2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

AMPLIFY: Actionless Motion Priors for Robot Learning from Videos

Jeremy A. Collins*[1], Loránd Cheng*[1], Kunal Aneja[1], Albert Wilcox[1], Benjamin Joffe[1,2], Animesh Garg[1]

***Abstract*— Action-labeled data for robotics is scarce and expensive, limiting the generalization of learned policies. In contrast, vast amounts of action-free video data are readily available, but translating these observations into effective policies remains a challenge. We introduce AMPLIFY, a framework that leverages large-scale video data by encoding visual dynamics into compact, discrete motion tokens derived from keypoint trajectories. Our modular approach separates visual motion prediction from action inference, decoupling the challenges of learning *what* motion defines a task from *how* robots can perform it. We train a forward dynamics model on abundant action-free videos and an inverse dynamics model on a limited set of action-labeled examples, allowing for independent scaling. Extensive evaluations demonstrate our dynamics model achieves over 2× better point track prediction accuracy compared to the prior state-of-the-art. In downstream policy learning, our dynamics predictions enable a 1.2-2.2× success rate improvement in low-data regimes, a 1.4× average improvement by learning from action-free human videos, and the first generalization to LIBERO tasks with zero in-distribution action data. Beyond robotic control, we find the latent dynamics learned by AMPLIFY to enhance video prediction quality. Our results present a novel paradigm leveraging heterogeneous data sources to build efficient, generalizable world models. More information can be found at amplify-robotics.github.io.**

I. INTRODUCTION

Recent successes in harnessing internet-scale data to train image and language foundation models have spurred an analogous push in robotics. In contrast with earlier methods that focused on achieving expert-level capabilities in narrow, controlled domains, recent efforts in robotics have aimed to generalize across tasks, object categories, object instances, environments, and the abundant variety of conditions present in the natural world [1], [2], [3], [4], [5], [6]. However, in order to train such generalist models, the typical behavior cloning (BC) approach requires prohibitively large amounts of action-labeled expert demonstrations. Datasets that are considered large-scale for robotics [1], [3], [7] take weeks or months to collect a few *hundred* hours of interaction data, falling far short of the roughly *one billion* hours of video data available on the internet. Therefore, methods that incorporate large-scale pre-training on these more abundant modalities tend to generalize better from limited action data [8], [9], [2]. Videos, in particular, contain rich priors on temporally-extended dynamics, behaviors, and semantics, which can be used to learn a predictive model of the world [10], [11], [12], [13], [14], [15], [16], [17] .

Prior work has leveraged video pre-training to learn representations using auxiliary tasks such as reward and value prediction [18], [19], [20], [21] or time-contrastive objectives [18], [22], [23]. While useful as representations, these methods only learn an encoder for static observations and do not explicitly model sequential dynamics. In contrast, model-based approaches can improve sample efficiency by separating the challenge of policy learning from learning dynamics [24]. Since videos contain rich priors over object and agent dynamics, model-based methods offer a promising avenue for learning from limited action data. One such approach is to train a full video prediction model to capture visual dynamics, which can act as a reference generator for downstream policies [10], [25]. However, predicting in pixel space is computationally intensive and costly to run at high frequencies, forcing these methods to make compromises like open-loop control [10] or partial denoising [25]. As a result, a number of works have aimed to learn *latent action* representations from videos using next-frame prediction [26], [27], [28] or latent consistency [29], efficiently modeling features that are predictive of the future. While this avoids high inference costs, these representations are still trained on image reconstruction/prediction objectives, capturing textural details or visually salient features that may not be relevant to policy learning.

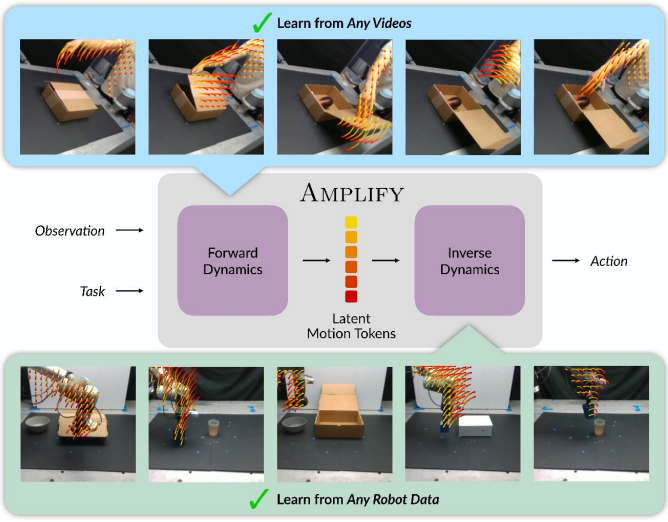


Fig. 1: Overview. AMPLIFY decomposes policy learning into forward and inverse dynamics, using latent keypoint motion as an intermediate representation. The forward model can be trained on *any* video data, while the inverse model can be trained on *any* interaction data. In contrast with behavior cloning (BC), AMPLIFY requires fewer demonstrations, can generalize to tasks for which we have *zero* action data, and can learn from human videos.

Motivated by the desire to capture motion rather than appearance, optical flow and keypoint tracking have emerged as appealing abstractions for extracting action information from videos without action labels. Recent advances in

* Equal contribution. [1]Georgia Tech [2]Georgia Tech Research Institute

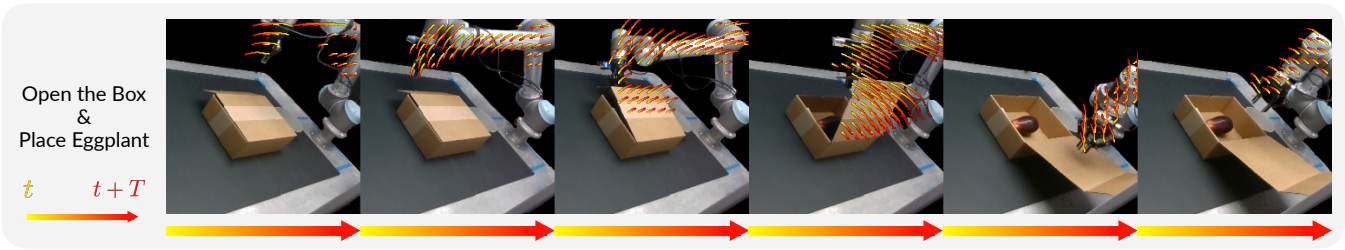


Fig. 2: Decoded keypoint trajectory predictions from AMPLIFY. Predicted zero-movement points are not shown.

computer vision have enabled efficient and precise pixel-level point tracking, even through occlusions and limited out-of-frame tracking [30], [31], [32], [33]. As these capabilities enable fine-grained capture of motion and scene dynamics, they have found applications in robotics for visual imitation learning [34] and tool use [35]. A number of prior works predict motion from images as optical flow [36], [37], [38] or by modeling the trajectories of specified keypoints [39], [40], [41], [42], [43], [44], [45], [46], [47]. However, many of these works still rely on prohibitively expensive video prediction models [38], [45], [48], [49], [50], object-centric mask extraction [41], [45], [51], calibrated cameras [44], or inefficient online planning [42], limiting their generality.

