proposed framework built upon the foundations of the VIP model on robotics tasks. We fine-tune VIP with three main modifications: data augmentation, task-specific sampling, and leveraging the Temporal Cycle-Consistency (TCC) loss[2]. Experiments show that our modification can considerably enhance VIP's representation ability on robotics tasks, and the representations, which serve to calculate the observation-to-goal embedding distance, can be used in the downstream robot learning process to realize few-shot imitation learning and cross-embodiment.

Our contributions are twofold:

1. We present a framework for scalable offline RL without labeling effort on the dataset. The presented techniques enable fast adaptation from human scenes to robotic videos.
2. We make the first attempt to incorporate task and object awareness directly from collected videos on value function pre-training, demonstrating superior performance than previous work.

2 Related work

2.1 Value-Implicit Pre-Training (VIP)

Value-Implicit Pre-Training (VIP)[9] uses the visual representation from a ResNet-50[7] as an implicit value function. It defines the implicit value as the Euclidean distance of the current frame's feature to the goal frame's representation vector: $V^*(\phi(o), \phi(g)) := -\|\phi(o) - \phi(g)\|_2$. The loss function can be derived from TD-learning with the implicit value:

$$\begin{aligned}\mathcal{L}(\phi) =& \mathbb{E}_{p(g)}\left[(1-\gamma)\mathbb{E}_{\mu_0(o;g)}\left[\|\phi(o) - \phi(g)\|_2\right]\right. \\ &\left. + \log \mathbb{E}_{(o,o';g)\sim D}\left[\exp\left(\|\phi(o) - \phi(g)\|_2 - \tilde{\delta}_g(o) - \gamma \left\|\phi\left(o'\right) - \phi(g)\right\|_2\right)\right]\right.\end{aligned} \tag{1}$$

The reward can be calculated from the temporal difference of the implicit value between each frame. Since it was trained on human demonstrations (Ego4D dataset[5]), its generalization ability on robotics manipulation tasks remains a question.

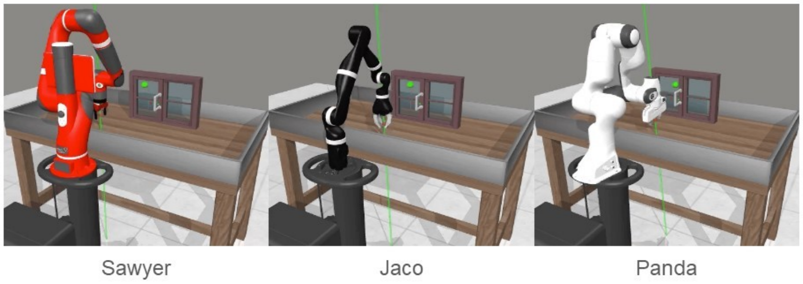
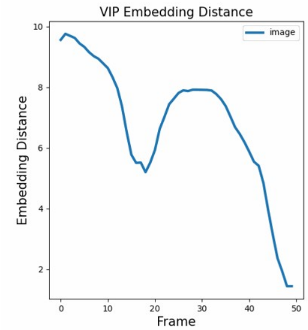


Figure 2: The three robot arms and our environment Meta-World.



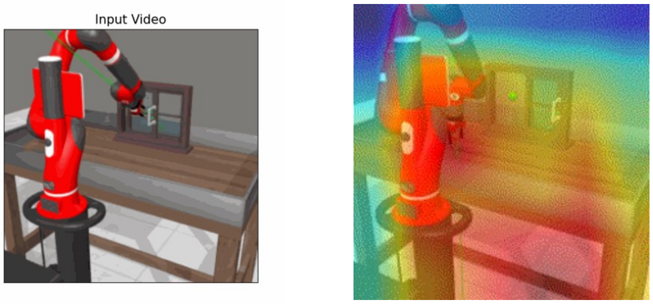


Fig. 3a: Vanilla VIP's embedding distance curve of the expert video on window opening task using Sawyer. Ideally, the curve should be monotonically decreasing.

Fig. 3b: Illustration of Grad-cam on vanilla VIP. The red color denotes a higher gradient response. It can be shown that the vanilla VIP lacks attention on the window.

