Mask2Act: Predictive Multi-Object Tracking as Video Pre-Training for Robot Manipulation

Junbo Zhang
zhangjb21@mails.tsinghua.edu.cn
Kaisheng Ma*
kaisheng@mail.tsinghua.edu.cn
Tsinghua University

Abstract

Imitation learning from demonstrations is a promising paradigm in robot learning. However, it requires a large amount of robot demonstration data, which is laborious to collect. To address this challenge, recent methods leverage action-free video data to extract skill knowledge and generate latent plans, which contain the predicted future states and guide policy learning. In this paper, we introduce Mask2Act, a novel framework that leverages action-free videos to train a predictive multi-object tracking model to predict future masks of any task-relevant objects in the video as the latent plans. The transition of object masks enriches the latent plans with accurate task-relevant motion knowledge and eliminates distracting information. The latent plans are subsequently utilized to guide policy learning with limited training data. Experiments across 50 manipulation tasks in 2 simulated environments show that our method significantly outperforms other video pre-training methods. Furthermore, Mask2Act inherently guides the model to focus on task-relevant knowledge rather than irrelevant background and distractor information, thus demonstrating superior capability in extracting skill knowledge from cross-task videos and generalizing to new tasks and environments.

1 Introduction

Imitation learning has demonstrated significant potential in robotic tasks. By collecting human demonstrations with observations and robot actions, a policy model can be trained to accomplish various robotic manipulation tasks. However, collecting such demonstrations is laborious, especially for action data [6, 7], which leads to a severe lack of generalization of the policy. Therefore, previous studies have investigated methods to learn from more readily available action-free video data, aiming to extract its transition dynamics knowledge and facilitate policy learning with a minimal amount of robot action data [8, 37, 40].

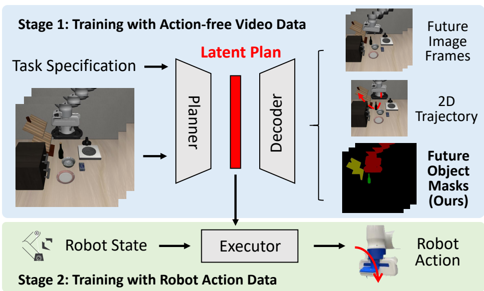

Figure 1: The ***planner*** is trained to generate latent plans with action-free video data. The ***executor*** is trained to produce actions conditioned on the latent plan.

 * Corresponding Author

To better leverage action-free video data, recent studies propose decoupling end-to-end deep imitation learning into two distinct stages: the planning and executing stage [12, 35, 37]. In the first stage, a *planner* is trained with large-scale video data to predict future transition dynamics based on current state and task goal. In the second stage, an *executor* is trained with limited robot action data to produce low-level actions based on generated plans from the *planner*. This paradigm can be abstracted as in Figure 1, where the *planner* generating latent plans is trained with various supervision signals. In this process, the intermediate plan representation serves as the bridge between the image frames and robot actions. Thus, the most critical aspect is: how to enrich the plan representation with a wealth of task-relevant motion knowledge and eliminate distracting information? Some studies propose directly predicting future image frames as plan representations [12, 13, 40]. However, this video prediction task requires reconstructing all image pixels, including texture, lighting, and distractor information, which is unrelated to robot motion. This significantly increases the computational cost for training and inference, often leading to hallucinations that confuse the policy [12, 39]. Other studies propose to encode future trajectories of the embodiment coordinates or key points into plan representations [15, 25, 36, 39]. However, only 2D trajectory information can be extracted from video data, which is visually ambiguous along depth axis and lacks information about the object morphology and motion details [15, 37].

In this paper, we propose a novel intermediate representation that better connects the image frames with the robot's motion in the physical world —— object mask. We formulate the training of the *planner* as a predictive multi-object tracking task, injecting the transition of task-relevant object masks into the plan representation. Specifically, we train the *planner* to take any object mask in the current frame as input and predict the object's mask in the future. As a prerequisite, it is feasible to obtain accurate task-relevant object masks from videos using current multi-modal foundation models, like Grounded-SAM [22, 26] (as shown in Figure 2 and previous work [20, 54]) . Most importantly, modeling object masks as the plan representations has three advantages over previous approaches, as discussed bellow and shown in Figure 2:

- ***Object-Centric Planning***: Object masks encapsulate the information that is task-relevant and motion-relevant within the scene, introducing the inductive bias of object shape/boundary consistency and reducing the impacts of visual distractors. In contrast, video prediction models unnecessarily expend computational resources on irrelevant distractor information and often produce unfaithful object boundaries and hallucination of movements [12, 35, 39].