Two of the most general keypoint modeling approaches are ATM [52] and Track2Act [51], which aim to learn a universal keypoint dynamics model to predict the future trajectories of arbitrary points in an image, and condition a policy on these predictions. However, Track2Act relies on the often unrealistic assumption of a goal image and restricts its output space to single-object rigid-body transformations. ATM, while more flexible in its representation, relies on unrealistic point-sampling heuristics during training that cannot be replicated during inference. In addition, neither ATM nor Track2Act learn a latent abstraction of keypoints, leaving them with high computational costs much like pixel-space video generation and potentially hindering generalization. Due to their high computational costs, Track2Act requires open-loop trajectory generation, and ATM only generates tracks for 32 points during policy inference, resulting in very coarse dynamics predictions.

In this paper, we investigate the use of *latent* keypoint motion as an abstraction for learning valuable action priors from action-free video data, combining the benefits of latent dynamics prediction with the explicit motion information captured in keypoint trajectories. We propose AMPLIFY: Actionless Motion Priors for Learning Inverse and Forward Dynamics, a three-stage framework that flexibly decouples dynamics modeling from policy learning. First, we learn a compact latent space for modeling the motion of a dense grid of keypoints. Second, we train a latent dynamics model to predict a sequence of latent motions based on the current observation. Finally, an inverse dynamics model learns to map predicted latent motions to low-level robot actions for execution. Notably, this modular approach allows the first two stages to be trained on *any* video data, while the inverse dynamics policy can be trained on *any* interaction data (Figure 1). We show that this has profound implications for policy generalization in Section III-B.

Through extensive real-world and simulated experiments, we evaluate both the accuracy and downstream utility of our latent dynamics model. Compared to state-of-the-art baselines, we observe that AMPLIFY leads to improved keypoint trajectory prediction, lowering mean-squared error by over $3\times$. We then demonstrate that these predictions are useful for control; conditioning the inverse dynamics policy on latent motions is a valuable prior that allows for more data-efficient learning and generalization to tasks for which we have *no action-labeled data*. Finally, we examine the versatility of our motion-based representations beyond control for tasks such as conditional video prediction.

In summary, we make the following key contributions:

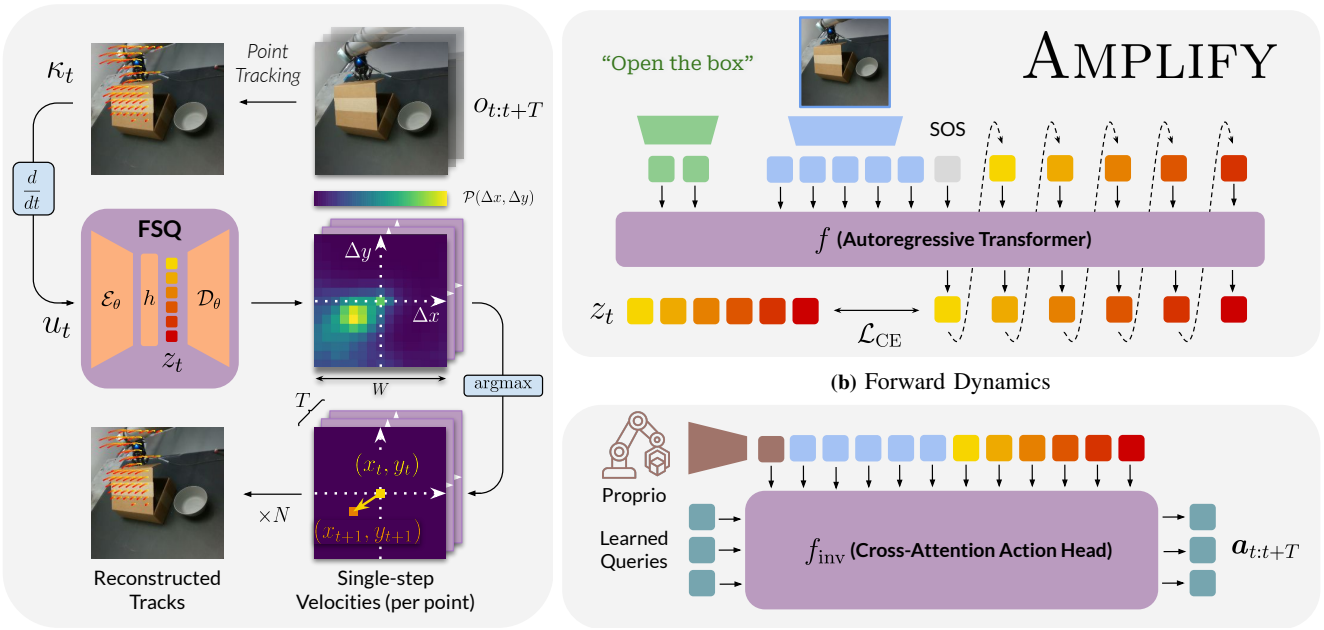
1) We present the first *latent* keypoint dynamics model and investigate crucial design choices.
2) We train a data-efficient and generalizable policy that can learn from action-free human data, outperforming SOTA BC baselines in few-shot settings.
3) We demonstrate the first generalization to LIBERO tasks with zero in-distribution action data, achieving a 60% success rate on held-out tasks while our best baseline achieves 2% success.
4) We demonstrate state-of-the-art keypoint prediction accuracy on three large-scale video datasets.
5) We apply latent motions to conditional video generation, outperforming baselines.

