Learning Video-Conditioned Policy on Unlabelled Data with Joint Embedding Predictive Transformer

Hao Luo[1], and Zongqing Lu[1,2⋆]

[1] School of Computer Science, Peking University [2] Beijing Academy of Artificial Intelligence

北京大学计算机学院
School of Computer Science

Lab Homepage | Scholar Page | Code

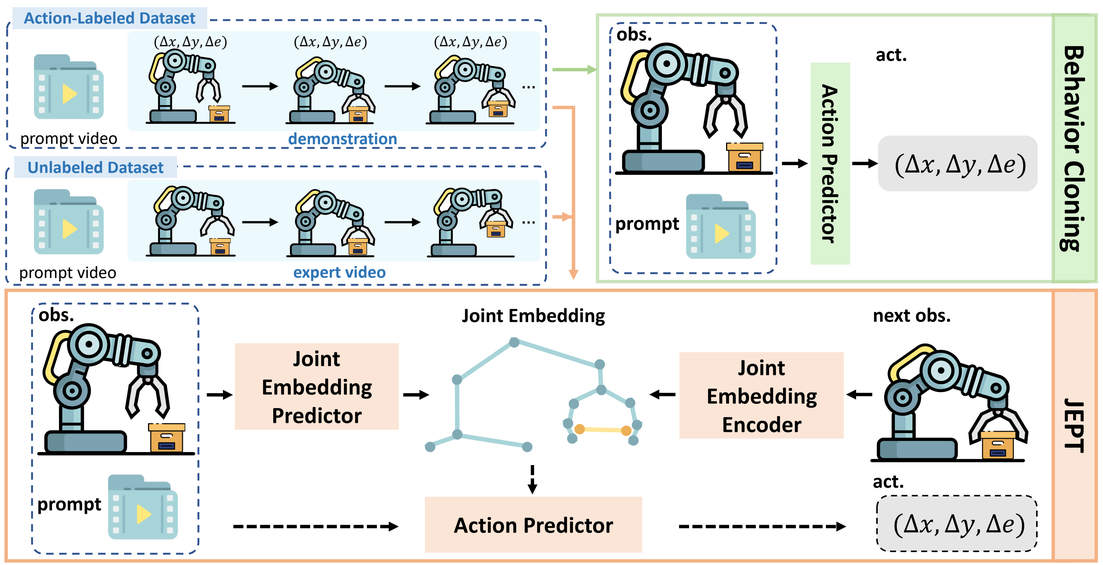
Background

Video-Conditioned Policy Learning

- A video-conditioned policy takes a prompt video depicting the task as a condition and executes the desired task in the dynamics encountered.
- Using a prompt video as a policy condition exhibits a better potential for flexibility and generalization on unseen tasks compared to other task specifications

Action-Labeled and Unlabeled Datasets Mixture

- **Visual imitation learning** requires a dataset of paired prompt videos and **expert demonstrations** with action labels.
- Try to alleviate the action annotation burden by additionally leveraging paired prompt videos and **expert videos without action labels**.



Method

Task Decomposition for Behavior Cloning

- **Q1:** How are the prompt videos supposed to manifest in the dynamics of the tasks?
- **Q2:** What actions are required to realize the given visual transitions?

Visual Transition Prediction

- To predict the **embeddings** of the **next observations**
- Visual transition is **task-specific** but is **learnable from both action-labeled and unlabeled data**.

Inverse Dynamics Learning

- To predict the **action** for the **desired visual transition**
- Inverse dynamics is **learnable merely from the action-labeled** but is **universally applicable across tasks**.

Joint Predictive Embedding

- Better abstract embeddings, better predictive generalization.