2.2 Video Pre-Training for Robotic (V-PTR)

Video Pre-Training for Robotic (V-PTR)[1] is the follow-up work of VIP. It improved the ability of generalization on robotics tasks via three phases of training. The first phase pre-trains an intent-conditioned value function (ICVF)[3], a general value function that can be trained on videos without action labels on the Ego4D dataset. The second phase refines the learned representation on a multi-task robot dataset (Bridge dataset) by pre-training a conservative Q-learning (CQL)[8]. The third phase fine-tunes the policy on the target dataset. The second phase requires labeled action for CQL. Data collection still remains a challenge for other robots.

3 Problem Formulation

In our initial proposal, we aim to use the embedding distance from VIP directly as the reward for robotics task policy training.

Meta-World For our training environment, we employ Meta-World[10] and select three robot arms—Sawyer, Jaco, and Panda (Figure 2). Our training dataset comprises videos of the 'Window Open' task executed by these robotic arms. To emulate real-world challenges in robot control, we adjust the arm settings, transitioning from end effector control to torque control.

Evaluation Metrics To assess the viability of our initial proposal, we feed the expert video into VIP and observe its outcome. As depicted in Fig. 3a, the result unveils a non-monotonically decreasing embedding distance curve. This irregular trend implies potential negative rewards during the training process, showing VIP was incompatible with direct application in robotics tasks. The subsequent failure of downstream training further underscores VIP's limitations in this context. We further analyze how VIP interprets the video with Grad-Cam (in Fig. 3b) and find that VIP focuses more on the robot arm than the task object. Thus, we speculate that given VIP's training on unsupervised data, it might lack essential domain-specific information crucial for effective downstream task performance.

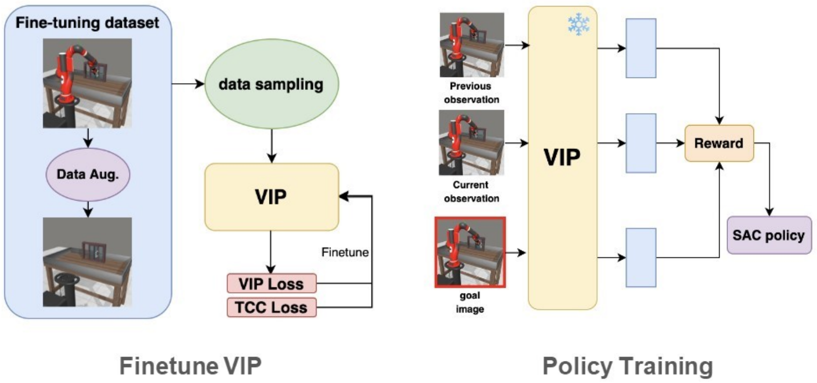


Figure 4: Our proposed framework. During the fine-tuning stage, data augmentation, task-specific data sampling, and TCC loss are applied to guide VIP to better transfer to the domain of robotic tasks. Afterward, we evaluate fine-tuned VIP by leveraging it as the value function to train the policy on downstream tasks.

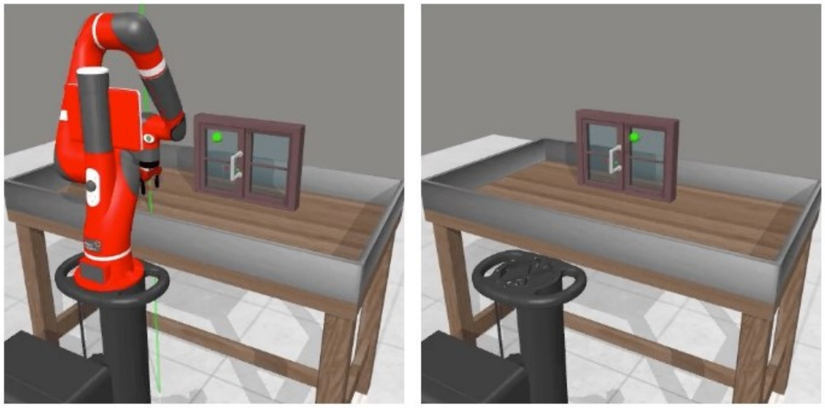


Figure 5: An illustration of a frame before and after data augmentation. For a video in the dataset, we remove the robots in it, with only target objects left, and add the processed video to the original dataset.

From the above results, we work on enhancing VIP's representation ability in robotics tasks so that the embedding distance it produces could be used in further training, ultimately realizing our goal of learning robotic tasks from videos.