II. METHOD

A. Problem Setup

We assume access to two types of data: a video dataset $\mathcal{V} = \{(o_t, g)\}$ and a dataset of robot interaction data $\mathcal{R} = \{(o_t, q_t, a_t)\}$ where $o \in \mathcal{O}$ are RGB image observations, $g \in \mathcal{G}$ is a goal (e.g., a language description), and $a \in \mathcal{A}, q \in \mathcal{Q}$ are the action and proprioceptive state of the robot, respectively[1]. Given these datasets, our aim is to learn the parameters of a visual control policy $\pi : \mathcal{O} \times \mathcal{Q} \times \mathcal{G} \to \mathcal{A} = f_{\text{inv}}(o_t, q_t, f(o_t, g))$ composed of a forward dynamics model $f : \mathcal{O} \times \mathcal{G} \to \mathcal{Z}$ that learns a *motion prior* in a latent space $\mathcal{Z}$ and an inverse dynamics model $f_{\text{inv}} : \mathcal{O} \times \mathcal{Q} \times \mathcal{Z} \to \mathcal{A}$ that maps the latent motion to a sequence of actions. Crucially, this decomposition allows for independent scaling of f and f_{inv} by training on $\mathcal{V}$ and $\mathcal{R}$, respectively. The following sections detail preprocessing (Sec. II-B), learning the latent

[1] $\mathcal{V}$ and $\mathcal{R}$ need not be disjoint in general, and any goal-directed interaction data (demonstrations) may be included in both $\mathcal{V}$ and $\mathcal{R}$. However, $\mathcal{V}$ may additionally contain non-robot videos and $\mathcal{R}$ may contain undirected action data such as exploration or play data.



(a) Motion Tokenization

(b) Forward Dynamics

(c) Inverse Dynamics

Fig. 3: Architecture. AMPLIFY consists of a three-stage decomposition: (a) a uniform grid of keypoint tracks are compressed into a discrete latent space using FSQ. For each timestep and each point, the decoder outputs a distribution in a local window centered around each point to reconstruct the instantaneous velocities, (b) a forward dynamics model is trained to predict the latent codes for the next T timesteps given an input image and task description, and (c) an inverse dynamics model decodes predicted latent motion tokens into an action chunk.

motion representation (Sec. II-C), and training the forward (Sec. II-D) and inverse (Sec. II-E) dynamics models.

B. Preprocessing Keypoint Tracks

We first augment $\mathcal{V} \rightarrow \mathcal{V}' = \{(o_t, \kappa_t, g)\}$ in a preprocessing step using the off-the-shelf point tracking model from [30] to obtain a set of keypoint tracks $\kappa_t \in \mathbb{R}^{T \times N \times 2}$ for each timestep t. More precisely, we initialize a 20×20 uniform grid of $N = 400$ points in each image o_t, then track the points through the next $T = 16$ frames $o_{t:t+T}$, capturing their 2-dimensional pixel coordinates. Although extracting specific task-relevant keypoints could potentially yield more informative predictions, we favor the uniform grid for its simplicity and generality, similar to [51], and find that it works effectively to model a variety of motions. Other works have attempted to select keypoints according to heuristics such as movement throughout the video [52], but we found that this led the model to learn spurious correlations and relies on unrealistic assumptions at test time. By reinitializing the grid of keypoints in each frame, we ensure no points are occluded and guarantee consistent coverage throughout every frame, even with moving cameras.

C. Motion Tokenization

Unlike prior keypoint-based methods which predict directly in pixel space [52], [51], [42], [45], we argue that learning to predict dynamics in a compressed latent space enables a more efficient and generalizable representation, similar to findings in model-based reinforcement learning [53], [54], [55]. To this end, we learn a compact discrete latent space from pre-processed keypoint trajectories using Finite Scalar Quantization (FSQ) [56], a drop-in replacement for vector-quantization [57]. FSQ employs an implicit codebook without auxiliary loss terms, avoiding representation collapse and resulting in better codebook utilization.

Figure 3a illustrates our tokenization scheme. We compute single-step velocities $u_t \in \mathbb{R}^{(T-1) \times N \times 2}$ from the pre-processed keypoint trajectories κ_t. Then, a keypoint encoder $\mathcal{E}_\theta : \mathbb{R}^{(T-1) \times N \times 2} \rightarrow \mathbb{R}^{b \times d}$ maps u_t to a b-length sequence $\tilde{z}_t$ of latent vectors $\tilde{z}_{t,i} \in \mathbb{R}^d$, which are quantized via FSQ to a sequence $z_t \in \mathbb{Z}^{b \times d}$ of discrete codes, and decoded by the keypoint decoder $\mathcal{D}_\theta : \mathbb{Z}^{b \times d} \rightarrow \mathbb{R}^{(T-1) \times N \times W^2}$ for reconstruction. Rather than just predicting the 2-dimensional pixel coordinate of each point directly, the decoder outputs a categorical distribution over W^2 classes representing a local $W \times W$ window of motions centered at the same point in the previous timestep. This imposes an inductive bias on the model toward next-keypoint predictions that are close to locations in the current timestep, and additionally better captures multimodal distributions compared to performing regression on the coordinates. In practice, W can be chosen based on the measured maximum point track velocity in the dataset. The keypoint encoder has a causally-masked transformer encoder architecture, and the keypoint decoder is an unmasked transformer decoder that cross-attends from a sequence of $T - 1$ learned positional encodings to the quantized codes from the encoder. Its outputs are then expanded across tracks and combined with learned track and view embeddings via a residual MLP. The encoder and decoder are jointly trained on $\mathcal{V}$ using a cross-entropy loss:

$$\mathcal{L}_{AE}(\theta) = \text{CE}\Big(\mathcal{D}_\theta\Big(h(\mathcal{E}_\theta(u_t))\Big), \omega_t\Big) \tag{1}$$

where $\omega_t = \Omega(u_t)$, $\Omega : \mathbb{R}^{(T-1) \times N \times 2} \rightarrow \mathbb{R}^{(T-1) \times N \times W^2}$ maps ground-truth pixel-space velocities to their correspond-

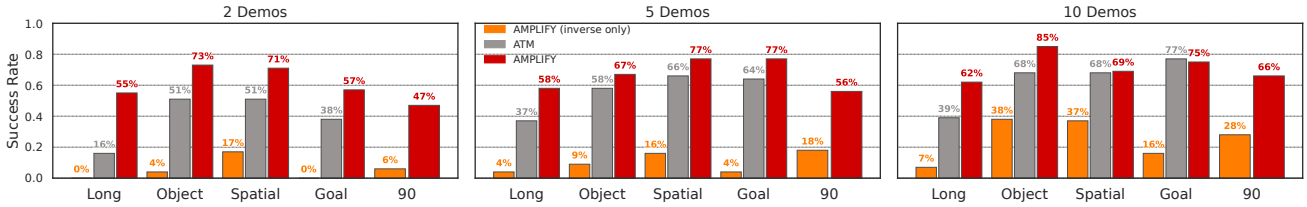


Fig. 4: LIBERO few-shot. Comparison of AMPLIFY against ATM [52] and a no-video-pre-training baseline. Our forward model is trained on all videos, and the inverse model is only trained on a limited number of demos. Success rates for all three methods are shown, except LIBERO-90, for which ATM does not report results.

ing class based on the displacement in the local $W \times W$ window, and h is the FSQ discretization function. When available, multi-view inputs are tokenized jointly into a single sequence of b codes. For simplicity, we do not include the view dimension in our notation.

