Video Prediction Policy: A Generalist Robot Policy with Predictive Visual Representations

Anonymous CVPR submission

Paper ID 9874

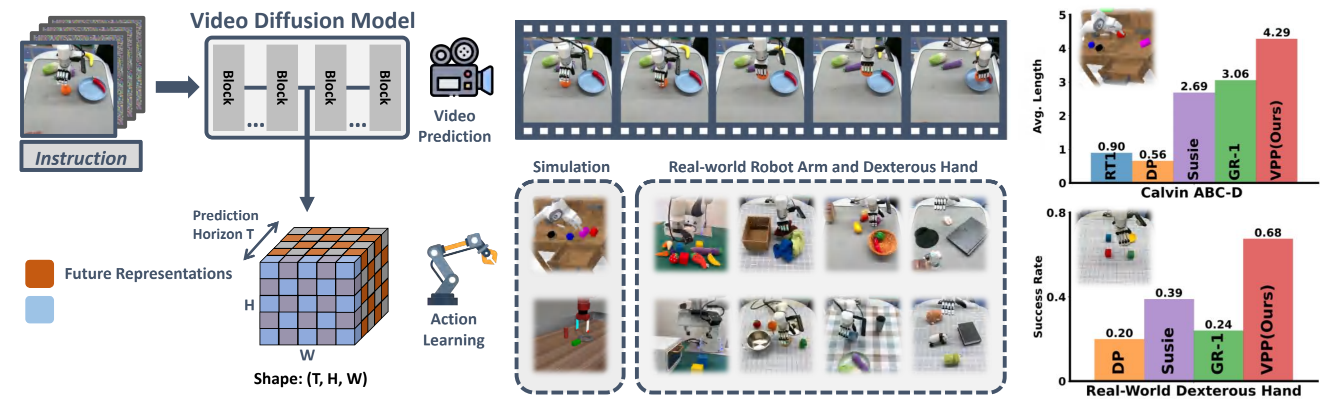


Figure 1. We observe that the visual representations within the video diffusion model explicitly capture both current and predicted future information. Our Video Prediction Policy, built on these representations, achieves consistent improvements across four benchmarks.

Abstract

Recent advancements in robotics have focused on developing generalist policies capable of performing multiple tasks. Typically, these policies utilize pre-trained vision encoders to capture crucial information from current observations. However, previous vision encoders, which trained on two-image contrastive learning or single-image reconstruction, can not perfectly capture the sequential information essential for embodied tasks. Recently, video diffusion models (VDMs) have demonstrated the capability to accurately predict future image sequences, exhibiting a good understanding of physical dynamics. Motivated by the strong visual prediction capabilities of VDMs, we hypothesize that they inherently possess visual representations that reflect the evolution of the physical world, which we term predictive visual representations. Building on this hypothesis, we propose the Video Prediction Policy (VPP), a generalist robotic policy conditioned on the predictive visual representations from VDMs. To further enhance these representations, we incorporate diverse human or robotic manipulation datasets, employing unified video-generation training objectives. VPP consistently outperforms existing methods across two simulated and two real-world benchmarks. Notably, it achieves a 28.1% relative improvement in the Calvin ABC-D benchmark compared to the previous state-of-the-art and delivers a 28.8% increase in success rates for complex real-world dexterous manipulation tasks.

1. Introduction

Building generalist robot policies capable of solving multiple tasks is an active area of research [8, 36]. Two essential components for constructing such generalist policies are action networks and vision encoders. One line of research focused on developing more advanced action networks, such as employing visual-language pre-trained models [7, 8, 28, 31, 58], training from scratch on diverse robotic datasets [49], incorporating auto-regressive [8] or diffusion architectures [16], and scaling up action networks [33]. Another line of work focuses on learning more effective visual representations [29, 41] for embodied tasks from egocentric video datasets [20, 21] via contrastive learning [45] or image reconstruction [24].

