Viewpoint-Invariant Latent Action Learning from Human Video Demonstrations

Jung Min Lee[1] jmleeluck@cml.snu.ac.kr
Dohyeok Lee[1] dohyeoklee@cml.snu.ac.kr
Jungwoo Lee[1*] junglee@snu.ac.kr

[1]Department of Electrical and Computer Engineering
Seoul National University

Abstract

Learning representations of visual transitions between consecutive frames in video enables robots to learn from both robot and human demonstrations. These representations, referred to as latent actions, capture the inherent state transition. However, continuously changing viewpoints in human videos introduce the ambiguity that hinders consistent modeling of latent actions. To address this issue, we propose **Vi**ew**P**oint-**I**nvariant **L**atent **A**ction, or ViPILA, a representation of visual transitions that is robust to the viewpoint variation from human videos without action labels and camera calibrations. Building on a theoretical analysis of viewpoint-invariance, we introduce a novel training objective that enforces consistency in latent actions across different viewpoints of the same state. The key idea is to enforce that a latent action inferred from one viewpoint can be used to reconstruct the observation from a different viewpoint, as long as the underlying state remains the same. We empirically demonstrate that the resulting viewpoint-invariant latent actions improve downstream manipulation policy learning in LIBERO simulation.

1 Introduction

Learning from videos (LfV) has emerged as a promising direction to scale robot learning without requiring costly, embodiment-specific datasets[1]. By leveraging large-scale human demonstration videos, LfV enables learning generalizable priors about behaviors and physical dynamics across diverse tasks and environments. However, applying human video demonstrations to robot learning faces fundamental challenges: the lack of explicit low-level action labels and the distributional shift introduced by changes in embodiment and viewpoint [1, 2].

Recent works [3, 4, 5, 6] have proposed modeling *latent actions*, compact representations that capture how visual observations change over time due to an agent's behavior, allow the policy to acquire manipulation skills from human video demonstration. It implies that learning latent actions enables policy to transfer across embodiments. However, viewpoint variation remains a major hurdle for learning from human videos and hinders learning latent actions, interrupting the network to understand of true dynamics [4, 5]. In particular, viewpoint changes in human videos not only cause the distribution shift but introduce ambiguity when interpreting the source of visual changes. A change in pixel values arises either from the agent's action or from a shift in viewpoint. We suppose that this uncertainty makes it difficult to disentangle action-induced dynamics from viewpoint-induced artifacts.

To address this, we propose an unsupervised learning framework that learns **Vi**ew**P**oint-**I**nvariant **L**atent **A**ction (ViPILA). Instead of enforcing the latent action to merely encode the pixel transitions between frames, ViPILA explicitly encourages the learned latent actions in the same state transition

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

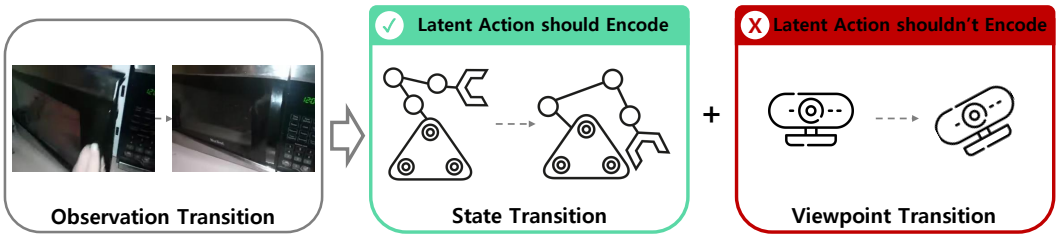

Figure 1: When learning latent actions from observation transitions, the data comprise both state transitions, which the latent action should encode, and viewpoint variation which it should not. A naïve pixel-reconstruction objective that models the observation transitions lacks a mechanism to distinguish these transitions. To exclude the influence of viewpoint changes during latent action learning, ViPILA encodes latent actions that are consistent across multiple viewpoints, thereby promoting viewpoint invariance.

to remain consistent under different camera viewpoints. The requirement for the consistency of latent actions across the viewpoints leads the dynamics model to focus on the underlying physical transitions rather than viewpoint-specific pixel changes. Our experiments show that it enables the model to predict the frame transition consistently, even under significant variations in the camera viewpoint.

We model a latent action learning framework theoretically and define the viewpoint-invariance in this framework. Also, we validate the viewpoint-invariance in terms of preventing spurious action, a counterfeit action that appears to induce a transition in the pixel space but corresponds to a different transition, or no valid transition, in the underlying physical state space. We found that spurious action is introduced since we cannot recover or estimate the inherent state from the pixel space, and the viewpoint-invariance can mitigate it. Based on this framework, we propose a novel training objective that makes latent action viewpoint-invariant. On top of this theoretical background, we empirically demonstrate the effectiveness of our framework. With the robot video dataset (without accessing action labels or camera extrinsics) and human video dataset, we demonstrate that the proposed framework induces the trained a latent action model to be viewpoint-invariant, which cannot be acquired from naïve multi-viewpoint training. In addition, to confirm the effectiveness of our framework in downstream tasks, we train latent action conditioned policy in a simulation benchmark. Our experiments demonstrate that viewpoint-invariant latent actions allow the policy to maintain robust performance even though the latent action is extracted from an unseen viewpoint in policy training. We summarize our main contributions as follows:

- We formulate the notion of viewpoint-invariant latent action and derive its theoretical properties under an inverse dynamics framework.
- We show that ViPILA can be learned from two independent single-view datasets without explicit synchronization in trajectories. This offers the potential to train ViPILA in heterogeneous manipulation videos, enabling dataset scaling.
- The policy conditioned on ViPILA outperforms the baseline when latent actions are inferred from viewpoints different from those used during training.