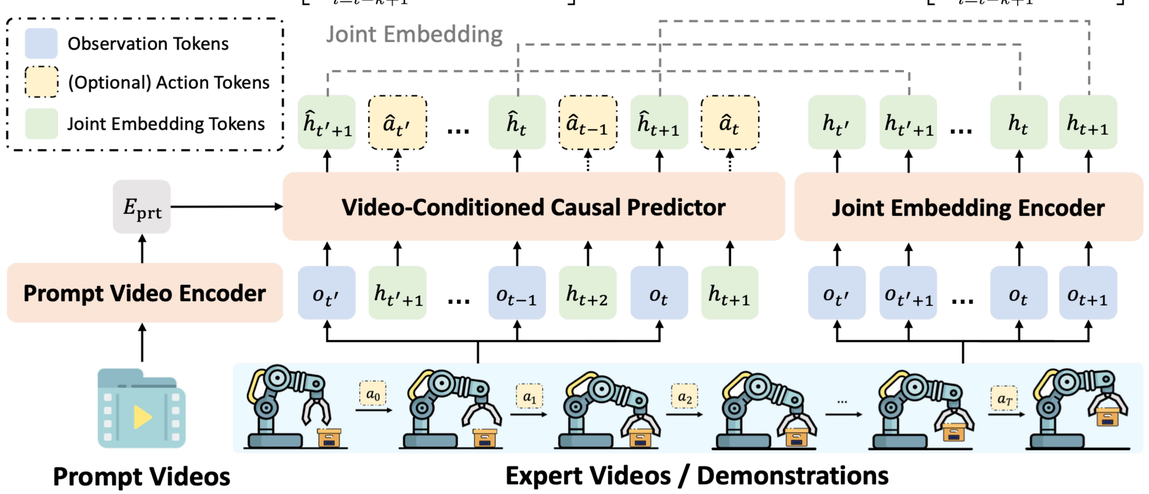
Joint Embedding Predictive Transformer

Visual Transition Loss:

$$\mathcal{L}_{\text{obs}} = \mathbb{E}_{(V,O)\sim\mathcal{D}_{\text{demo}}\cup\mathcal{D}_{\text{vid}}}\left[\frac{1}{k}\sum_{i=t-k+1}^{t}\left\|h_{i+1}-\hat{h}_{i+1}\right\|_2\right]$$

Inverse Dynamics Loss:

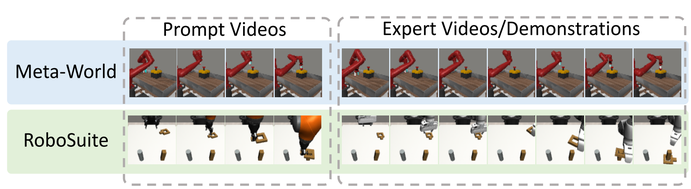
$$\mathcal{L}_{\text{act}} = \mathbb{E}_{(V,O,A)\sim\mathcal{D}_{\text{demo}}}\left[\frac{1}{k}\sum_{i=t-k+1}^{t} a_i \log \hat{a}_i\right]$$

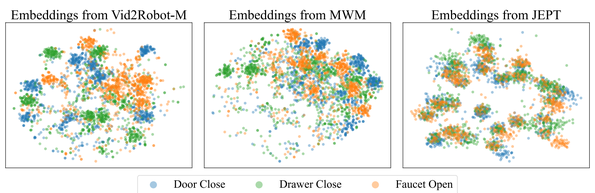


Main Experimental Results

Task	Vid2Robot-D	Vid2Robot-M	BC+IDM	JEPT+MWM	DT*	JEPT
$\mathcal{T}_{\text{demo}}$	51.8	46.3	49.5	44.5	**61.3**	51.3
$\mathcal{T}_{\text{vid}}$	4.7	28.7	8.3	21.3	13.0	**31.7**
Handle Press	10.0	22.0	8.0	18.0	6.0	**28.0**
Lever Pull	0.0	4.0	0.0	4.0	0.0	**10.0**
Plate Slide Back	4.0	8.0	0.0	2.0	0.0	**14.0**
Faucet Open	4.0	14.0	4.0	14.0	0.0	**22.0**
Seen Average	28.2	37.5	28.9	32.9	37.1	**41.5**
Unseen Average	4.5	12.0	3.0	9.5	1.5	**18.5**

Task	Vid2Robot-D	Vid2Robot-M	BC+IDM	JEPT+MWM	DT*	JEPT
$\mathcal{T}_{\text{demo}}$	26.8	20.7	26.7	24.3	**36.3**	27.2
$\mathcal{T}_{\text{vid}}$	4.0	11.6	8.4	10.8	2.0	**12.8**
Panda Lift	4.0	16.0	6.0	22.0	8.0	**38.0**
Sawyer Lift	0.0	2.0	0.0	6.0	0.0	**16.0**
IIWA Lift	0.0	4.0	0.0	2.0	0.0	**12.0**
UR5e Lift	0.0	0.0	0.0	0.0	0.0	**8.0**
Seen Average	15.5	16.1	17.5	17.6	19.2	**20.1**
Unseen Average	1.0	5.5	1.5	7.5	2.0	**18.5**





- JEPT achieves **superior overall performance** and **better generalization** on unseen tasks compared with either direct visual imitation learning or conducting two subtasks with other embeddings.
- JEPT effectively **leverages the mixture dataset** to enhance the performance on the unlabeled and unseen tasks.
- JEPT learns **more distributionally consistent embeddings** compared with baselines.
- More ablations and analyses are available in the paper due to the limited space.