In this paper, we focus on the visual representation learning. We observe that previous vision encoders, which are pre-trained using contrastive learning between two frames or single-frame reconstruction, fail to adequately capture the physical dynamics inherent in sequential video datasets. Recently, powerful video diffusion models (VDMs) [6, 10,

CVPR
#9874

26, 27, 56], trained with direct video generation objectives on much larger datasets, have demonstrated the ability to generate continuous image sequences and exhibit a strong understanding of the physical world. Inspired by the strong prediction capabilities of VDMs, we hypothesize that they can better capture the physical dynamics within video datasets and inherently contain valuable visual representations that reflect the dynamics and evolution of objects. Moreover, we observe that the visual representations within VDMs are structured with shape (T, H, W), explicitly representing 1 current step and $(T-1)$ predicted future steps, where H and W correspond to the height and width of single image representation. In contrast, previous vision encoders do not explicitly capture future representations. A comparison is visualized in Figure 2. Based on this distinction, we refer to these latent variables within the video diffusion model as “predictive visual representations”. In the experiment part, we also visualize these predictive representations and find they contain valuable temporal information that reflects the evolution of the physical world.

Our key insight is that these predictive visual representations are highly informative for downstream action learning, as they capture the movement of objects, including the robot itself. Moreover, the ability to predict can be learned from both internet-scale video datasets and various robotic datasets using a consistent video generation loss, enabling us to transfer physical knowledge from large-scale internet datasets to specific robotic systems.

Building on this insight, we introduce the **V**ideo **P**rediction **P**olicy (**VPP**), which employs a two-stage learning process: First, we finetune a text-guided video prediction (TVP) model [14, 22] from pre-trained video diffusion model [6] using various manipulation datasets, including ego-centric human manipulation [20], open-source robotic datasets [42], and self-collected robot data. This training aims to obtain a controllable video generation model that enhances prediction capabilities in the manipulation domain. Second, we develop a multi-task generalist robot policy conditioned on the predictive representations within the TVP model. Given that the predictive representations in the TVP model remain high-dimensional, with the shape (T, H, W), we employ a video former to distill essential information across spatial and temporal dimensions, followed by a widely used diffusion policy [16] to output actions.

In experiments, our Video Prediction Policy (VPP) consistently outperform other baseline algorithms across two simulated [39, 57] and two real-world settings, demonstrating the effectiveness of our approach. Notably, the VPP achieves a 28.1% improvement in the Calvin ABC→D benchmark [39] compared to the previous SOTA method [30]. Additionally, VPP shows a 28.8% improvement in success rate over the strongest baseline, Susie [5], in complex real-world scenarios involving dexterous hand manipulation. Our contributions can be summarized as follows:

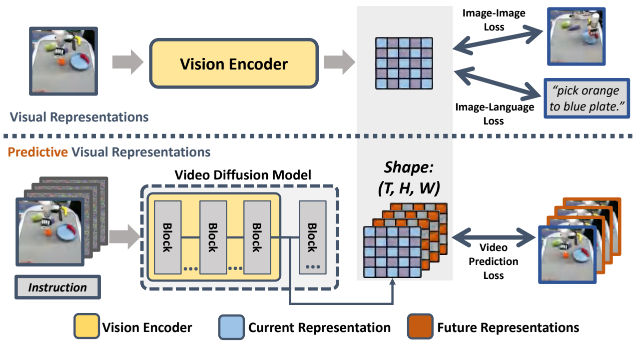


Figure 2. We use the video diffusion model as a vision encoder to obtain the predictive representations that explicitly express both current and sequential future frames. Previous vision encoders did not have explicit future representations.

1. To the best of our knowledge, we are the first to leverage the visual representations inside video diffusion models. We find that these representations explicitly express predicted future frames, which we refer to as “predictive visual representations”.
2. We introduce a novel generalist robotic policy, the Video Prediction Policy, by fine-tuning a TVP model in the manipulation domain and then learning actions conditioned on predictive visual presentations in the TVP model.
3. We demonstrate the superior performance of our approach in both simulated and real-world environments, highlighting its versatility.