2 Related Works

Latent Action Learning from Video. Recent progress in video-based robot learning has explored how to extract meaningful representations from large-scale human demonstration videos to support downstream tasks. Some works aim to learn visual priors such as human poses [7], object affordances [8, 9], or trajectory information [10, 11]. Another line of work focuses on learning latent actions, abstraction of temporal transition in video, in an unsupervised manner by modeling visual transitions across video frames [3, 4, 5, 6, 12, 13]. These methods typically encode differences between consecutive frames to model temporal dynamics without access to action labels. Various works show that latent actions extracted from the large corpus of video are effective for cross-embodiment control. For instance, LAPA [3], UniVLA[5], and ViLLA[6] proposed the framework for learning an embodiment-agnostic vision-language-action (VLA) model by training a discrete codebook of latent

actions. In addition, UniSkill[4] shows that such representations can serve as cross-embodiment skill abstractions and effectively improve skill-conditioned policy.

While these works demonstrate that latent actions can serve as intermediate cross-embodiment representations, they do not examine how viewpoint variation in videos influences the construction of latent actions. In contrast, we investigate the impact of viewpoint changes on latent action learning and introduce ViPILA based on this insight.

Learning Viewpoint-Invariance from Human Video. In the realm of computer vision, learning viewpoint-invariant representations from human videos is widely studied[14, 15, 16]. Recently, Viewpoint Rosetta Stone [17] proposes a soft alignment between two independent single-view videos to learn viewpoint-invariant representations, opening potential to learn viewpoint invariance at scale.

In robot learning, the viewpoints in open-source robot datasets are highly limited[18], which motivates leveraging human manipulation videos to learn viewpoint-invariant representations. R3M[19] uses diverse human manipulation videos which feature a wide range of viewpoints, to train a universal visual encoder for robotic manipulation. Due to the diverse viewpoints present in human videos, the model demonstrates improved generalization across unseen viewpoints.

Viewpoint-Invariant Representation in Robot Learning. When using images as observations of the policy, viewpoint variation in the camera is one of the main bottlenecks of the generalizability of trained policy[20, 21, 22]. To attain a policy generalizable to the viewpoint variation, one line of work leverages 3D-aware representation extracted from the scene, using point clouds[23, 20, 24], or Neural Radiance Fields (NeRFs)[25, 26]. Another line of work injects the 3D inductive bias through a novel view synthesis model via data augmentation[21, 22]. In addition, viewpoint-robust policies are studied in the field of a world model[27]. For example, ReViWo [28] introduces a world model that disentangles viewpoint-dependent and viewpoint-independent features via contrastive learning. MV-MWM[29] uses viewpoint randomization to encourage viewpoint-invariance.

These methods, however, primarily focus on learning viewpoint-invariant state or observation features. In contrast, our work aims to achieve viewpoint-invariance at the latent action level. Enforcing viewpoint-invariance at the latent action level is particularly important for enabling policy transfer across embodiments. Furthermore, many of these approaches require camera extrinsic annotations or images of the same scene captured from multiple viewpoints, which is impractical to collect at scale in the real-world. In contrast, our method relies on neither explicit viewpoint annotations nor exhaustive multi-view coverage. Instead, it can synthesize training pairs from unpaired single-view samples, substantially reducing data collection cost. To our knowledge, our method is the first to enforce viewpoint-invariance directly in the latent action space without expensive annotations.

3 Method

In this section, we provide the theoretical framework for ViPILA learning. We first formulate the latent action learning setup in LfV and show how ViPILA can be beneficial when learning from large-scale human video datasets by mitigating spurious actions. Finally, we propose a simple framework that enforces viewpoint-invariance in dynamics learning. Extended theoretical analysis and proofs are provided in Appendix A.

3.1 Problem Formulation

Consider a state space $\mathcal{S}$ and an action space $\mathcal{A}$. Let the *forward dynamics* $f : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ map the current state $s_t \in \mathcal{S}$ and action $a_t \in \mathcal{A}$ to the next state $s_{t+1} \in \mathcal{S}$:

$$s_{t+1} = f(s_t, a_t) \tag{1}$$

In general, the true state s_t is not directly observable in various scenarios, including human videos. Such partial observability often results from the loss of depth information in monocular video and occlusions or improper camera angles that prevent complete capture of the manipulator and other task-relevant entities. Instead, we only observe a sequence of images rendering the state over time. Suppose we have a collection of the viewpoints $\mathcal{T}$ in demonstrations. Given a viewpoint $T \in \mathcal{T}$, we define a sensor model $h : \mathcal{T} \times \mathcal{S} \to \mathcal{O}$ that maps the underlying state s to the observed image o:

$$o = h_T(s) \tag{2}$$

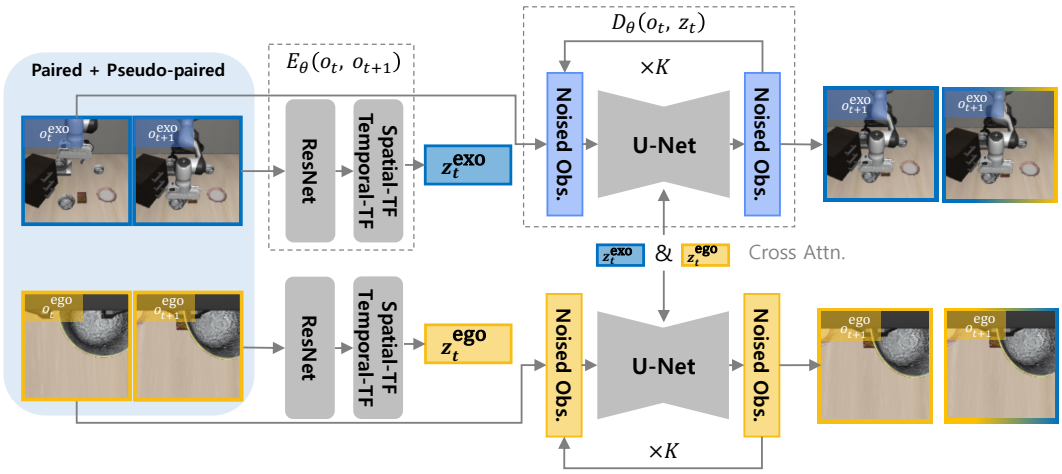

Figure 2: Architecture of ViPILA learning. Given observation transitions (o_t, o_{t+1}) from egocentric and exocentric viewpoints, observation inverse dynamics $E_\theta(o_t, o_{t+1})$ encode the latent action z_t from each viewpoint (**left**). Then, observation forward dynamics $D_\theta(o_t, z_t)$ reconstructs the next observation o_{t+1} from the current observation o_t and either latent action from egocentric view or exocentric view (**right**).

Note that h_T is generally a non-injective function over the state space $\mathcal{S}$.

We can also define the forward dynamics in the observation space $\mathcal{O}$. Given an observation o_t and action a_t, the *observation forward dynamics* $D : \mathcal{O} \times \mathcal{A} \to \mathcal{O}$ maps to the next observation:

$$o_{t+1} = D(o_t, a_t) \tag{3}$$

To explicitly model viewpoint changes by the agent's behavior, we define the *viewpoint dynamics* $\Phi : \mathcal{T} \times \mathcal{A} \to \mathcal{T}$, which describes how the viewpoint $T_t \in \mathcal{T}$ changes to $T_{t+1} \in \mathcal{T}$ in response to action a_t:

$$T_{t+1} = \Phi(T_t, a_t) \tag{4}$$

For instance, if $T_t = \Phi(T_t, a_t)$ for $\forall T_t \in \mathcal{T}, \forall a_t \in \mathcal{A}$, it describes the static camera. Note that, in general, the viewpoint T_t is also unknown like the state s_t. Then, we can define a viewpoint trajectory $\tau_T = (T_0, a_0, T_1, a_1, \ldots, a_{H-1}, T_H)$, where $T_{t+1} = \Phi(T_t, a_t)$ for $0 \le t < H$. Consider another trajectory $\tau_{T'}$, which follows a different viewpoint dynamics Φ' but shares the same action sequence:

$$\tau_{T'} = (T'_0, a_0, T'_1, a_1, \ldots, a_{H-1}, T'_H) \tag{5}$$

where $T'_{t+1} = \Phi'(T'_t, a_t)$. We refer to $\tau_{T'}$ as the *paired trajectory* of τ_T. Based on this, we define the set of paired viewpoints $\mathcal{P}_{T_t}$ as the set of all viewpoints that can be reached at timestep t by replaying the same action sequence from T_0 to T_t under Φ', starting from an arbitrary initial viewpoint $T'_0 \in \mathcal{T}$:

$$\mathcal{P}_{T_t} = \left\{ T'_t \in \mathcal{T} \middle| \exists T'_0 \in \mathcal{T} \text{ s.t. } T'_{k+1} = \Phi'(T'_k, a_k) \text{ for } 0 \le k < t \right\} \tag{6}$$

3.2 Theoretical Analysis of Viewpoint-Invariance & Spurious Actions

Define the *inverse dynamics* $g : \mathcal{S} \times \mathcal{S} \to \mathcal{A}$ as:

$$g(s_t, s_{t+1}) = \left\{ a_t \in \mathcal{A} \middle| f(s_t, a_t) = s_{t+1} \right\} \tag{7}$$

Similarly, define the *observation inverse dynamics* $E : \mathcal{O} \times \mathcal{O} \to \mathcal{A}$ as:

$$E(o_t, o_{t+1}) = \left\{ a_t \in \mathcal{A} \middle| D(o_t, a_t) = o_{t+1} \right\} \tag{8}$$

Note that the observation inverse dynamics E depends not only on the underlying state s_t but on the viewpoint T_t from which the observation is rendered. For simplicity, we refer to the observation

inverse dynamics model E as the inverse dynamics model (IDM). Then, we define viewpoint-invariance of the observation inverse dynamics E as follows:

Definition 3.1. *The observation inverse dynamics E is said to be viewpoint-invariant on $\mathcal{V} \subseteq \mathcal{T}$ if*

$$E(h_{T_t}(s_t), h_{T_{t+1}}(s_{t+1})) = E(h_{T'_t}(s_t), h_{T'_{t+1}}(s_{t+1})) \tag{9}$$

holds for all $s_t, s_{t+1} \in \mathcal{S}$, $T_t, T_{t+1} \in \mathcal{V}$, $T'_t \in \mathcal{P}_{T_t}$, and $T'_{t+1} \in \mathcal{P}_{T_{t+1}}$.

Thus, if an action changes both state and viewpoint, the observation inverse dynamics E must assign it consistently across the viewpoints.