D. Forward Dynamics (Actionless Motion Prior)

After training the motion tokenizer, we train an autoregressive transformer $f(o_t, g)$ to predict the tokenized motion sequence z_t corresponding to the video $o_{t:t+T}$ based on the current observation and task description. Image observations are encoded and projected into the embedding space of the transformer using the flattened feature map from a pre-trained ResNet-18 [58] to generate $7 \times 7 = 49$ vision tokens per image. The task description is encoded with T5-Small [59], and its summary token is used as the language token. These conditioning tokens are then concatenated with a start of sequence (SOS) token and the latent motion tokens to predict the next tokens in the sequence (Figure 3b). A block-causal attention mask is used, where the conditioning tokens are non-causal and the motion tokens are causally masked. We use a cross-entropy loss on the predicted codes without decoding to full keypoint trajectories, and only back-propagate gradients to the dynamics model while the tokenizer remains frozen (Equation 2). $\mathtt{sg}$ refers to the stop-gradient operator.

$$\mathcal{L}_{\text{forward}} = \text{CE}\Big(f(o_t, g), \mathtt{sg}(h(\mathcal{E}_\theta(u_t)))\Big) \tag{2}$$

E. Inverse Dynamics

Finally, we learn an inverse dynamics model $f_{\text{inv}}(o_t, q_t, z_t)$ that decodes latent motion tokens into a distribution over action chunks $\boldsymbol{a}_t = a_{t:t+T}$, as shown in Figure 3c. Importantly, this module is not task-conditioned and instead acts as a general reference follower trained on any interaction data $\mathcal{R}$. The model uses a transformer decoder with a sequence of learned tokens that cross-attend to image tokens, a linear projection of proprioceptive state, and motion tokens to produce a sequence of m action tokens. These action tokens are fed into an action head to output a distribution over length-T action chunks. Following BAKU [60], we opt for an isotropic Gaussian prior on the action distribution. The inverse dynamics model is trained with a negative log-likelihood (NLL) loss with a temporal discount γ to reduce the impact of predictions towards the end of the sequence.

$$\mathcal{L}_{\text{inv}} = -\sum_{\tau=t}^{t+T-1} \gamma^{\tau-t} \cdot \log p\left(a_\tau \mid \mu_{\tau-t}, \sigma_{\tau-t}\right) \tag{3}$$

where $\mu_{\tau-t} = f^{\mu}_{\text{inv}}(o_t, q_t, z_t)[\tau - t]$ and $\sigma_{\tau-t} = \exp(f^{\sigma}_{\text{inv}}(o_t, q_t, z_t)[\tau - t])$ are the predicted mean and standard deviation. The inverse dynamics model can be trained on ground truth quantized tokens $z_t = h(\mathcal{E}_\theta(u_t))$, but in practice, we observe improved performance when conditioning the action decoder on the predicted outputs $\hat{z}_t$ of the forward dynamics model during training. Both the motion tokenizer and the forward dynamics model are frozen for this stage. The keypoint decoder $\mathcal{D}_\theta$ is not used, as we condition f_{inv} on latent motions rather than decoded tracks.

F. Inference

During inference, the forward dynamics model takes the current observation and task at each timestep t and autoregressively predicts a sequence of latent motion tokens $\hat{z}_t = f(o_t, g)$. The inverse dynamics model then ingests these tokens, along with image and proprioception tokens, into an action chunk $\boldsymbol{a}_t = f_{\text{inv}}(o_t, q_t, \hat{z}_t)$. Following ACT [61], we use temporal ensembling to aggregate information over predicted action chunks with the same temporal discount γ.

III. Experiments

We evaluate AMPLIFY guided by two main axes of investigation: **quality** of dynamics prediction (Sec. III-A) and **utility** of predictions for policy learning (Sec. III-B) and conditional video generation (Sec. III-C). All models were trained on a single NVIDIA L40S GPU for 24–48 hours, and inference runs at ≈ 10 Hz on a single RTX 3090.

A. Quality of Forward Dynamics Prediction

We test the prediction accuracy of our forward dynamics model on a combination of three simulated and real-world video datasets, including both human and robot data: BridgeData v2 [62], a large-scale robot dataset consisting of over 60k real-world rollouts of diverse manipulation tasks in 24 different environments; Something-Something v2 [63], a video dataset consisting of over 220,000 videos of humans performing everyday manipulation tasks with a variety of objects and primitive motion categories; and LIBERO [64], a benchmark of 130 diverse simulated robotic manipulation tasks, from which we use the observations from 6500 demonstration rollouts as a video dataset.

We compare to ATM [52] and Track2Act [51], two state-of-the-art keypoint trajectory prediction approaches. For fair comparison with ATM, we adapt its evaluation setup to use the same 20×20 uniform grid of 400 points used by AMPLIFY. In addition, on BridgeData v2 we compare track prediction accuracy to a baseline of first predicting videos with Seer

TABLE I: Dynamics Prediction. AMPLIFY achieves 3.7× better MSE and 2.5× better pixel accuracy compared to ATM, and a 4-26% improvement over Track2Act, which uses a goal image, and Seer, which requires full video prediction.

Method	LIBERO			Bridge	SSv2
Metric	MSE ↓	Δ_{AUC} ↑	Pixel Acc. ↑	Δ_{AUC} ↑	Δ_{AUC} ↑
ATM [52]	0.022	0.767	0.250	–	–
Track2Act [51]	–	–	–	0.770	0.700
Seer [65]	–	–	–	0.914	–
AMPLIFY	**0.006**	**0.913**	**0.629**	**0.968**	**0.725**

[65], then applying CoTracker [30] to the initial set of points and tracking through the generated videos. Since our forward dynamics model predicts in latent space, we use the decoder from the Motion Tokenization stage for fair comparison in pixel space. We measure performance on normalized tracks ($\kappa \in [-1, 1]$) using (1) Mean Squared Error (MSE); (2) Pixel-Wise Accuracy (Pixel Acc.), which measures the percentage of predictions that are pixel-perfect compared to ground-truth; and (3) Δ_{AUC}, originally used by point tracking methods [32], [30], and later used for track point prediction by Track2Act.