2. Related Works

Visual Representation Learning for Robotics. Self-supervised learning (SSL) techniques, such as contrastive [13, 15], distillation-based [2, 11], and reconstructive [3, 24], have achieved significant advancements in visual representation learning. Prior research has shown that these SSL techniques enable vision encoders to produce effective representations for embodied AI tasks [12, 43, 46, 54, 55], capturing both high-level semantic and low-level spatial information. Notably, methods like R3M [41], vip [37], VC-1 [38], and Voltron [29] have specifically focused on embodied tasks by innovating pre-training approaches on human manipulation video datasets [20, 21]. However, regardless of the training objective, the learned vision encoders primarily focus on extracting pertinent information from current observations without explicitly predicting future states. In contrast, our Video Prediction Policy leverages predictive representations within video prediction models to explicitly encapsulate both current and predicted future frames.

Future Prediction for Embodied Control Tasks. Existing research also explores the use of future prediction to enhance policy learning [4, 5, 18, 51]. For example, SuSIE [5] conditions its control policy on a predicted future keyframe generated by InstructPix2Pix [9], while UniPi [18] learns the inverse dynamics between two generated frames. These methods typically rely on a single future prediction step to determine actions, which may not accurately capture the complexities of physical dynamics. Additionally, they often operate in raw pixel space, which contains much irrelevant information. GR-1 [51] generates subsequent frames and actions in an autoregressive manner. However, it only generates one image per forward pass, and its prediction quality lags behind that of diffusion-based methods. Furthermore, GR-1 does not leverage pre-trained video foundation models. In contrast, VPP leverages an intermediate representation fine-tuned from a pre-trained video diffusion model, which captures continuous future trajectories to more effectively inform policy learning.

Visual Representation inside Diffusion Models. Diffusion models have achieved remarkable success in the image and video generation tasks [6, 48]. Typically trained as denoisers, diffusion models predict original images from noisy inputs [25]. Research has shown that **image diffusion models** can also function effectively as vision encoders [23, 34, 53], generating meaningful visual representations. These representations have been proven to be linear-separable for discrimination tasks [53], invaluable for semantic segmentation [34], and versatile for embodied tasks [23]. However, the capabilities of representations within **video diffusion models** have not been extensively explored. Our findings suggest that variables within VDMs have a unique predictive property not present in other visual representations, making them especially useful for sequential embodied control tasks.

3. Preliminaries

Video Diffusion Models. The core idea of diffusion models is to continuously add Gaussian noise to make video sequences a Gaussian and leverage the denoising process for generating videos. Let x_0 represent a real video sample, the forward process aims to add Gaussian noise and result in a set of noisy data, i.e., $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbb{I})$, where x_t and α_t indicate the noisy data and noise amplitude at the timestep t. Let $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, the above process can be simplified as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t \, . \tag{1}$$

The reverse process starts from the most noisy sample x_T can be described in a variational approximation of the probabilities $q(x_{t-1}|x_t)$, as follows:

$$p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\mu_\theta(x_t, t), (1-\bar{\alpha}_{t-1})\mathbb{I}). \tag{2}$$

where $\mu_\theta(x_t, t) = (x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t))/\sqrt{\bar{\alpha}_t}$ is a learnable neural network to estimate x_{t-1}. Further, in text-guided video generation, the denoising process learns the noise estimator $\epsilon_\theta(x_t, c)$ to approximate the score function $\sqrt{1-\bar{\alpha}_t}\nabla_{x_t} \log p_\psi(x_t|c)$, controlling the video generation based on the initial frame and language prompt.

Diffusion Policy. The diffusion model has also proven effective in action learning, known as diffusion policy [16]. The diffusion policy aims to denoise the action sequence $a_i = (\widehat{a}_i, \widehat{a}_{i+1}, ..., \widehat{a}_{i+m})$ based on observations s_i and instruction. Chi et al. [16] point out that diffusion policy is capable of expressing complex multimodal action distributions and stabilizing training. Recent work [47] further enhances the diffusion policy by incorporating the advanced diffusion transformer (DiT) block [44], a technique we also adopt in the Video Prediction Policy to improve performance.

4. Video Prediction Policy

In this section, we describe the two-stage learning process of the Video Prediction Policy, shown in Figure 3. Initially, we train the Text-guided Video Prediction (TVP) model across diverse manipulation datasets to harness physical knowledge from internet data; subsequently, we design networks to aggregate predictive visual representations inside the TVP model and output final robot actions.