Proposition 3.2. *For any $s_t, s_{t+1} \in \mathcal{S}$ and any $T_t \in \mathcal{T}$,*

$$a_t \in g(s_t, s_{t+1}) \Longrightarrow a_t \in E(h_{T_t}(s_t), h_{\Phi(T_t, a_t)}(s_{t+1})) \tag{10}$$

Proposition 3.2 shows that any action which is valid in the true state space is also considered valid by the observation inverse dynamics E. However, since the converse does not generally hold, E may overestimate the set of valid actions—it can include the actions that appear plausible in observation space but do not correspond to actual state transitions. We define such action by *spurious action*. This overestimation arises from the non-injective nature of the sensor model h_T, which can map different states to the same visual observations, introducing ambiguity in interpreting pixel changes. This uncertainty is similar to the perceptual aliasing problem in POMDPs[30].

From this consideration, we introduce the notion of *non-spuriousness*, which ensures that any action predicted by E is truly valid in the underlying state space.

Definition 3.3. *The observation inverse dynamics E is said to be non-spurious on $\mathcal{V} \subseteq \mathcal{T}$ if and only if*

$$a_t \in E(h_{T_t}(s_t), h_{T_{t+1}}(s_{t+1})) \Longrightarrow a_t \in g(s_t, s_{t+1}) \tag{11}$$

holds for all $s_t, s_{t+1} \in \mathcal{S}$ and $T_t, T_{t+1} \in \mathcal{V}$.

While the non-spuriousness is desirable, directly enforcing it is intractable without access to true states or viewpoints. Instead, we show that enforcing viewpoint-invariance on E suffices for non-spuriousness. The following theorem formalizes this connection.

Theorem 3.4. *Let $\mathcal{C}$ be defined as:*

$$\mathcal{C} = \left\{ T \in \mathcal{T} \,\middle|\, \exists T_1, T_2, \cdots, T_n \in \mathcal{P}_T \text{ s.t. } (h_{T_1}, \cdots, h_{T_n}) \text{ is jointly-injective in } \mathcal{S} \right\}$$

If the observation inverse dynamics E is viewpoint-invariant on $\mathcal{T}$, then it is non-spurious on $\mathcal{C}$.

Theorem 3.4 formalizes how viewpoint-invariance helps eliminate spurious actions. The key idea is that given the state transition if there exists at least one viewpoint that can clearly distinguish this transition through the camera, then the observation inverse dynamics E, estimating the action in that view, rejects actions that lead to incorrect transitions. In other words, the model should contain non-spurious action from that viewpoint.

Now, if E is viewpoint-invariant, the same action prediction must hold across all viewpoints. Therefore, any action that is invalid from a certain viewpoint must also be rejected in all other viewpoints. This means that enforcing viewpoint-invariance enables E to filter its predictions to actions that truly induce a given state transition, even for viewpoints that cannot themselves distinguish that transition.

3.3 Viewpoint-Invariance Latent Action Learning

We now present an unsupervised learning framework for training viewpoint-invariant observation inverse dynamics in practice. We use a paired dataset $\mathcal{D}^{\text{paired}}$ consisting of trajectories $\tau_{\text{paired}} = \left\{(o_t^{\text{ego}}, o_t^{\text{exo}})_{t=1}^H, \ell\right\}$, where o_t^{ego} and o_t^{exo} are time-synchronized egocentric and exocentric observations, and ℓ is a language embedding describing the action in the trajectory. Note that the language ℓ could be obtainable reliably by automatic speech recognition (ASR) or video captioning models[31, 32]. We also consider unpaired datasets: $\mathcal{D}^{\text{ego-only}}$ and $\mathcal{D}^{\text{exo-only}}$, containing trajectories of the form $\tau_{\text{ego}} = \left\{(o_t^{\text{ego}}, \varnothing)_{t=1}^H, \ell_{\text{ego}}\right\}$ and $\tau_{\text{exo}} = \left\{(\varnothing, o_t^{\text{exo}})_{t=1}^H, \ell_{\text{exo}}\right\}$, respectively.

Let D_θ and E_θ be the observation forward and inverse dynamics models, parameterized by θ. For samples from $\mathcal{D}^{\text{paired}}$ and a metric $d : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$, we define the unsupervised loss $\mathcal{L}_{\text{paired}}$ as:

$$\mathcal{L}_{\text{paired}}(\theta) = \underbrace{\frac{1}{2}\left[\sum_{a_t \in E_\theta\left(o_t^{\text{ego}}, o_{t+1}^{\text{ego}}\right)} d\Big(o_{t+1}^{\text{ego}}, D_\theta(o_t^{\text{ego}}, a_t)\Big) + \sum_{a'_t \in E_\theta\left(o_t^{\text{exo}}, o_{t+1}^{\text{exo}}\right)} d\Big(o_{t+1}^{\text{exo}}, D_\theta(o_t^{\text{exo}}, a'_t)\Big)\right]}_{\text{self-viewpoint}}$$

$$+ \underbrace{\frac{1}{2}\left[\sum_{a'_t \in E_\theta\left(o_t^{\text{exo}}, o_{t+1}^{\text{exo}}\right)} d\Big(o_{t+1}^{\text{ego}},\ D_\theta(o_t^{\text{ego}}, a'_t)\Big) + \sum_{a_t \in E_\theta\left(o_t^{\text{ego}}, o_{t+1}^{\text{ego}}\right)} d\Big(o_{t+1}^{\text{exo}},\ D_\theta(o_t^{\text{exo}}, a_t)\Big)\right]}_{\text{cross-viewpoint}}$$

If $\mathcal{L}_{\text{paired}} = 0$, then the inverse dynamics model E_θ becomes the viewpoint-invariant for the viewpoints of o_t^{ego} and o_t^{exo}. We provide theoretical justification for this loss function in Appendix B. We call the first two terms in $\mathcal{L}_{\text{paired}}$ the self-viewpoint terms, and the latter two terms the cross-viewpoint terms.