- ***Rich Motion Information***: The transition of object masks retains both translation and rotation motion information in original image. And the accurate description of the objects' morphology reflects a prosperous state of task completion, facilitating the policy model to decode robot actions. In contrast, 2D trajectories depicted from video contain only the translation information, lack descriptions of object morphology and suffer from the ambiguities of depth dynamics and occlusions [15, 37].

- ***Potential for Generalization***: The transition of task-relevant object masks abstracts high-level planning and object manipulation knowledge, which is vital for skill learning. It excludes the background and appearance differences of different scenes and objects, making generalization easier between different tasks and environments.

With the object masks extracted from the video, we first pre-train the *planner* to track any objects and generate the plan representations that can be decoded into corresponding masks in future frames. Next, we train the *executor* with minimal robot action data guided by predicted object masks. By predicting future object masks as latent plans, the model is guided to focus on task-relevant objects and their morphology rather than visual distractors and background information. Experiment results on 40 tasks from LIBERO [24] and 10 tasks from Meta-World [52] show that Mask2Act significantly outperforms other video pre-training methods, indicating that modeling object masks enriches the plan representation with the most comprehensive visuomotor knowledge. Moreover, experiments of cross-task video pre-training and generalization tests with new tasks and unseen environments demonstrate that Mask2Act is more efficient at learning skill knowledge from cross-task data and beneficial for generalization. Our contributions can be summarized as follows:

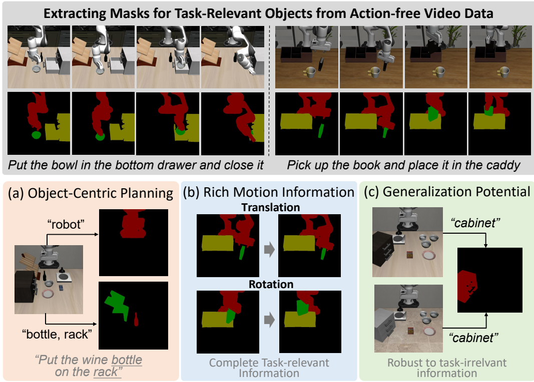

Figure 2: **Top:** Examples of object masks extracted from videos. **Bottom:** Three advantages of modeling object masks as the latent plans.

- We propose modeling object masks as latent plans to connect image frames with robot actions, which contain richer task-relevant object and motion information.
- With extensive experiments across 50 manipulation tasks in 2 simulated environments, we demonstrate that Mask2Act makes a better use of action-free video data than other video pre-training methods.
- By training the model to focus on task-relevant knowledge and neglect distracting information, Mask2Act effectively learns from cross-task videos and generalizes to new tasks and environments.

2 Related Work

2.1 Hierarchical imitation learning.

Imitation learning is effective at learning various policies from robot demonstration data. However, the requirement of a large amount of training data limits its ability to learn general skills. To mitigate this issue, hierarchical training framework proposes first to learn a latent plan from play data [11, 16], and next learn low-level visuomotor control from a small number of teleoperated demonstrations [17, 28]. Sub-goals and robot skills can be extracted in the latent plans to enable transfer and meta-learning [14, 17, 30]. Recent works propose an effective planner-executor structure [3, 12, 35, 37, 39], a two-stage training framework to learn rich latent plans from robot play data. This paper studies how to enrich the latent plans with task-relevant motion knowledge that better promote policy learning.

Prior works learn an inverse dynamics model from video and produce pseudo action labels for policy learning [1, 34]. The effectiveness of such methods is severely constrained by