4.1. Text-guided Video Prediction (TVP) Model for Robot Manipulation.

Recent advancements have focused on training general video generation models using extensive online video datasets, which encode abundant prior knowledge about the physical world's dynamics. However, we notice that these models are not fully controllable and fail to yield optimal results in specialized domains such as robot manipulation. To address this, we fine-tune the general video generation model into a specialized "Manipulation TVP Model" to enhance prediction accuracy.

We chose the open-sourced Stable Video Diffusion (SVD) model [6] with 1.5 billion parameters as our foundation. we observe that the open-sourced SVD model conditions only on initial-frame images s_0 without incorporating language instructions l. We augment the model to incorporate CLIP [45] language features l_{emb} using cross-attention layers. Furthermore, we adjust the output video resolution to 16×256×256 to optimize training and inference efficiency. Despite these modifications, we preserve the other components of the original pre-trained SVD framework to retain its core capabilities. We denote this modified version as V_θ. In this setup, the initial observation s_0 is concatenated channel-wise with each predicted frame as a condition. Then model V_θ is trained with diffusion objective, re-

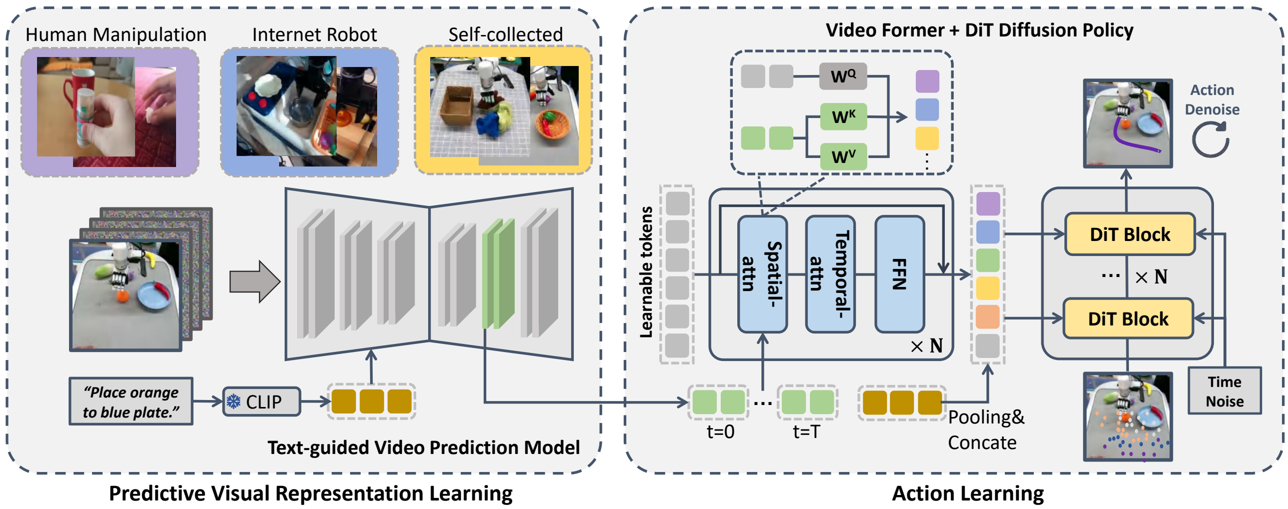


Figure 3. Video Prediciton Policy first trains a text-guided video prediction (TVP) model for manipulation domain, starting from pre-trained video foundation model. Subsequently, it learns actions based on the predictive representations internal to the TVP model.

constructing the full video sequence $x_0 = s_{0:T}$ in dataset D from noised samples $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$:

$$\mathcal{L}_D = \mathbb{E}_{x_0 \sim D, \epsilon, t} \| V_\theta(x_t, l_{emb}, s_0) - x_0 \|^2 \tag{3}$$

The video prediction objective offers a unified interface that directly generates future visual sequences, enabling the TVP model to harness physical knowledge from diverse datasets. These include internet-based human manipulation datasets D_H, publicly available robot manipulation data D_R, and also self-collected datasets D_C. Given the varying quality and scale of these datasets, we introduce specific coefficients λ to appropriately balance the influence of different dataset types:

$$\mathcal{L}_{video} = \lambda_H \mathcal{L}_{D_H} + \lambda_R \mathcal{L}_{D_R} + \lambda_C \mathcal{L}_{D_C} \tag{4}$$

Then we froze the fine-tuned manipulation TVP models in downstream action learning.