To leverage unpaired data, we adopt pseudo-pairing between $\mathcal{D}^{\text{ego-only}}$ and $\mathcal{D}^{\text{exo-only}}$ using soft alignment inspired by [17]. Let $|\mathcal{D}^{\text{exo}}| < |\mathcal{D}^{\text{ego}}|$. Using a pretrained video encoder $Z(\cdot)$ and a cosine similarity function $\text{sim}(\cdot, \cdot)$, we pair each exocentric observation sequence $\tau_{\text{exo}} \sim \mathcal{D}^{\text{exo-only}}$ with its most similar egocentric observation sequence $\tilde{\tau}_{\text{ego}}$:

$$\tilde{\tau}_{\text{ego}} = \underset{\tau_{\text{ego}} \sim \mathcal{D}^{\text{ego-only}}}{\arg\max}\ \text{sim}\left([Z(\tau_{\text{exo}});\ \ell_{\text{exo}}],\ [Z(\tau_{\text{ego}});\ \ell_{\text{ego}}]\right) \tag{12}$$

Then, for egocentric trajectory $\tilde{\tau}_{\text{ego}} = \left\{(\tilde{o}_t^{\text{ego}},\ \varnothing)_{t=1}^{H},\ \ell_{\text{ego}}\right\}$ that has the highest cosine similarity to the τ_{exo}, we form a dataset $\mathcal{D}^{\text{pseudo-paired}}$ with the trajectory $\tau_{\text{pseudo-paired}} = \left\{(\tilde{o}_t^{\text{ego}}, o_t^{\text{exo}})_{t=1}^{H}, \ell_{\text{exo}}, \ell_{\text{ego}}\right\}$ and apply the same loss $\mathcal{L}_{\text{paired}}$, weighted by the similarity score between language descriptions:

$$\mathcal{L}_{\text{pseudo-pair}}(\theta) = \text{sim}(\ell_{\text{exo}}, \ell_{\text{ego}}) \times \mathcal{L}_{\text{paired}}(\theta) \tag{13}$$

Since the narrations in paired dataset $\mathcal{D}^{\text{paired}}$ with same trajectory is identical across the viewpoint ($\ell_{\text{ego}} = \ell_{\text{exo}}$), $\text{sim}(\ell_{\text{ego}}, \ell_{\text{exo}}) = 1$. Thus, we can rewrite the objective, including both paired and pseudo-paired dataset as follows:

$$\mathcal{L}_{\text{w-vi}}(\theta) = \text{sim}(\ell_{\text{ego}}, \ell_{\text{exo}}) \times \mathcal{L}_{\text{paired}}(\theta) \tag{14}$$

for paired trajectory τ_{paired} and pseudo-paired trajectory $\tau_{\text{pseudo-paired}}$. We define the $\mathcal{L}_{\text{w-vi}}$ as *weighted viewpoint-invariance loss*.

3.4 Implementation

To implement E_θ we use ResNet[33] and several ST-transformer layers[34]. For D_θ, we use a U-Net with diffusion objective, following the off-the-shelf image-editing model[35]. We replace d in $\mathcal{L}_{\text{paired}}$ with the noise prediction loss as follows:

$$d(o_{t+1}, D_\theta(o_t, a_t)) \Longrightarrow \mathbb{E}_{\tau \sim \text{Uniform}\{1,\ldots,K\}, \epsilon \sim \mathcal{N}(0,I)}\left[\|\epsilon - \epsilon_\theta(\tilde{o}_{t+1}^{(\tau)}, \tau | o_t, a_t)\|^2\right] \tag{15}$$

where $\tilde{o}_{t+1}^{(\tau)}$ forward-noised target at diffusion timestep $\tau \in [1, K]$. We implement D_θ by denoising Gaussian noise with ϵ_θ. The overall architecture of ViPILA is summarized in Figure 2.

With the trained observation inverse dynamics model E_θ, we can train the policy π_ϕ conditioned on the latent action extracted from E_θ. Given observations o_t, o_{t+1} sampled from the robot manipulation demonstration $\mathcal{D}_{\text{robot}}$, the inverse dynamics model extracts the latent action representation $z_t^{\mathcal{A}} = E_\theta(o_t, o_{t+1})$.[2] Then, the behavior cloning policy π_ϕ is trained by maximizing log-likelihood on the dataset $\mathcal{D}_{\text{robot}}$:

$$\phi^\star = \underset{\phi}{\arg\max}\ \mathbb{E}_{(o_t, a_t, o_{t+1}) \sim \mathcal{D}_{\text{robot}}}\left[\log \pi_\phi(a_t | o_t, z_t^{\mathcal{A}})\right], \quad z_t^{\mathcal{A}} = E_\theta(o_t, o_{t+1}) \tag{16}$$

[2]We rewrite latent action to $z_t^{\mathcal{A}}$ for clarity.

4 Experiments

We empirically validate ViPILA and evaluate its effectiveness in robot manipulation through simulation benchmarks. We aim to address the following research questions: **(1)** Can the proposed framework successfully induce viewpoint-invariant observation inverse dynamics model E? **(2)** Is a pseudo-paired dataset advantageous for promoting viewpoint-invariance in E? **(3)** Do ViPILA improve policy performance, especially when the viewpoint extracting latent action differs from that used during training?

4.1 Experiment Setup

Datasets. Table 1 summarizes the datasets for latent action learning. Paired data come from LIBERO [36] and H2O [37], while unpaired data are from BridgeData V2 [38] (exo-only robots) and Something-Something V2 [39] (ego-only humans). To balance scales (220K vs. 60K trajectories), we match each BridgeData V2 trajectory with the most similar egocentric video from Something-Something V2.