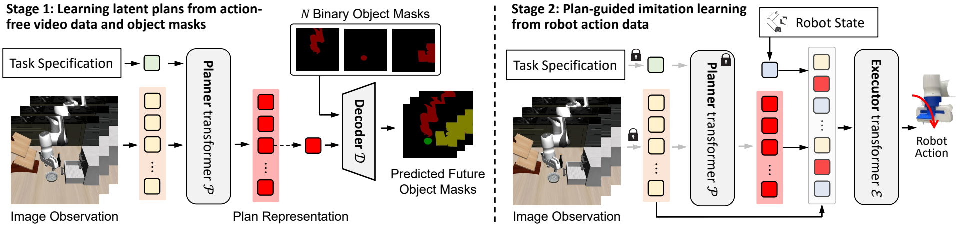

Figure 3: Two-stage training with planner-executor framework and formulating the planner as a multi-object tracking model. **Stage 1:** We train the planner to take any object mask as input and predict the object's mask in future frames as latent plans. **Stage 2:** Conditioned on latent plans, we learn a task-agnostic policy (executor) with robot action data.

the quantity of robot action data, which is laborious to collect. Recent works have demonstrated the great potential of the planner-executor architecture for learning from actionless video data. The planner is trained to predict future latent plans with various supervision signals extracted from video frames. A line of works leverages the video prediction model to reconstruct future image frames as the plan representations [12, 13, 18, 23, 35, 40, 47, 48]. However, video prediction requires extensive computational cost and often leads to hallucinations and unfaithful object shapes. Another line of works predict future flow or trajectory of 2D points as the plan representations [4, 15, 36, 37, 39, 45, 53]. However, 2D point trajectory is visually ambiguous along the depth axis and lacks the information about object morphology and motion details [15, 37]. In this paper, we propose to use object masks as plan representations and formulate the planner as a multi-object tracking model, which better models the object-centric motion knowledge related to the task.

2.2 Generalizable policy learning.

Generalization over new tasks [3, 8], environments [2], object instances [41, 50] and robot morphologies [6, 33, 39] is crucial for the application of robots. Recent works have explored several generalization factors in manipulation tasks [2, 42, 43], such as background, distractors, lighting and camera pose. Recent video pre-training methods [35, 40, 44] also show the capability of extracting knowledge from cross-task videos and generalize to new tasks. By modeling task-relevant object morphology as latent plans and excluding distracting information of background and appearance, our method shows great potential for learning from cross-task videos and generalizing to new tasks and environments.

3 Method

We adopt the widely-used planner-executor structure as in previous works [35, 37, 39]. First, the *planner* predicts future latent plans based on current state and task specification, benefiting from learning through video-only data that are easy to obtain. In our method, we model the transition of task-relevant object masks into the plan representations to enrich a wealth of task-relevant motion information. Second, the *executor* takes the predicted plan representations and produces the robot actions, which can efficiently learn accurate policies from limited robot action data. An overview of Mask2Act framework is shown in Figure 3.

In Mask2Act framework, we formulate the *planner* as a predictive multi-object tracking model: Given any object masks in the current state, the model learns to predict new masks corresponding to each object in the future frames. Note that the number of objects that can be tracked simultaneously is arbitrary in the multi-object tracking framework. We first describe how to extract the task-relevant objects' masks from the pre-training video data. Next, we describe how to train the *planner* as a multi-object tracking model to obtain plan representations enriched with task-relevant motion information. Finally, we describe how to train the *executor* with limited robot action data with the guidance of plan representations.

3.1 Extracting Task-Relevant Object Masks from Video Data

We first introduce the pipeline for extracting object masks from all video frames for pre-training. Specifically, we first used an open-vocabulary detector, Grounding-DINO [27], to obtain bounding boxes and feed them into Segment Anything (SAM) [22] to obtain the robot mask and object masks referenced in the task description for *a single video frame*. We ensure that the task-relevant objects are visible in this annotated video frame, and the robot is always considered an object in the scene. Then, we employ a Video Object Segmentation (VOS) model, XMem [9], to take the single annotated frame and track the corresponding objects in all the other frames, which is proved to be effective in robotic video data [20, 38, 55]. This way, we obtain the 2D segmentation masks for target objects in all video frames. The extracted object mask for frame i is depicted as $M_i \in \mathbb{R}^{H\times W}$. Each element of M_i is an integer ranging from 0 to N, where 0 represents the background, and N represents N target objects. We can further improve the quality of object segmentation and tracking with more advanced models such as S2M [10, 55] and Track-Anything [46]. In practice, we find that the current approach can produce sufficiently accurate masks, as shown in Figure 2. XMem can effectively handle occlusions, and even when the annotated frame contains occlusions or incomplete masks, it can complete the masks during the tracking for other frames.