4.2. Action Learning Conditioned on Predictive Visual Representation

TVP Model as Vision Encoder. After training the TVP model specifically for manipulation tasks, it can accurately predict future sequences based on image observations and instructions. However, denoising an entire video sequence is highly time-consuming and may lead to open-loop control issues, as discussed in [18]. Moreover, videos in their original pixel format often contain excessive, irrelevant information that can interfere with effective decision-making.

To address these concerns, we employ the video diffusion model primarily as a "vision encoder" rather than a "denoiser" by performing only a single forward step. Our insight is that the first forward step, while not yielding a clear video, still provides a rough trajectory of future states and valuable guidance. This insight is verified in our experiment section and shown in Fig 5. Specifically, we concatenate the current image s_0 with the final noised latent $q(x_{t'}|x_0)$ (typically white noise) and input this combination into the TVP model. We then directly utilize the latent features $F_m \in \mathbb{R}^{T \times W \times H \times C}$ in m^{th} layer of the video diffusion model V_θ:

$$F_m = V_\theta(x_{t'}, l_{emb}, s_0)_{(m)} \tag{5}$$

For a robot with multiple camera views, such as a third-view and a wristed camera, we predict the future for each view independently, denoted as $F_m^{static}, F_m^{wrist}$.

Video Former. These predictive representations within the video diffusion model are still high-dimensional, as they express a sequence of image features. To efficiently aggregate representations across spatial, temporal, and multi-view dimensions, we use a Video Former to consolidate this information into a fixed number of tokens. The Video Former initializes $T \times L$ learnable tokens $Q_{[0:T,0:L]}$, performing spatial-temporal attention on each corresponding frame in the predictive representations, followed by feed-forward layers. Formally, this branch can be expressed as follows where i is the index of frame:

$$\begin{aligned} Q' &= \{\text{Spat-Attn}(Q[i], (F_m^{static}[i], F_m^{wrist}[i]))\}_{i=0}^{T} \\ Q'' &= \text{FFN}(\text{Temp-Attn}(Q')). \end{aligned} \tag{6}$$

Action Generation. After the Video-Former aggregates the Predictive feature into learnable tokens Q'', a diffusion policy is employed as the action head to generate the action sequence $a_0 \in A$ based on Q''. We integrate the aggregated presentation Q'' into diffusion transformer blocks using cross-attention layers. The diffusion policy aims to reconstruct the original actions a_0 from noised action $a_k = \sqrt{\bar{\beta}_k}a_0 + \sqrt{1-\bar{\beta}_k}\epsilon$, where ϵ represents white noise, and $\bar{\beta}_k$ is the noisy coefficient at step k. This step can be interpreted as learning a denoiser D_ψ to approximate the noise ϵ and minimize the following loss function:

$$\mathcal{L}_{\text{diff}}(\psi; A) = \mathbb{E}_{a_0,\epsilon,k}\|D_\psi(a_k, l_{emb}, Q'') - a_0\|^2 \quad (7)$$

In real-world dexterous hand manipulation tasks, where $a = \{a^{xyz} \in R^3, a^{rot} \in R^3, a^{finger} \in R^{12}\}$,we use coefficients to balance the loss contributions from end-effector movement, rotational actions, and finger movements. Therefore, the optimization loss function for the diffusion policy can be written as:

$$\begin{aligned}\mathcal{L}_{\text{policy}}(\psi; A) =&\omega_{xyz}\mathcal{L}_{\text{diff}}(\psi; a^{xyz}) + \omega_{rot}\mathcal{L}_{\text{diff}}(\psi; a^{rot}) \\ &+\omega_{finger}\mathcal{L}_{\text{diff}}(\psi; a^{finger})\end{aligned} \quad (8)$$

5. Experiments

In this section, we conduct extensive experiments on both simulated and real-world robotic tasks to evaluate the performance of the video prediction policy (VPP). The simulated environments include the CALVIN benchmark [39] and MetaWorld benchmark [57], while the real-world tasks encompass Panda arm manipulation and XHand dexterous hand manipulation. Our aim to answer the following questions:

1. Can VPP achieve a higher success rate in manipulation tasks with predictive visual representations?
2. How do the video pre-training and internet manipulation datasets enhance the performance of VPP?
3. How does predictive representation compare to previous visual representations?
4. Which layer of the video diffusion model provides the most effective predictive visual representations?