Table 1: Summary of datasets.

	Paired	**Unpaired**
Robot	LIBERO	BridgeData V2
Human	H2O	Something-Something V2

Baselines. We compare the observation inverse dynamics model trained with $\mathcal{L}_{\text{w-vi}}$ (VIPILA) against two variants: (SELF-VIEW) trained only with the self-viewpoint term to assess the role of cross-view consistency, and (PAIR-ONLY) trained without pseudo-paired data to examine its impact. We further evaluate the learned latent actions in downstream manipulation by training diffusion policies conditioned on the latent actions from SELF-VIEW and VIPILA.

Evaluation. We use small validation splits for dynamics evaluation and assess policies on three unseen LIBERO [36] pick-and-place tasks. To test robustness to viewpoint shifts, we train policies using latent actions from an exocentric viewpoint and evaluate with either egocentric or exocentric prompts at inference. During inference, a small subset of demonstrations is used as prompts, and at each timestep the policy receives the current observation o_t and a latent action from the inverse dynamics model with the target o_{t+1} samplied eight framed ahead. (See Appendix D.)

4.2 Experiment Results

Does the Inverse Dynamics Model Learn Viewpoint-Invariant Latent Actions?

Table 2: Comparison of average PSNR (dB), pixel MSE of latent actions under two models: VIPILA vs SELF-VIEW. Format: (viewpoint of latent action extraction) → (viewpoint of rendering).

Metrics	**Ego → Ego**	**Exo → Ego**	**Exo → Exo**	**Ego → Exo**
PSNR ↑				
VIPILA	**14.72**	**14.71**	19.36	**19.31**
SELF-VIEW	14.45	13.51	**19.89**	17.01
MSE ↓				
VIPILA	**0.042**	**0.042**	0.014	**0.014**
SELF-VIEW	0.053	0.059	**0.012**	0.031

Table 2 presents a quantitative comparison between VIPILA and SELF-VIEW on the validation dataset. VIPILA consistently outperforms its counterpart in cross-viewpoint settings (e.g., Ego → Exo and Exo → Ego), achieving higher PSNR and lower MSE. We provide the details of the dynamics learning in Appendix C.

Furthermore, VIPILA maintains stable performance across both self- and cross-view predictions, whereas SELF-VIEW degrades notably under viewpoint shifts. This indicates that self-veiw training alone is insufficient to induce viewpoint-invariance even with multi-view images. In contrast, the explicit viewpoint-invariance constraint in VIPILA enables the model to generalize across viewpoints, mitigating overfitting to specific camera viewpoints. Interestingly, VIPILA also surpasses SELF-VIEW in the (Ego → Ego) setting, suggesting that viewpoint-invariance benefits even nominal

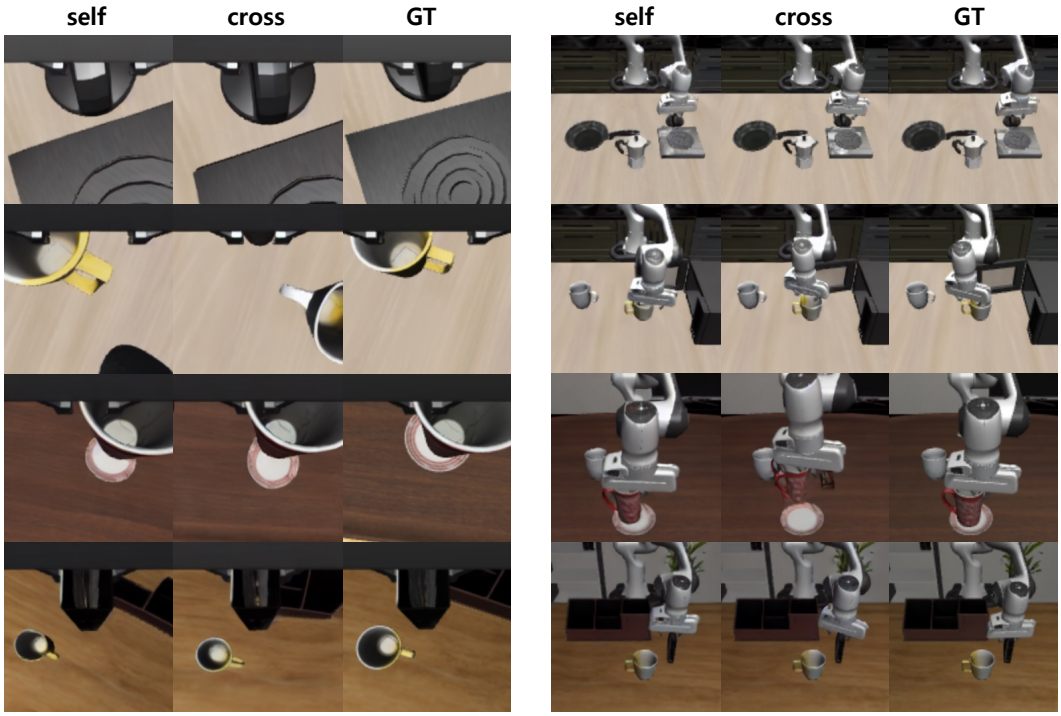

Figure 3: Samples of dynamics prediction results. **self** refers to the prediction in the same viewpoint, while **cross** refers to the prediction in a different viewpoint.

self-view predictions. Specifically, egocentric observations often experience continuous variation due to the camera motion, and learning a viewpoint-invariant representation allows the model to better capture the underlying dynamics despite these shifts.