Results are summarized in Table I, demonstrating that AMPLIFY consistently leads to more accurate predictions, even though the forward dynamics model is only trained on a latent consistency loss rather than pixel-space prediction objectives. On the LIBERO dataset, we achieve over twice the pixel-wise accuracy of ATM, and we outperform Track2Act (which, unlike our method, has access to goal images) on their chosen Δ_{AUC} metric across BridgeData v2 and Something-Something v2. We attribute this success to several design choices, including the compression of motion into a compact latent space, thus improving efficiency and generalization; the prediction of discrete tokens to leverage the expressive power of autoregressive transformers; and the use of local-window pixel space classification, which gives our forward dynamics model the ability to model rich multi-modal distributions of motion and capture fine-grained dynamics.

B. Utility of Predicted Latent Motions for Policy Learning

Beyond prediction accuracy, we examine whether video pre-training using AMPLIFY can provide a useful prior for policy learning in both real-world and simulated experiments. Specifically, we evaluate AMPLIFY along four dimensions measuring (1) few-shot learning, (2) cross-embodiment transfer, (3) generalization, and (4) in-distribution performance. We evaluate performance using success rates on all five subsets of LIBERO, as well as a set of 3 real-world tasks: "Put the Rubik's Cube on the Box" (`Place Cube`), "Stack the Green and Blue Cups in the Orange Cup" (`Stack Cups`), and "Open the Box and Move the Eggplant into the Bowl" (`Open Box & Place Eggplant`).

Few-Shot Learning – We study whether AMPLIFY can learn from fewer action-labeled demonstrations by training the forward model on all videos, while the inverse model is only trained on 4%, 10%, or 20% of the 50 demonstrations available for each of the subsets of LIBERO. In Figure 4, we compare AMPLIFY with ATM, trained on all videos and the same subsets of action data, as well as a variant of AMPLIFY that does not condition on motion tokens to predict actions. Both AMPLIFY and ATM consistently outperform the no-pre-training variant, indicating that in low-data regimes, video pre-training on keypoint dynamics provides a strong prior for data-efficient policy learning. In addition, AMPLIFY achieves stronger performance than ATM on nearly every subset, suggesting that a latent motion representation has higher utility for action prediction than conditioning the policy directly on pixel-space track predictions. This seems to be especially true at the extreme low end–when provided with only 2 demonstrations per task, AMPLIFY achieves an average $1.94\times$ improvement over ATM.

TABLE II: Cross-Embodiment Transfer. By leveraging human video demonstrations to train the forward dynamics model, AMPLIFY outperforms Diffusion Policy on real-world tasks.

Method	Place Cube			Stack Cups			Box/Eggplant			Avg.
	5	10	All	5	10	All	5	10	All	
Diffusion Policy [66]	0.6	0.5	**0.9**	**0.3**	0.5	0.5	**0.1**	0.2	0.2	0.42
AMPLIFY (DP head)	**0.7**	**0.9**	**0.9**	**0.3**	**0.6**	**1.0**	**0.1**	**0.3**	**0.4**	**0.58**

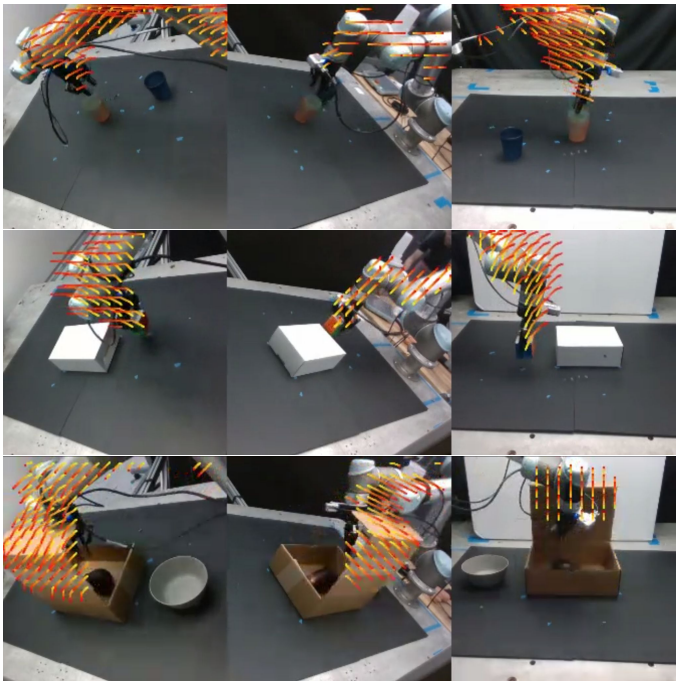


Fig. 5: Decoded predictions from AMPLIFY. Rows show different tasks and columns show different camera viewpoints at the same time step. We use three static RGB cameras as input observations for both human and robot data. Because AMPLIFY jointly tokenizes across viewpoints, its predictions generally exhibit 3D consistency.

Cross-Embodiment Transfer – Since the forward dynamics model can be trained on any observation data, we study whether videos of humans demonstrating a task can be used to improve policy learning. We train the forward dynamics model on both human and robot video data, while the inverse dynamics model is trained only on the action-labeled robot data. This setup highlights how the two stages can be decoupled to scale independently, unlike behavior cloning methods that cannot effectively harness action-free data. We evaluate success rates on three real-world tasks of varying

TABLE III: Zero-shot task generalization from LIBERO 90 to unseen LIBERO subsets. We are the first to report non-trivial success on LIBERO without using *any* action data from the target tasks. Compared to the best BC baseline, AMPLIFY provides a 27× average improvement.

Method	LIBERO Long	LIBERO Object	LIBERO Spatial	LIBERO Goal
Diffusion Policy [66]	0.00	0.00	0.00	0.00
QueST [67]	0.07	0.00	0.01	0.01
BAKU [60]	0.06	0.00	0.00	0.00
AMPLIFY (w/o tracks)	0.00	0.00	0.00	0.02
AMPLIFY	**0.52**	**0.80**	**0.69**	**0.41**

TABLE IV: In-Distribution performance on LIBERO. When provided with full action data, AMPLIFY remains competitive with various state-of-the-art baselines.