3.2 Learning Object-Centric Latent Plans By Multi-Object Tracking

The *planner* is designed to learn from action-free video data and predict the latent plan that encodes the future motion information. This plan representation is concretized into different forms in different methods, such as image frames and flow. In this paper, it is trained to decoded into future object masks extracted from the video. Specifically, at time step t, the *planner* $\mathcal{P}$ takes the latest h steps of observations $o_{t-h:t}$ and the task specification g as inputs. It then outputs a sequence of plan representations $p_{t-h:t}$, each of which corresponds to the predicted plans of future states from its current time step:

$$p_{t-h:t} = \mathcal{P}(g, o_{t-h:t}). \tag{1}$$

The predicted plan representations are trained to track the target objects in future frames. Taking the time step t as an example, given a target object mask $M_t \in \mathbb{R}^{H\times W}$ in current time step as introduced above, we first translate it into n binary object masks, $m_t^n \in \mathbb{R}^{2\times H\times W}, n = 1,2,...,N$, where 0 represents the background and 1 represents the object, N is the number of target objects. The object masks in the next future frame are generated via a decoder $\mathcal{D}$, which takes binary object masks and the plan representation in current time step as inputs:

$$\hat{m}_{t+1}^n = \mathcal{D}(p_t, m_t^n),\ n = 1,2,...,N, \tag{2}$$

where $\hat{m}_{t+1}^n \in \mathbb{R}^{H\times W}$ contains the predicted probability of each pixel belonging to the target object. We then use the soft-aggregation [32] used in multi-object tracking scenarios to fuse the predicted logits from different objects into one logit:

$$(\hat{m}_{t+1}^1, \hat{m}_{t+1}^2, ..., \hat{m}_{t+1}^N) \rightarrow \hat{M}_{t+1} \in \mathbb{R}^{(N+1)\times H\times W}, \tag{3}$$

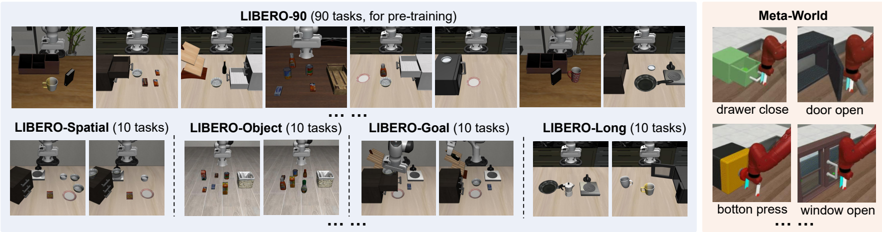

Figure 5: Visualization of different LIBERO task suites and Meta-World tasks.

where each value represents the probability of each pixel belonging to the background or the N objects. The mask generation process of each future Δt step remains consistent with the above. With the ground-truth future masks extracted beforehand, we use bootstrapped cross entropy loss and dice loss with equal weighting following multi-object tracking methods [9, 49] to train the planner $\mathcal{P}$ and decoder $\mathcal{D}$. This way, we leverage the video and extracted object masks to inject rich task-relevant motion information into the latent plan.

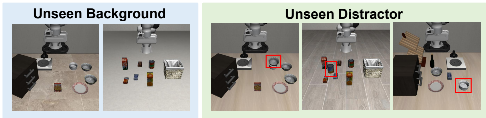

Figure 4: The environments for generalization tests with unseen backgrounds and distractors.

3.3 Policy Learning Guided by Latent Plans

After training the *planner* with video data, the *executor* is trained to infer actions based on robot data and the predicted latent plans. At time step t, given the latest h steps of observations $o_{t-h:t}$, robot states $s_{t-h:t}$ and the predicted plan representations produced by the planner, the executor $\mathcal{E}$ infers an action a_t at the current step:

$$a_t = \mathcal{E}\left(o_{t-h:t},\ s_{t-h:t},\ p_{t-h:t}\right). \tag{4}$$

This policy model is trained in a task-agnostic way following PLEX [35] and MimicPlay [37]. The task specification for multi-task learning is no longer needed since the task knowledge is already incorporated in latent plans. The executor is trained with MSE loss. In this way, the executor has specific role of inferring inverse dynamics from the future plans.

3.4 Implementation Details

We implement *planner* and *executor* with GPT-style transformer structure following PLEX [35] and GR-1 [40]. We process different input modalities with corresponding encoders, using ResNet-18 [19] as image encoders and linear layers as robot state encoders.