Figure 3 shows qualitative samples of the dynamics model on the LIBERO dataset. Each column present (i) the self-viewpoint prediction, (ii) the cross-viewpoint prediction, and (iii) the ground truth. While minor spatial misalignments appear in the cross-viewpoint predictions, the overall scene layout, object identities, and motion trends are preserved, demonstrating successful transfers of latent actions across viewpoints.

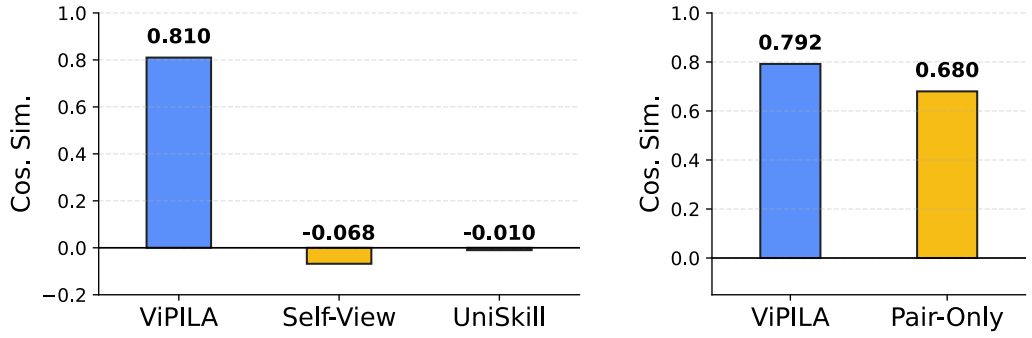

Figure 4: Comparison of cosine similarity. (Left): evaluated on the whole validation set. (Right): evaluated on the paired dataset included in the validation set, excluding the pseudo-paired dataset.

We further compute the cosine similarity between latent actions extracted from egocentric and exocentric views for the same state. As a baseline, we also evaluate the cosine similarity of latent actions from UniSkill[4] in same the manner. As the Figure 4 (left) demonstrates, ViPILA yields

higher cosine similarity than baselines, indicating that the latent representations are aligned across viewpoints. We provide additional analysis of dynamics learning in Appendix C.3.

Is Pseudo-paired Dataset Beneficial to the Inverse Dynamics Model? To investigate the impact of the pseudo-paired dataset, we compare the cosine similarity of latent actions from the models trained with and without the pseudo-paired dataset. Figure 4 (right) demonstrates that VIPILA achieves a significantly higher cosine similarity (0.792) than PAIR-ONLY (0.680). This improvement demonstrates that pseudo-pairing effectively regularizes the inverse dynamics model, promoting consistent latent action across viewpoints and reinforcing viewpoint-invariance in the learned representation. See Appendix C.3 for the next observation prediction result.

Are Viewpoint-Invariant Latent Actions Advantageous to the Robot Learning? Table 3 shows the results of the policy training. We present both the AUC (area under the success rate curve) and the success rate, where the success rate is formatted as (maximum success rate)/(average success rate of last 4 checkpoints). The policy conditioned on VIPILA performs comparably to SELF-VIEW when the prompt and training viewpoint align. However, when they differ, the policy with VIPILA attains promising performance, whereas the policy with SELF-VIEW fails to complete any of the evaluation tasks. This improvement results from the consistent representation VIPILA provides to the policy, enabling the policy to leverage the learned knowledge regardless of changes in prompt viewpoint. In contrast, SELF-VIEW produces viewpoint-dependent latent actions, hindering the reuse of learned behaviors at inference. This results demonstrate the effectiveness of VIPILA for viewpoint transfer in robot manipulation.

Table 3: AUC and average success rate on three LIBERO pick-and-place tasks. **Exo** and **Ego** indicate the viewpoint from which the prompt is provided.

IDM	**Task 1**		**Task 2**		**Task 3**		**Ave.**	
	Exo	**Ego**	**Exo**	**Ego**	**Exo**	**Ego**	**Exo**	**Ego**
VIPILA								
AUC↑	0.72	**0.46**	**0.63**	**0.15**	**0.76**	**0.41**	**0.70**	**0.34**
Succ. Rate↑	**1.00/1.00**	**1.00/0.67**	**1.00**/0.75	**1.00/0.25**	**1.00/0.83**	**0.67/0.33**	**1.00/0.86**	**0.89/0.42**
SELF-VIEW								
AUC↑	**0.76**	0.00	0.59	0.00	0.65	0.00	0.67	0.00
Succ. Rate↑	1.00/0.92	0.00/0.00	**1.00/0.83**	0.00/0.00	**1.00**/0.67	0.00/0.00	1.0/0.81	0.00/0.00

5 Conclusion and Future Works

We present an unsupervised learning framework for learning ViPILA. First, we formalize the latent action learning setup and establish the rationale for viewpoint-invariant latent actions in terms of spurious actions. Then, we propose a weighted viewpoint-invariance loss which is theoretically grounded but widely applicable. We empirically show that our loss function enforces the learned latent actions to remain consistent across different viewpoints for the same underlying state. Furthermore, we find that incorporating an unpaired dataset with a similarity-based pseudo-pairing mechanism improves the alignment of latent actions. In addition, we demonstrate the effectiveness of the ViPILA by conditioning the diffusion-based policy in simulation.

Limitations and future works. While our framework is applicable in principle to real-world scenarios, the current study evaluates its effectiveness only in the simulation. Applying ViPILA to real robot systems remains an important direction for future work. Additionally, we use fixed and heuristic frame intervals to extract latent actions. Leveraging an adaptive interval could improve the expressiveness and temporal alignment of latent actions. Lastly, prior work has demonstrated the utility of discretized latent actions in VLA models. Exploring ViPILA in the discrete setting is a promising avenue for future research.