5.1. Simulated Benchmarks Experiments

Environmental Setups. We consider the CALVIN [39] and MetaWorld [57] simulated environments. CALVIN is a challenging benchmark focused on evaluating the instruction-following capability of robotic policies for long-horizon manipulations. As depicted on the left side of Figure 4, it encompasses four environments, denoted ABCD. We utilize the most challenging ABC→D setting, where robots are trained with standard datasets collected from environments ABC and tested in the unseen environment D. MetaWorld features a Sawyer robot performing various manipulation tasks and is widely used to evaluate the precision and dexterity of robotic policies. As shown on the right of Figure 4, it includes 50 tasks with a rich array of operating objects at different levels of difficulty [46]. We collected 50 trajectories for each task using the official Oracle policy as our training dataset.

Baselines. We mainly consider two types of baselines, methods with direct action learning and methods related to future prediction:

- RT-1 [7]. A direct action learning robot policy that integrates semantic information using Efficient-Net with FiLM-conditioning, followed by token learners for action learning.
- Diffusion Policy [16]. A direct action learning policy with novel action diffusers.
- Robo-Flamingo [32]. A direct action learning policy that leverages a pre-trained LLM, incorporating visual information into each layer in a flamingo style [1].
- Uni-Pi [18]. Begins by learning a video prediction model to generate future sequences and then learns an inverse kinematics model between two frames to determine actions.
- MDT [47]. Learns a diffusion transformer policy along with an auxiliary mae loss to reconstruct one masked future frame.
- Susie [5]. Uses a fine-tuned InstructPix2Pix [9] model to generate a goal image and learns a downstream diffusion policy conditioned on the goal image.
- GR-1 [51]. Learns video and action sequences jointly using an auto-regressive transformer. During policy execution, GR-1 outputs one future frame followed by one action.

Additionally, we include the 3D Diffuser Actor [30] baseline on the Calvin benchmark, as it is the previous state-of-the-art method on this benchmark, although it additionally uses depth image with camera pose unlike other methods.

Video Prediction Policy Training Details. We first train a controllable text-guided video prediction model for the manipulation domain on various datasets as described in Figure 3. Our experiments include 193,690 human manipulation trajectories from the Something-Something-V2 datasets [20] and 179,074 high-quality trajectories from internet robotic manipulation datasets [7, 17, 19, 28, 40, 42]. This stage also includes downstream task datasets, such as the official Calvin ABC dataset and Metaworld dataset, and self-collected datasets on real-world robots. Given the varying scales and quality of different robot datasets, we apply varying sampling probabilities similar to the approach used in [49]. Detailed dataset scales and sample ratios are available in the Appendix 2. The video model training process takes two days on eight NVIDIA A100 GPUs. Subsequent action learning for each robot takes approximately 6-12 hours on four NVIDIA A100 GPUs.