4 Method

4.1 Framework Overview

Our framework is shown in Figure 4. It comprises upstream VIP fine-tuning and downstream SAC policy training for robotic tasks. During the fine-tuning stage, three proposed techniques are applied: data augmentation, Sampling Function, and TCC loss. Afterward, following VIP, we evaluate the quality of generated rewards by training a SAC agent on robotic tasks in the Meta-World.

4.2 Data Augmentation

Upon detailed analysis of VIP's performance through Grad-Cam, it becomes evident that the model focuses on the robot arm rather than the task object. To address this observation, we propose to remove the robot arm from the training video, thereby compelling VIP to redirect its attention to the task. This concept led to a modification of the Meta-World environment, as illustrated in Figure 5.

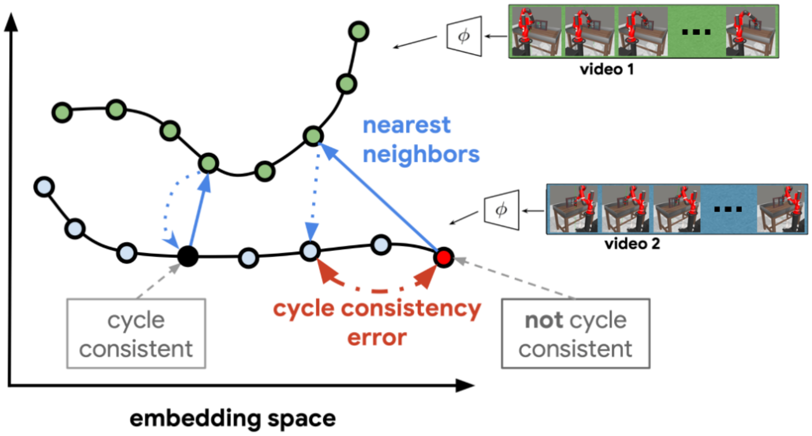


Figure 6: Illustration of TCC loss. Ideally, two videos with the same action trajectories should possess a one-to-one feature mapping regardless of the view of the videos and the color/shape of the robots. A cycle-consistency error arises if such a one-to-one relation is violated in some frames.

4.3 Temporal Cycle-Consistency Loss

During the VIP fine-tuning stage, a task in the dataset may contain videos from different views, robots with various shapes/colors, and even videos with only task objects by our proposed data augmentation. In this case, given any pairs of videos with the same task, their frame features, generated by VIP's ResNet-50, should possess a one-to-one mapping. For example, the feature of a frame from the left showing that Saywer has just touched the window should be close to that from the right view but relatively far from the frame where Saywer has opened the window. Inspired by this, we adopt the TCC[2] loss in the fine-tuning stage. In detail, K cycles will be randomly sampled and gone through within videos of a batch. For each cycle, a set of T videos are chosen in order, say $S = \{v_1, ..., v_T\}$. Then, after randomly selecting an initial frame f_1 in v_1, we jump to the next frame by:

$$f_2 = \arg\max_{f_i^2} ||\phi(f_1) - \phi(f_i^2)||_2, \tag{2}$$

where $\phi(\cdot)$ is the VIP ResNet-50. Finally, we calculate the come-backed frame $\hat{f}_1$ by finding the nearest frame in v_1 with respect to v_T with formula (2). The index difference between f_1 and $\hat{f}_1$ is served as the cycle-consistency error.

4.4 Sampling Strategy

In the original framework of VIP, four frames are sampled randomly in the video as the starting frame, current observation frame, next observation frame, and goal frame. However, this may lead to skewed attention on the robot and target object. For instance, in the window opening task, the window is closed most of the time and opened only at last few frames. This incurs VIP to simply focus on the moving of the robot and treat the window as background since the state of the window does not change for most of the sampled observations. To tackle this issue, we alter the sampling function from randomness to exponential and beta distribution, as shown in Figure 7. This simple technique drives VIP to focus more on the window because samples of the moving window have a higher probability of being chosen.