Method	Long	90	Object	Spatial	Goal
Diffusion Policy [66]	0.73	0.67	0.70	0.79	0.83
QueST [67]	0.67	0.89	–	–	–
BAKU [60]	**0.86**	**0.90**	–	–	–
UniPi [10]	0.06	–	0.60	0.69	0.12
ATM [52]	0.44	0.63	0.81	0.79	0.59
AMPLIFY (IDM only)	0.76	0.83	0.64	**0.83**	**0.92**
AMPLIFY (Full)	0.77	**0.90**	**0.93**	0.77	**0.92**

difficulty, using Diffusion Policy as the BC baseline. We use a UR5 robotic arm equipped with three static RGB camera views (Figure 5). For fair comparison, we replace the Gaussian head used in other experiments with a Diffusion Policy head in the inverse dynamics model. This ensures that the only difference between the two approaches is whether the predictions from our forward dynamics model are used to condition the policy. Similarly to the previous section, we evaluate AMPLIFY in both the few-shot setting and the full data setting. Results in Table II demonstrate that AMPLIFY can effectively leverage additional human data to learn common dynamics between human and robot motions, and use the predicted latent motions to improve policy learning. The average improvements of $1.32\times$, $1.4\times$, and $1.5\times$ indicate a more prominent gap as task complexity increases.

Generalization – Observing that AMPLIFY excels in learning from *limited* action data, we now turn to a setting where *no* action data is available for target tasks. Given *only observations* of target tasks, as well as a dataset of out-of-distribution interaction data, we evaluate how well AMPLIFY can solve the target tasks zero-shot. This challenging setting requires methods to both learn a good abstraction of the mapping from observations to actions, and also generalize that abstraction to predict correct actions on new tasks. To test this setting, we train the forward dynamics model on videos and text from all subsets of LIBERO, and train the inverse dynamics model and BC baselines on actions from *only* LIBERO 90. We then evaluate on four LIBERO target suites (Long, Object, Spatial, Goal), specifically designed to test different categories of generalization [64]. We find that BC methods completely fail in this scenario, achieving near-zero success rates (Table III). We attribute this failure to two main shortcomings of BC: (1) the supervised imitation objective has no incentive to learn generalizable representations, and (2) BC has no mechanism for harnessing additional data that may be informative, such as videos. In contrast, AMPLIFY attains an average 60.5% success rate on target tasks, approaching the success rates of models that were directly trained on the target tasks. This success highlights the value of latent dynamics prediction as a versatile interface for learning general priors from action-free videos. In addition, it suggests that training a general reference following inverse dynamics model may be a more generalizable objective compared to imitation learning.

TABLE V: Video Prediction. Conditioning AVDC [38] on predicted motion tokens from our dynamics model improves generated video quality on BridgeData v2.

Method	PSNR ↑	LPIPS ↓	SSIM ↑
AVDC [38]	15.93	**0.16**	0.56
AVDC + AMPLIFY	**16.40**	0.19	**0.59**

In-Distribution Performance – We evaluate AMPLIFY in an in-distribution setting, training both the forward and inverse dynamics models on only demonstration data. We compare to state-of-the-art approaches with and without video pre-training. Results in Table IV indicate that AMPLIFY, even without additional video data, is competitive with SOTA behavior cloning methods and outperforms video pre-training methods trained with (ATM) and without (UniPi) keypoint tracks. In this setting, we observe that since there is sufficient information to learn tasks to a high degree *without* video pre-training, standard BC methods tend to match or outperform approaches using pre-training. However, few-shot and task generalization experiments demonstrate that prior approaches underperform in limited data regimes and do not generalize effectively to new tasks, while AMPLIFY remains performant.

C. Utility of Predicted Latent Motions for Conditional Video Generation

To demonstrate the utility of predicting keypoint trajectories beyond robotic control, we condition a video prediction model [38] on the latent motion tokens predicted by our forward dynamics model. We find that conditioning a video prediction model on our latent motion tokens leads to a modest improvement in generation quality (Table V). Compared to a baseline model that does not use track inputs, our approach yields better performance on two of three metrics. This improvement suggests that our latent motion representation captures rich, structured dynamics that improve not only control tasks but also the fidelity of generated videos.

D. Ablations

We summarize key ablations in Table VI. For the Motion Tokenizer, replacing point track coordinate regression with a local-window classification objective yields a clear gain in track quality. This supports the hypothesis that discrete, local prediction better captures fine-grained multi-modal motion than direct regression, which tends to undershoot due to the prevalence of zero-motion in the data and averaging between modes. Prediction horizon shows a consistent, stage-dependent trade-off. Shorter horizons help representation and prediction: The Motion Tokenizer's Δ_{AUC} is highest at 4 steps and declines as the horizon grows, and the Forward

TABLE VI: Ablations on LIBERO-10. Chosen setting and best result in **bold**.

Motion Tokenizer		
Horizon	4	**0.985** Δ_{AUC}
	8	0.961 Δ_{AUC}
	16	0.919 Δ_{AUC}
Decoder loss	MSE reg.	0.883 Δ_{AUC}
	Local-window cls.	**0.919** Δ_{AUC}
Codebook size	512	0.912 Δ_{AUC}
	1024	0.919 Δ_{AUC}
	2048	**0.921** Δ_{AUC}
Forward Dynamics Model		
Horizon	4	**0.757** Pixel Acc.
	8	0.678 Pixel Acc.
	16	0.613 Pixel Acc.
Vision encoder	**ResNet-18**	0.613 Pixel Acc.
	ResNet-50	**0.621** Pixel Acc.
	DINOv2	**0.621** Pixel Acc.
Inverse Dynamics Model		
Horizon	4	0.36 Succ. Rate
	8	0.64 Succ. Rate
	16	**0.75** Succ. Rate
Action head	**Gaussian (Transformer)**	**0.74** Succ. Rate
	Diffusion (U-Net)	**0.74** Succ. Rate
	Flow Matching (DiT)	0.73 Succ. Rate

Dynamics Model shows the same trend in Pixel Accuracy. In contrast, longer horizons help control: the Inverse Dynamics Model achieves its best Success Rate at 16 steps, indicating that longer action chunks provide useful disambiguation for action recovery. Increasing Motion Tokenizer codebook size yields diminishing returns; Δ_{AUC} rises as the codebook size is changed from 512 to 2048, suggesting that larger vocabularies are helpful but not significantly so. The choice of vision backbone has a comparatively small effect; ResNet-18 trails ResNet-50 and DINOv2 by only a narrow margin, indicating limited sensitivity to encoder capacity for the evaluated tasks. Finally, action-head choice produces near-parity: a simple Gaussian policy head reaches 74% Success Rate, matching a diffusion U-Net (74%) and slightly exceeding flow matching with a DiT head (73%).