Acknowledgments and Disclosure of Funding

We are grateful to Yerin Kim for designing the illustrations and providing valuable advice on their presentation and to Seungyub Han for writing. This work is in part supported by the National Research Foundation of Korea (NRF, RS-2024-00451435(20%), RS-2024-00413957(20%)),

Institute of Information & communications Technology Planning & Evaluation (IITP, RS-2021-II212068(10%), RS-2025-02305453(15%), RS-2025-02273157(15%), RS-2025-25442149(10%) RS-2021-II211343(10%)) grant funded by the Ministry of Science and ICT (MSIT), Institute of New Media and Communications(INMAC), and the BK21 FOUR program of the Education, Artificial Intelligence Graduate School Program (Seoul National University), and Research Program for Future ICT Pioneers, Seoul National University in 2025.

References

[1] Robert McCarthy, Daniel C. H. Tan, Dominik Schmidt, Fernando Acero, Nathan Herr, Yilun Du, Thomas G. Thuruthel, and Zhibin Li. Towards generalist robot learning from internet video: A survey, 2024.

[2] Dingzhe Li, Yixiang Jin, Yong A, Hongze Yu, Jun Shi, Xiaoshuai Hao, Peng Hao, Huaping Liu, Fuchun Sun, and Bin Fang. What foundation models can bring for robot learning in manipulation : A survey. *CoRR*, abs/2404.18201, 2024.

[3] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos, 2024.

[4] Hanjung Kim, Jaehyun Kang, Hyolim Kang, Meedeum Cho, Seon Joo Kim, and Youngwoon Lee. Uniskill: Imitating human videos via cross-embodiment skill representations. *arXiv preprint arXiv:2505.08787*, 2025.

[5] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

[6] Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, Jianyu Chen, and Jiang Bian. villa-x: Enhancing latent action modeling in vision-language-action models. *arXiv preprint arXiv: 2507.23682*, 2025.

[7] From One Hand to Multiple Hands: Imitation Learning for Dexterous Manipulation from Single-Camera Teleoperation. Qin, yuzhe and su, hao and wang, xiaolong, 2022.

[8] Homanga Bharadhwaj, Abhinav Gupta, Vikash Kumar, and Shubham Tulsiani. Towards generalizable zero-shot manipulation via translating human interaction plans, 2023.

[9] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. 2023.

[10] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision (ECCV)*, 2024.

[11] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning, 2023.

[12] Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[13] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments, 2024.

[14] Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Proceedings of the 1st Workshop and Challenge on Comprehensive Video Understanding in the Wild*, CoVieW'18, page 3, New York, NY, USA, 2018. Association for Computing Machinery.

[15] Pierre Sermanet, Corey Lynch, Jasmine Hsu, and Sergey Levine. Time-contrastive networks: Self-supervised learning from multi-view observation. *arXiv preprint arXiv:1704.06888*, 2017.

[16] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina González, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbeláez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shout, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19383–19400, 2024.

[17] Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. Viewpoint rosetta stone: Unlocking unpaired ego-exo videos for view-invariant representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[18] Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Muhammad Zubair Irshad, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. In *Conference on Robot Learning (CoRL)*, Munich, Germany, 2024.

[19] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation, 2022.

[20] Yifeng Zhu, Zhenyu Jiang, Peter Stone, and Yuke Zhu. Learning generalizable manipulation policies with object-centric 3d representations. In *7th Annual Conference on Robot Learning*, 2023.

[21] Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Exaug: Robot-conditioned navigation policies via geometric experience augmentation. *arXiv preprint arXiv:2210.07450*, 2022.

[22] Stephen Tian, Blake Wulfe, Kyle Sargent, Katherine Liu, Sergey Zakharov, Vitor Guizilini, and Jiajun Wu. View-invariant policy learning via zero-shot novel view synthesis. *arXiv*, 2024.

[23] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[24] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. *arXiv:2306.14896*, 2023.

[25] Danny Driess, Ingmar Schubert, Pete Florence, Yunzhu Li, and Marc Toussaint. Reinforcement learning with neural radiance fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[26] Dongseok Shim, Seungjae Lee, and H Jin Kim. Snerl: Semantic-aware neural radiance fields for reinforcement learning.

[27] David Ha and Jürgen Schmidhuber. World models. 2018.

[28] Jing-Cheng Pang, Nan Tang, kaiyuan Li, Yuting Tang, Xin-Qiang Cai, Zhen-Yu Zhang, Gang Niu, Sugiyama Masashi, and Yang Yu. Learning view-invariant world models for visual robotic manipulation. In *International Conference on Learning Representations (ICLR)*, 2025.

[29] Younggyo Seo, Junsu Kim, Stephen James, Kimin Lee, Jinwoo Shin, and Pieter Abbeel. Multi-view masked world models for visual robotic manipulation. *Proceedings of Machine Learning Research*, 202:30613–30632, 2023. Publisher Copyright: © 2023 Proceedings of Machine Learning Research. All rights reserved.; 40th International Conference on Machine Learning, ICML 2023 ; Conference date: 23-07-2023 Through 29-07-2023.

[30] Brian Sallans. Learning factored representations for partially observable markov decision processes. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[31] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[32] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[34] Mingxing Xu, Wenrui Dai, Chunmiao Liu, Xing Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint arXiv:2001.02908*, 2020.

[35] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

[36] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023.

[37] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, October 2021.

[38] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

[39] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense, 2017.

[40] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *arXiv preprint arXiv:2212.04501*, 2022.

[41] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[42] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[43] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.