During the training of *planner* $\mathcal{P}$, all image features are input to the transformer modules and translated into corresponding plan tokens. Then, the decoder $\mathcal{D}$ concatenates the mask features and plan features to obtain fused features. Note that the number of objects in the mask M_t provided to the decoder $\mathcal{D}$ can be arbitrary as in XMem [9]. We set the maximum number of tracked objects in one task to 3 in LIBERO and 2 in Meta-World, including the robot. During the training of *executor* $\mathcal{E}$, the *planner* $\mathcal{P}$ and image encoder are frozen. During inference, the decoder $\mathcal{D}$ is discarded as in MimicPlay [37].

Method	Latent Plan	LIBERO				Meta-World
		Spatial	Object	Goal	Long	MT10
BC-R3M	-	52.16 ± 7.9	47.56 ± 5.1	38.68 ± 3.3	13.08 ± 2.2	63.44 ± 3.3
PLEX	Image	69.24 ± 2.3	52.44 ± 6.8	21.68 ± 2.0	14.84 ± 3.1	67.92 ± 3.8
MimicPlay	Trajectory	60.52 ± 2.6	62.04 ± 4.2	63.00 ± 5.1	17.88 ± 1.9	67.08 ± 3.1
UniPi	Image	72.04 ± 3.5	60.52 ± 2.7	28.20 ± 2.9	10.96 ± 1.8	70.12 ± 4.7
ATM	Trajectory	66.92 ± 1.6	61.12 ± 5.3	64.84 ± 4.4	24.36 ± 4.8	71.76 ± 3.9
Mask2Act (Ours)	Mask	$\mathbf{76.12 \pm 2.3}$	$\mathbf{68.76 \pm 5.7}$	$\mathbf{75.16 \pm 3.8}$	$\mathbf{30.64 \pm 6.1}$	$\mathbf{78.24 \pm 5.1}$

Table 1: Multi-task learning results on four LIBERO task suites and Meta-World MT10. Mask2Act extracts richer task-relevant knowledge from video data compared with other methods that modeling images or point trajectories as plan representations.

4 Experiments

We first conduct experiments to answer the following questions: 1) Is modeling the transition of object masks better encoding motion information and more effective at learning from action-free video compared with previous video pre-training method? 2) How does our method learn from cross-task pre-training videos and generalize to new tasks and unseen environments during inference? Next, we provide ablation and visualization results to better comprehend our method.

4.1 Experimental Setup

Environments and Tasks. We conduct experiments on the LIBERO [24] and Meta-World [52] benchmark. LIBERO has 130 language-conditioned manipulation tasks with various objects and environments. LIBERO has four task suites shown in Figure 5: **LIBERO-Spatial**, **LIBERO-Object**, **LIBERO-Goal**, and **LIBERO-100**. Meta-World is a collection of 50 manipulation tasks. We consider the official benchmarks of **Multi-Task 10 (MT10)** and **Meta-Learning 10 (ML10)** for video pre-training and cross-task learning.

We train multi-task policies for all benchmarks. Our model is compatible with any form of task specification, such as goal image [37], videos of a task accomplishment [29, 51] and language description [12]. In our experiments, we use goal images following PLEX. Each single task has 50 expert demonstrations containing two camera views, robot states, robot actions and a language instruction. In the first training stage, we utilize each task's 50 demonstration data for action-free video pre-training. We only predict the latent plans on the third-person view following MimicPlay [37] because it contains all the scene information necessary for task planning. In the second stage, we utilize 10 demonstration trajectories with robot states and actions for each task to train a multi-task policy.

Baseline Methods. We evaluate five video pre-training baselines within the planner-executor architecture, which encode different intermediate representations into the latent plan: 1) BC-R3M: a vanilla behavioral cloning baseline with the same policy architecture, in which the plan representations are masked with zero. The pre-trained R3M [31] weights are used to initialize the image encoder. 2) PLEX[35]: a method that learns from video data and noisy action data. The planner of PLEX predicts future image features as latent plans instead of raw image frames. In our experiments, we do not use the noisy action data ($\mathcal{D}_{\text{ttd}}$ in [35]) for training. 3) MimicPlay [37]: a method that predicts the future trajectory of the human hand or robot end-effector as latent plans. Since the pre-training video data in our experiments only contain robot demonstrations, we formulate the future trajectory of the robot as the 2D end-effector trajectory following the simulation experiments in MimicPlay [37] because 3D locations cannot be obtained solely from video data. We denote this baseline as *MimicPlay-robot*. 4) UniPi [12]: a method that trains a text-conditioned video diffusion model to predict future image frames as latent plans. We follow the implementation of UniPi in AVDC [23], which significantly saves the computational cost and remains high fidelity. 5) ATM [39]: a method that predicts the future trajectory of sampled points of current image as latent plans. It uses Co-Tracker [21] to obtain ground-truth 2D point trajectories from video data. The generated image frames or 2D point trajectories serve as latent plan representations, which are only used in the training of the planner.