5 Experiment

5.1 Fine-tune VIP

Initially, we employ an expert policy in tasks involving an open window on the Sawyer robot, capturing video frames at a resolution of 480 x 480 from two distinct viewing angles. Subsequently, we disentangle the Sawyer from its viewpoint by setting the alpha value to 0 in Metaworld, creating

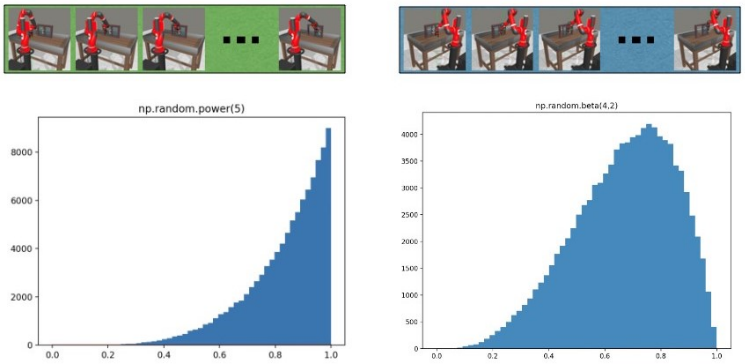


Figure 7: Illustration of exponential distribution $Exp(\alpha)$ and beta distribution $Beta(\alpha, \beta)$. A larger sampling possibility denotes harder samples. The exponential distribution assumes frame difficulty increases with the latter states; the beta distribution assumes the timing where the robot and object interact has the greatest difficulty.

four expert trajectory video frames. Among these, two frames depict visible robot arms, while the remaining frames exclusively feature the target object. For the task-specific sampling, we sample the first of four frames with exponential distribution with a rate parameter of 5. In terms of TCC loss, We set number of cycles $K = 20$ and number of videos per cycle $T = 4$ for all experiments.

We adopt the released VIP pre-trained weights, with ResNet50 as the backbone and linear layers to project image features to a 1024-dimension space. The model is then fine-tuned for 100k steps, with batch size = 4 and learning rate = 10e-5. Other hyper-parameters follow the VIP open-source code.

5.2 Policy Training

In the Meta-World environment, the Sawyer robot is characterized by 7 Degrees of Freedom (DoF). Its action space encompasses 8 dimensions, enabling control over each joint's torque as well as the gripper's torque. The observation space is structured with a shape of 51. To formulate the reward function, we leverage the fine-tuned VIP obtained from the preliminary stage. This reward function calculates the temporal difference in embedding distance between the visual observations at the current timestep and the preceding timestep. Our training strategy involves initializing the Soft Actor-Critic (SAC)[6] policy from scratch, executing training for 2 million timesteps across 9 parallel environments.

5.3 Vanilla VIP

Figure 8 shows that the vanilla VIP can rapidly attain a nearly zero embedding distance. However, this seemingly successful outcome masks a critical issue: the policy converges to a sub-optimal solution. Specifically, the Sawyer robot moves its end effector to the right side of the window and gets stuck there. The Grad-CAM[4] visualization further exhibits that the vanilla VIP fails to consistently focus on the window. This lack of attention hinders the task of guiding the agent to conduct necessary adjustments to alter the window states.

5.4 VIP with TCC Loss

Figure 9 reveals an absence of a consistent downward trend of the embedding distance across the entire trajectory. This characteristic poses a challenge for the agent learning process, particularly in terms of achieving a coherent direction. Grad-CAM further highlights that fine-tuned VIP does not prioritize the window or the Sawyer robot. This apparent lack of focus exacerbates the difficulty in guiding the agent.

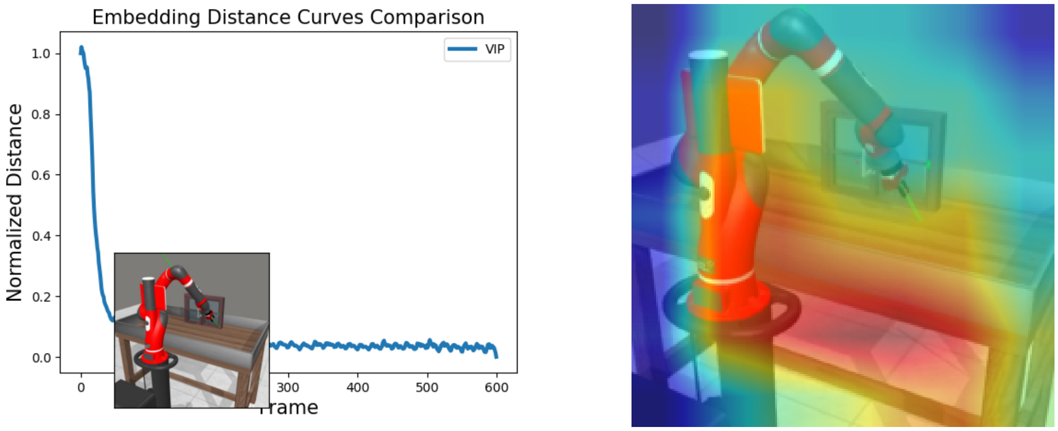


Fig. 8a: VIP curve

Fig. 8b: Grad-cam

Figure 8: VIP fine-tuned on augmented robot arm and object with $Exp(5)$ sampling.