IV. Conclusion

In this work, we introduced AMPLIFY, a framework that leverages large-scale action-free video data and a small amount of interaction data to significantly enhance robotic policy performance. By decoupling the learning of *what* constitutes a task from *how* to execute it, our approach efficiently utilizes heterogeneous data sources. Our key insight lies in representing scene dynamics through compact latent motion tokens derived from keypoint trajectories, which enables higher efficiency and improved performance compared to pixel-level reconstruction methods. Experimental results show that AMPLIFY consistently outperforms baselines in the limited action data regime and in zero-shot generalization settings. Moreover, the versatility of our latent representation extends beyond control, proving useful in tasks such as conditional video prediction. Our findings demonstrate the promise of harnessing large-scale human video data to inform robotic control policies and pave the way for more scalable, generalizable, and efficient robot learning.

References

[1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn *et al.*, "Rt-1: Robotics transformer for real-world control at scale."

[2] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[3] A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," *arXiv preprint arXiv:2310.08864*, 2023.

[4] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn *et al.*, "pi0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[5] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong *et al.*, "Fast: Efficient action tokenization for vision-language-action models," *arXiv preprint arXiv:2501.09747*, 2025.

[6] L. X. Shi, B. Ichter, M. Equi, L. Ke, K. Pertsch, Q. Vuong *et al.*, "Hi robot: Open-ended instruction following with hierarchical vision-language-action models," *arXiv preprint arXiv:2502.19417*, 2025.

[7] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," 2024.

[8] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair *et al.*, "OpenVLA: An Open-Source Vision-Language-Action Model," Jun. 2024, arXiv:2406.09246 [cs]. [Online]. Available: http://arxiv.org/abs/2406.09246

[9] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn *et al.*, "pi_0: A Vision-Language-Action Flow Model for General Robot Control," Nov. 2024, arXiv:2410.24164. [Online]. Available: http://arxiv.org/abs/2410.24164

[10] Y. Du, M. Yang, B. Dai, H. Dai, O. Nachum, J. B. Tenenbaum *et al.*, "Learning Universal Policies via Text-Guided Video Generation," Nov. 2023, arXiv:2302.00111 [cs]. [Online]. Available: http://arxiv.org/abs/2302.00111

[11] R. McCarthy, D. C. H. Tan, D. Schmidt, F. Acero, N. Herr, Y. Du *et al.*, "Towards Generalist Robot Learning from Internet Video: A Survey," Jun. 2024, arXiv:2404.19664 [cs]. [Online]. Available: http://arxiv.org/abs/2404.19664

[12] O. Rybkin, K. Pertsch, K. G. Derpanis, K. Daniilidis, and A. Jaegle, "Learning what you can do before doing anything," Feb. 2019, arXiv:1806.09655 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1806.09655

[13] NVIDIA, N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker *et al.*, "Cosmos World Foundation Model Platform for Physical AI," Mar. 2025, arXiv:2501.03575 [cs]. [Online]. Available: http://arxiv.org/abs/2501.03575

[14] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko *et al.*, "Imagen video: High definition video generation with diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2210.02303

[15] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou *et al.*, "Hunyuanvideo: A systematic framework for large video generative models," *arXiv preprint arXiv:2412.03603*, 2024.

[16] Veo-Team, :, A. Gupta, A. Razavi, A. Toor, A. Gupta *et al.*, "Veo 2," 2024. [Online]. Available: https://deepmind.google/technologies/veo/veo-2/

[17] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada *et al.*, "Lumiere: A space-time diffusion model for video generation," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.

[18] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "Vip: Towards universal visual reward and representation via value-implicit pre-training," *arXiv preprint arXiv:2210.00030*, 2022.

[19] D. Ghosh, C. Bhateja, and S. Levine, "Reinforcement Learning from Passive Data via Latent Intentions," Apr. 2023, arXiv:2304.04782 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2304.04782

[20] C. Bhateja, D. Guo, D. Ghosh, A. Singh, M. Tomar, Q. Vuong *et al.*, "Robotic Offline RL from Internet Videos via Value-Function Pre-Training," Sep. 2023, arXiv:2309.13041 [cs]. [Online]. Available: http://arxiv.org/abs/2309.13041

[21] N. Dashora, D. Ghosh, and S. Levine, "Viva: Video-trained value functions for guiding online rl from diverse data," *arXiv preprint arXiv:2503.18210*, 2025.

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

[23] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," *arXiv preprint arXiv:2203.12601*, 2022.

[24] T. M. Moerland, J. Broekens, A. Plaat, C. M. Jonker *et al.*, "Model-based reinforcement learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 16, no. 1, pp. 1–118, 2023.

[25] Y. Hu, Y. Guo, P. Wang, X. Chen, Y.-J. Wang, J. Zhang *et al.*, "Video prediction policy: A generalist robot policy with predictive visual representations," *arXiv preprint arXiv:2412.14803*, 2024.

[26] J. Bruce, M. D. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes *et al.*, "Genie: Generative interactive environments," in *Forty-first International Conference on Machine Learning*, 2024.

[27] S. Ye, J. Jang, B. Jeon, S. Joo, J. Yang, B. Peng *et al.*, "Latent action pretraining from videos," *arXiv preprint arXiv:2410.11758*, 2024.

[28] Y. Chen, Y. Ge, Y. Li, Y. Ge, M. Ding, Y. Shan *et al.*, "Moto: Latent Motion Token as the Bridging Language for Robot Manipulation," Dec. 2024, arXiv:2412.04445 [cs]. [Online]. Available: http://arxiv.org/abs/2412.04445

[29] Z. Cui, H. Pan, A. Iyer, S. Haldar, and L. Pinto, "Dynamo: In-domain dynamics pretraining for visuo-motor control," *Advances in Neural Information Processing Systems*, vol. 37, pp. 33 933–33 961, 2024.

[30] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht, "CoTracker: It is better to track together," 2023.

[31] Q. Wang, Y.-Y. Chang, R. Cai, Z. Li, B. Hariharan, A. Holynski *et al.*, "Tracking everything everywhere all at once," in *International Conference on Computer Vision*, 2023.

[32] C. Doersch, Y. Yang, M. Vecerik, D. Gokay, A. Gupta, Y. Aytar *et al.*, "Tapir: Tracking any point with per-frame initialization and temporal refinement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 061–10 072.