4.2 Experimental Results

Learning from action-free video data and limited robot action data. The results of four LIBERO task suites and Meta-World's Multi-Task 10 (MT10) benchmark are shown in Table 4.1. Our method of modeling multiple object masks as plan representations significantly outperforms all the video pre-training baselines on all task suites. The comparison with MimicPlay-robot and ATM indicates that the transition of object masks encodes richer task-relevant motion information than the 2D trajectories. Besides, predicting the whole image contents as latent plans as PLEX and UniPi shows significant flaws in more complex tasks, such as LIBERO-Goal and LIBERO-Long, in which the scene and object morphology change more dramatically. In this situation, video prediction may fail to accurately model the task-relevant motion information, leading to unfaithful object shapes or hallucinations. In contrast, object tracking in Mask2Act introduces better object shape consistency and generates object masks with high fidelity, as shown in the visualization in Figure 8.

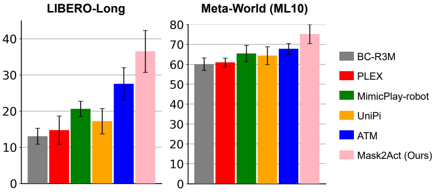

Figure 6: Cross-task video pre-training on LIBERO-Long and Meta-World ML10.

Learning from cross-task pre-training video data. To explore the ability to extract skill knowledge from larger-scale and richer video data and generalize to new tasks, we further conduct experiments in which we pre-train the *planner* on cross-task video and fine-tune the *executor* on different downstream tasks. For LIBERO, the video data from LIBERO-90 serves as the data source for pre-training, and the long-horizon tasks of LIBERO-Long are used for downstream fine-tuning and evaluation. For Meta-World, we consider the official Meta-Learning 10 (ML10) split consisting of 10 tasks for pre-training and 5 tasks for fine-tuning and evaluation. In this scenario, the pre-training data encompasses a wider variety of objects, tasks, and layouts, while the unseen downstream tasks share some generalizable skills and knowledge with the pre-training data. The results in Figure 6 indicate that Mask2Act better captures the task-relevant motion and object manipulation knowledge from cross-task video and has better potential for generalizing to various downstream tasks. Modeling the 2D point trajectories or image frames as latent plans are confronted with the problem of losing motion details and unnecessarily focusing on task-irrelevant information.

Generalizing to unseen environments. By modeling object masks as the intermediate representation between image and motion, the model is trained to focus on task-relevant information and ignore the distracting factors. This facilitates the ability to generalize in unseen environments during inference. To further investigate this aspect, we modify the LIBERO simulation environments to change the background (e.g., the texture and color of table and floor), and add task-related distractors into the scene as shown in Figure 4. We add the distractor that is "task-related", e.g., an additional *bowl* in the task of "*pick up the black bowl from table center and place it on the plate.*". We test the policy model trained with different plan representations in these unseen environments. The results in Figure 7

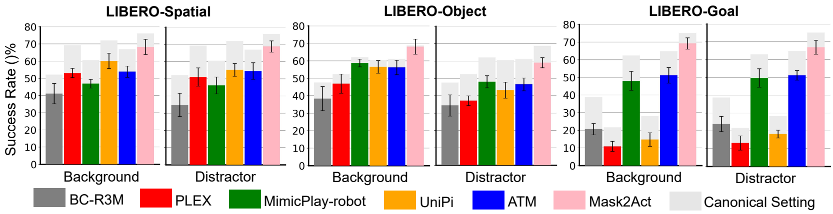

Figure 7: Results of generalization experiments. The gray bars show original results in seen canonical settings.

show that our method maintains a success rate superior to other video pre-training methods in all generalization test, and also has the least performance degradation compared with the canonical settings. This indicates that the object-centric prior effectively guides the model to focus on task-relevant objects, enhancing its ability to handle background changes and mitigate the impact of visual distractors.