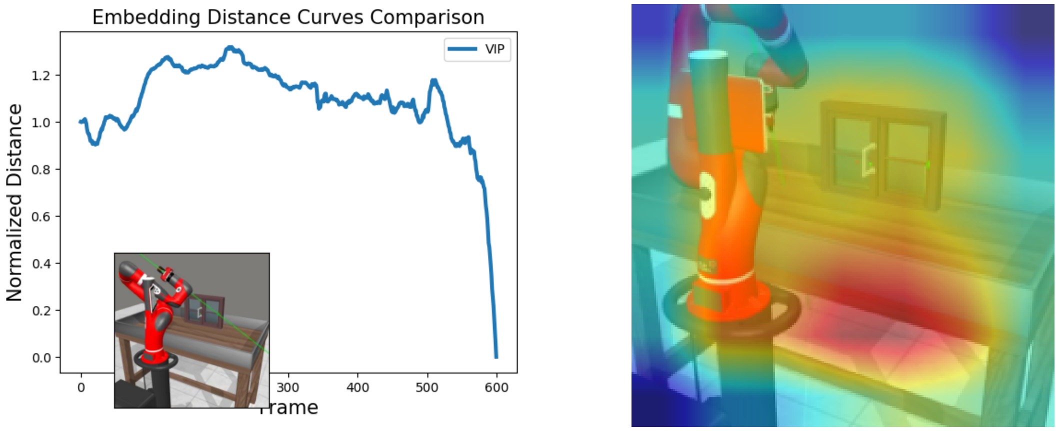


Fig. 9a: VIP curve

Fig. 9b: Grad-cam

Figure 9: VIP fine-tuned with TCC loss and $Exp(5)$ sampling.

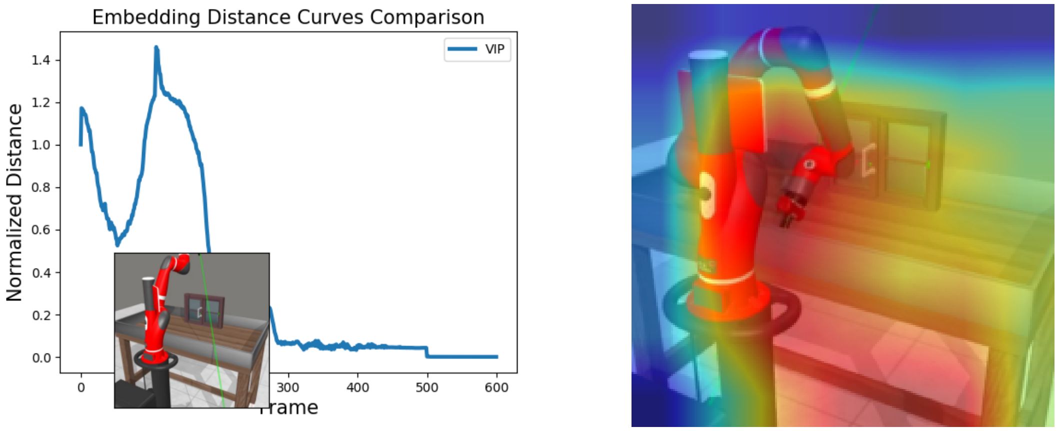


Fig. 10a: VIP curve

Fig. 10b: Grad-cam

Figure 10: VIP fine-tuned with data augmentation and $Exp(5)$ sampling.