[33] Y. Xiao, Q. Wang, S. Zhang, N. Xue, S. Peng, Y. Shen *et al.*, "Spatialtracker: Tracking any 2d pixels in 3d space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[34] M. Vecerik, C. Doersch, Y. Yang, T. Davchev, Y. Aytar, G. Zhou *et al.*, "RoboTAP: Tracking Arbitrary Points for Few-Shot Visual Imitation," Aug. 2023, arXiv:2308.15975 [cs]. [Online]. Available: http://arxiv.org/abs/2308.15975

[35] Z. Qin, K. Fang, Y. Zhu, L. Fei-Fei, and S. Savarese, "Keto: Learning keypoint representations for tool manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 7278–7285.

[36] C. Yuan, C. Wen, T. Zhang, and Y. Gao, "General flow as foundation affordance for scalable robot learning," *arXiv preprint arXiv:2401.11439*, 2024.

[37] L.-H. Lin, Y. Cui, A. Xie, T. Hua, and D. Sadigh, "FlowRetrieval: Flow-Guided Data Retrieval for Few-Shot Imitation Learning," Oct. 2024, arXiv:2408.16944. [Online]. Available: http://arxiv.org/abs/2408.16944

[38] P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, and J. B. Tenenbaum, "Learning to Act from Actionless Videos through Dense Correspondences," *arXiv:2310.08576*, 2023.

[39] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang *et al.*, "Egomimic: Scaling imitation learning via egocentric video," 2024. [Online]. Available: https://arxiv.org/abs/2410.24221

[40] J. Gao, Z. Tao, N. Jaquier, and T. Asfour, "K-VIL: Keypoints-Based Visual Imitation Learning," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3888–3908, Oct. 2023, conference Name: IEEE Transactions on Robotics. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10189175

[41] X. Fang, B.-R. Huang, J. Mao, J. Shone, J. B. Tenenbaum, T. Lozano-Pérez *et al.*, "Keypoint Abstraction using Large Models for Object-Relative Imitation Learning," Oct. 2024, arXiv:2410.23254. [Online]. Available: http://arxiv.org/abs/2410.23254

[42] C. Gao, H. Zhang, Z. Xu, C. Zhehao, and L. Shao, "Flip: Flow-centric generative planning as general-purpose manipulation world model," in *ICLR*, 2025.

[43] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, "kpam: Keypoint affordances for category-level robotic manipulation," in *The International Symposium of Robotics Research*. Springer, 2019, pp. 132–157.

[44] I. Guzey, Y. Dai, G. Savva, R. Bhirangi, and L. Pinto, "Bridging the Human to Robot Dexterity Gap through Object-Oriented Rewards," Oct. 2024, arXiv:2410.23289. [Online]. Available: http://arxiv.org/abs/2410.23289

[45] M. Xu, Z. Xu, Y. Xu, C. Chi, G. Wetzstein, M. Veloso *et al.*, "Flow as the Cross-Domain Manipulation Interface," Jul. 2024, arXiv:2407.15208 [cs]. [Online]. Available: http://arxiv.org/abs/2407.15208

[46] S. Haldar and L. Pinto, "Point policy: Unifying observations and actions with key points for robot manipulation," *arXiv preprint arXiv:2502.20391*, 2025.

[47] V. Liu, A. Adeniji, H. Zhan, S. Haldar, R. Bhirangi, P. Abbeel *et al.*, "Egozero: Robot learning from smart glasses," 2025. [Online]. Available: https://arxiv.org/abs/2505.20290

[48] H. Bharadhwaj, D. Dwibedi, A. Gupta, S. Tulsiani, C. Doersch, T. Xiao *et al.*, "Gen2Act: Human Video Generation in Novel Scenarios enables Generalizable Robot Manipulation," Sep. 2024. [Online]. Available: https://arxiv.org/abs/2409.16283v1

[49] S. Li, Y. Gao, D. Sadigh, and S. Song, "Unified video action model," *arXiv preprint arXiv:2503.00200*, 2025.

[50] C. Zhu, R. Yu, S. Feng, B. Burchfiel, P. Shah, and A. Gupta, "Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets," *arXiv preprint arXiv:2504.02792*, 2025.

[51] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation," *arXiv preprint arXiv:2405.01527*, 2024.

[52] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao *et al.*, "Any-point Trajectory Modeling for Policy Learning," Feb. 2024, arXiv:2401.00025 [cs]. [Online]. Available: http://arxiv.org/abs/2401.00025

[53] N. Hansen, X. Wang, and H. Su, "Temporal difference learning for model predictive control," in *ICML*, 2022.

[54] N. Hansen, H. Su, and X. Wang, "TD-MPC2: Scalable, Robust World Models for Continuous Control," Mar. 2024, arXiv:2310.16828 [cs]. [Online]. Available: http://arxiv.org/abs/2310.16828

[55] A. Scannell, M. Nakhaei, K. Kujanpää, Y. Zhao, K. S. Luck, A. Solin *et al.*, "Discrete codebook world models for continuous control," *arXiv preprint arXiv:2503.00653*, 2025.

[56] F. Mentzer, D. Minnen, E. Agustsson, and M. Tschannen, "Finite Scalar Quantization: VQ-VAE Made Simple," Oct. 2023, arXiv:2309.15505 [cs]. [Online]. Available: http://arxiv.org/abs/2309.15505

[57] A. v. d. Oord, O. Vinyals, and K. Kavukcuoglu, "Neural Discrete Representation Learning," May 2018, arXiv:1711.00937 [cs]. [Online]. Available: http://arxiv.org/abs/1711.00937

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, iSSN: 1063-6919.

[59] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[60] S. Haldar, Z. Peng, and L. Pinto, "Baku: An efficient transformer for multi-task policy learning," *arXiv preprint arXiv:2406.07539*, 2024.

[61] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware."

[62] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning*. PMLR, 2023, pp. 1723–1736.

[63] R. Goyal, S. E. Kahou, V. Michalski, J. Materzyńska, S. Westphal, H. Kim *et al.*, "The "something something" video database for learning and evaluating visual common sense," 2017. [Online]. Available: https://arxiv.org/abs/1706.04261

[64] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu *et al.*, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *arXiv preprint arXiv:2306.03310*, 2023.

[65] X. Gu, C. Wen, W. Ye, J. Song, and Y. Gao, "Seer: Language instructed video prediction with latent diffusion models," *arXiv preprint arXiv:2303.14897*, 2023.

[66] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," *arXiv preprint arXiv:2303.04137*, 2023.

[67] A. Mete, H. Xue, A. Wilcox, Y. Chen, and A. Garg, "QueST: Self-Supervised Skill Abstractions for Learning Continuous Control," Sep. 2024, arXiv:2407.15840 [cs]. [Online]. Available: http://arxiv.org/abs/2407.15840