Generalizing to novel object instances. To further investigate the benefit of tracking task-relevant objects as latent plans, we test the policy model with novel object instances during inference. We choose all 6 cabinet-related tasks and 14 plate-related tasks from the LIBERO and replace the target objects with novel cabinet and plate instances from the object assets. Results in Table 4.2 demonstrate the benefit of tracking targets in latent plans. By tracking the target object during planning, the model is guided to focus on generalizable knowledge of object morphology rather than its appearance. Without tracking the object, the policy is fully conditioned on the images, which suffers in generalizing to unseen objects.

Tracking Target	Novel Cabinet	Novel Plate
×	51.73 ± 3.8	40.97 ± 4.9
✓	**66.53** ± **4.3**	**51.31** ± **3.7**

Table 2: Generalizing to novel objects with/without tracking target objects.

4.3 Analysis and Ablation Study

In Figure 8, we visualize the initial task states and predicted object masks from the *planner*. Our method produces high-fidelity object masks, which capture accurate object shape consistency and object morphology under conditions of contact and occlusion. These object masks enrich the latent plans with rich task-relevant motion and skill knowledge, providing good guidance for data-efficient policy learning.

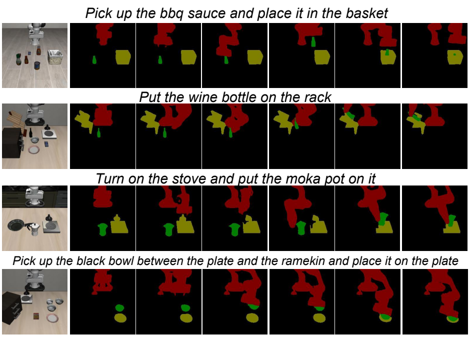

Figure 8: Visualization of the initial task states and predicted object masks.

For ablation, we evaluate our model with different numbers of objects being tracked in one task. We randomly select one or two objects to be tracked in a demonstration trajectory. Besides, we explore the impact of different numbers of robot data used in each task's training.

References

[1] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. 2022.

[2] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *First Workshop on Out-of-Distribution Generalization in Robotics at CoRL 2023*, 2023.

[3] Homanga Bharadhwaj, Abhinav Gupta, Vikash Kumar, and Shubham Tulsiani. Towards generalizable zero-shot manipulation via translating human interaction plans. 2024.

[4] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation. *arXiv preprint arXiv:2405.01527*, 2024.

[5] Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X Lee, Maria Bauza, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, et al. Robocat: A self-improving foundation agent for robotic manipulation. *arXiv preprint arXiv:2306.11706*, 2023.

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2023.

[8] Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. Learning video-conditioned policies for unseen manipulation tasks. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[9] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022.

[10] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5559–5568, 2021.

[11] Zichen Jeff Cui, Yibin Wang, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. From play to policy: Conditional behavior generation from uncurated robot data. In *The Eleventh International Conference on Learning Representations*, 2023.

[12] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 2023.

[13] Yilun Du, Sherry Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, Leslie Pack Kaelbling, et al. Video language planning. In *The Twelfth International Conference on Learning Representations*, 2023.

[14] Chongkai Gao, Yizhou Jiang, and Feng Chen. Transferring hierarchical structures with dual meta imitation learning. In *Conference on Robot Learning*, 2023.

[15] Jiayuan Gu, Sean Kirmani, Paul Wohlhart, Yao Lu, Montserrat Gonzalez Arenas, Kanishka Rao, Wenhao Yu, Chuyuan Fu, Keerthana Gopalakrishnan, Zhuo Xu, et al. Rt-trajectory: Robotic task generalization via hindsight trajectory sketches. In *The Eleventh International Conference on Learning Representations*, 2024.

[16] Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. In *Conference on Robot Learning*, 2020.

[17] Kourosh Hakhamaneshi, Ruihan Zhao, Albert Zhan, Pieter Abbeel, and Michael Laskin. Hierarchical few-shot imitation with skill transition models. In *International Conference on Learning Representations*, 2022.