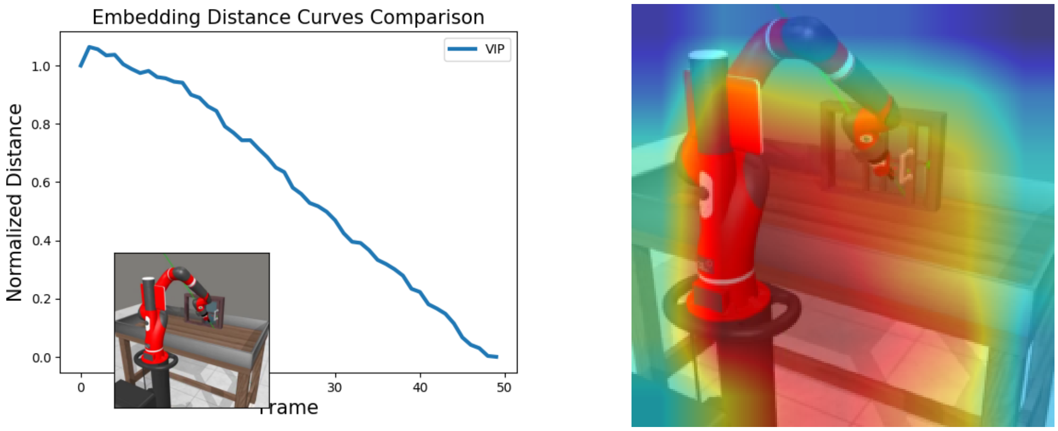


Fig. 11a: VIP curve

Fig. 11b: Grad-cam

Figure 11: VIP finetuned with TCC loss, data augmentation, and $Exp(5)$ sampling.

5.5 VIP with Data Augmentation

Examining Figure 10, despite an initial downward trend in the embedding curve, a subsequent increase in distance is observed. This shows a potential challenge in maintaining the desired trajectory. The Grad-CAM visualization provides additional insights, indicating that the fine-tuned VIP exhibits increased awareness of both the window and the Sawyer robot. However, the absence of TCC loss to align the original and augmented videos poses a notable difficulty. This deficiency makes it challenging for the agent to effectively learn the correct pose of the robot arm while performing the task of opening the window. Addressing this misalignment becomes crucial for refining the policy.

5.6 VIP with TCC Loss, Data Augmentation, and $Exp(5)$ Sampling

With the TCC loss, data augmentation, and exponential sampling all combined, our policy successfully achieves the task of window opening. Figure 11 illustrates a clear and monotonic decrease in embedding distance throughout the expert trajectory, indicating a potentially smoothing training process for the downstream agent. The Grad-CAM heatmap further substantiates this progress, with an enhanced focus on the window. These positive outcomes underscore the effectiveness of our proposed approach in enhancing VIP's ability to guide an agent to perform the targeted manipulation task.

6 Conclusion

Our proposed method demonstrates superior performance compared to the original VIP with evaluations on downstream robot tasks. The success of our approach can be attributed to the incorporation of task awareness into VIP through a combination of data augmentation, an exponential sampling function, and TCC. Specifically, our framework enables robots to learn from videos without explicit actions, rewards, and labels. The infusion of task awareness facilitates our method's success in cross-embodiment scenarios, where robots can learn from the videos of other robots. This breakthrough holds profound implications, as it allows our method to generalize across diverse robotic platforms, paving the way for seamless adaptability to any robot in future applications.

7 Limitation and Future Directions

While our method performs well in simulation, there are still room for improvement before applying it to real world. First of all, we now only apply data augmentation on offline dataset. This limits the generalizability for downstream RL tasks. We plan to apply data augmentation and TCC loss also during online training to make the VIP adapt its object awareness dynamically on downstream RL tasks. Moreover, the sampling strategy now is rule-based prioritised on the last few frames. However, in the real world, robot-object interactions may happen in different frames for different tasks. We plan to design a new prioritised sampling algorithm based on TCC loss and VIP loss, aiming to find key/hard frames automatically. With the both improvements applied, we believe that VIP could capture the most essential movement of the robot arm and the task object.

References

[1] C. Bhateja, D. Guo, D. Ghosh, A. Singh, M. Tomar, Q. Vuong, Y. Chebotar, S. Levine, and A. Kumar. Robotic offline rl from internet videos via value-function pre-training, 2023.

[2] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman. Temporal cycle-consistency learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[3] D. Ghosh, C. A. Bhateja, and S. Levine. Reinforcement learning from passive data via latent intentions. In *International Conference on Machine Learning*, pages 11321–11339. PMLR, 2023.

[4] J. Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

[5] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[7] 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 (CVPR)*, June 2016.

[8] A. Kumar, A. Zhou, G. Tucker, and S. Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

[9] 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. *arXiv preprint arXiv:2210.00030*, 2022.

[10] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2019.