[18] Haoran He, Chenjia Bai, Ling Pan, Weinan Zhang, Bin Zhao, and Xuelong Li. Large-scale actionless video pre-training via discrete diffusion for efficient policy learning. *arXiv preprint arXiv:2402.14407*, 2024.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *Conference on Robot Learning*, 2023.

[21] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023.

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[23] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to act from actionless videos through dense correspondences. In *The Twelfth International Conference on Learning Representations*, 2023.

[24] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. 2023.

[25] Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499, 2023.

[28] Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. Learning latent plans from play. In *Conference on robot learning*, 2020.

[29] Zhao Mandi, Fangchen Liu, Kimin Lee, and Pieter Abbeel. Towards more generalizable one-shot visual imitation learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2434–2444. IEEE, 2022.

[30] Ajay Mandlekar, Danfei Xu, Roberto Martín-Martín, Silvio Savarese, and Li Fei-Fei. Learning to generalize across long-horizon tasks from human demonstrations. *arXiv preprint arXiv:2003.06085*, 2020.

[31] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *6th Annual Conference on Robot Learning*, 2022.

[32] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9226–9235, 2019.

[33] Gautam Salhotra, I-Chun Arthur Liu, and Gaurav S Sukhatme. Learning robot manipulation from cross-morphology demonstration. In *Conference on Robot Learning*, 2023.

[34] Karl Schmeckpeper, Annie Xie, Oleh Rybkin, Stephen Tian, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Learning predictive models from observation and interaction. In *European Conference on Computer Vision*, 2020.

[35] Garrett Thomas, Ching-An Cheng, Ricky Loynd, Felipe Vieira Frujeri, Vibhav Vineet, Mihai Jalobeanu, and Andrey Kolobov. Plex: Making the most of the available data for robotic manipulation pretraining. In *Conference on Robot Learning*, 2023.

[36] Mel Vecerik, Carl Doersch, Yi Yang, Todor Davchev, Yusuf Aytar, Guangyao Zhou, Raia Hadsell, Lourdes Agapito, and Jon Scholz. Robotap: Tracking arbitrary points for few-shot visual imitation. *arXiv preprint arXiv:2308.15975*, 2023.

[37] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. In *Conference on Robot Learning*, 2023.

[38] Yixuan Wang, Zhuoran Li, Mingtong Zhang, Katherine Rose Driggs-Campbell, Jiajun Wu, Li Fei-Fei, and Yunzhu Li. D^3 fields: Dynamic 3d descriptor fields for zero-shot generalizable robotic manipulation. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

[39] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. In *Robotics: Science and Systems (RSS)*, 2024.

[40] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *The Twelfth International Conference on Learning Representations*, 2023.

[41] Wenke Xia, Dong Wang, Xincheng Pang, Zhigang Wang, Bin Zhao, and Di Hu. Kinematic-aware prompting for generalizable articulated object manipulation with llms. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[42] Annie Xie, Lisa Lee, Ted Xiao, and Chelsea Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation. *arXiv preprint arXiv:2307.03659*, 2023.

[43] Eliot Xing, Abhinav Gupta, Sam Powers, and Victoria Dean. Kitchenshift: Evaluating zero-shot generalization of imitation-based policy learning under domain shifts. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[44] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery. In *Conference on Robot Learning*, pages 3536–3555, 2023.

[45] Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface. *arXiv preprint arXiv:2407.15208*, 2024.

[46] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023.

[47] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

[48] Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Video as the new language for real-world decision making. *arXiv preprint arXiv:2402.17139*, 2024.

[49] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, pages 2491–2502, 2021.

[50] Qiaojun Yu, Junbo Wang, Wenhai Liu, Ce Hao, Liu Liu, Lin Shao, Weiming Wang, and Cewu Lu. Gamma: Generalizable articulation modeling and manipulation for articulated objects. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[51] Tianhe Yu, Chelsea Finn, Sudeep Dasari, Annie Xie, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *Robotics: Science and Systems*, 2018.

[52] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, 2020.

[53] Chengbo Yuan, Chuan Wen, Tong Zhang, and Yang Gao. General flow as foundation affordance for scalable robot learning. *arXiv preprint arXiv:2401.11439*, 2024.

[54] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. In *8th Annual Conference on Robot Learning*.

[55] Yifeng Zhu, Zhenyu Jiang, Peter Stone, and Yuke Zhu. Learning generalizable manipulation policies with object-centric 3d representations. In *Conference on Robot Learning*, 2023.