Category	Method	Annotated Data	Tasks completed in a row					
			1	2	3	4	5	Avg. Len ↑
Direct Action Learning Method	RT-1 [7]	100%ABC	0.533	0.222	0.094	0.038	0.013	0.90
	Diffusion Policy [16]	100%ABC	0.402	0.123	0.026	0.008	0.00	0.56
	Robo-Flamingo [32]	100%ABC	0.824	0.619	0.466	0.331	0.235	2.47
Future Prediction Related Method	Uni-Pi [18]	100%ABC	0.560	0.160	0.080	0.080	0.040	0.92
	MDT [47]	100%ABC	0.631	0.429	0.247	0.151	0.091	1.55
	Susie [5]	100%ABC	0.870	0.690	0.490	0.380	0.260	2.69
	GR-1 [51]	100%ABC	0.854	0.712	0.596	0.497	0.401	3.06
3D Method	3D Diffuser Actor [30]	100%ABC	0.938	0.803	0.662	0.533	0.412	3.35
Ours	**VPP (ours)**	100%ABC	**0.957**	**0.912**	**0.863**	**0.810**	**0.750**	**4.29**
Data Efficiency	MDT [47]	10%ABC	0.408	0.131	0.034	0.008	0.001	0.58
	GR-1 [51]	10%ABC	0.672	0.371	0.198	0.108	0.069	1.41
	VPP (ours)	10%ABC	**0.878**	**0.746**	**0.632**	**0.540**	**0.453**	**3.25**

Table 1. Zero-shot long-horizon evaluation on the Calvin ABC→D benchmark where agent is asked to complete five chained tasks sequentially. The Video Prediction Policy demonstrates a significant improvement in the average task completion length.

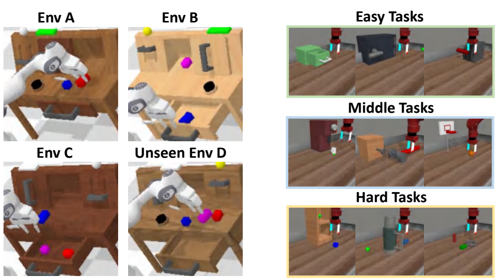


Figure 4. CALVIN and Metaworld benchmarks.

Task Level (Numbers)	Easy (28 tasks)	Middle (11 tasks)	Hard (11 tasks)	Average ↑ (50 tasks)
RT-1	0.605	0.042	0.015	0.346
Diffusion Policy	0.442	0.062	0.095	0.279
Susie	0.560	0.196	0.255	0.410
GR-1	0.725	0.327	0.451	0.574
VPP (ours)	**0.818**	**0.493**	**0.526**	**0.682**

Table 2. Success rate on 50 Metworld tasks which require precise control.

Video Prediction Policy Execution Details. To enhance the control frequency of robots, we assign most of the parameters to the video former part, which has approximately 300M parameters, while the diffusion policy head contains only 20M parameters. The policy execution involves running the video diffusion model and video former for one forward step, and the lightweight diffusion transformer policy denoises the action for 10 steps conditioned on learnable tokens. This design allows us to run the entire video prediction policy process at 7-10 Hz on a local machine equipped with an NVIDIA RTX-4090 GPU. Following the original diffusion policy paper [16], we also output 6∼10 action steps in one VPP forward step, further improving control frequency.

Quantitative Results. The comparisons on the Calvin benchmark are shown in Table 1. Results for Robo-Flamingo, Susie, GR-1, and 3D Diffuser Actors are recorded from their original papers. The MDT result is run on official implementation. The RT-1 result is sourced from [32] and the Uni-Pi result from [5]. We also ran the Diffusion Policy based on the official open-source codebase with CLIP language conditions. Our proposed Video Prediction Policy significantly improved the previous state-of-the-art result from an average task completion length of 3.35 to 4.29 without using any point cloud or depth input. Even with only 10% of the annotated Calvin ABC data used for training, our method still achieved a length of 3.25, which exceeds the results of related methods using full data. Furthermore, the Video Prediction Policy also achieved the best performance in the MetaWorld benchmark with 50 tasks, outperforming the strongest GR-1 baseline by 10.8% in average success rate.

5.2. Analysis of Predictive Visual Representations

Our video prediction policy has achieved significant improvements in simulated experiments with predictive representations. In this part, we conduct various experiments to verify the effectiveness of these predictive representations.

Visualizations of Predictive Representations. Since we use the video prediction model as a vision encoder and perform a single forward pass to obtain predictive representations, we are curious about the quality of these representations. In Figure 5 , we visualize the ground truth future, single-step predictions, and 30-step denoised predictions. Although the single-step prediction does not capture every detail with perfect accuracy, it still conveys valuable information related to robotic manipulation, such as the movement of objects and the robot arm, which effectively supports downstream action learning.

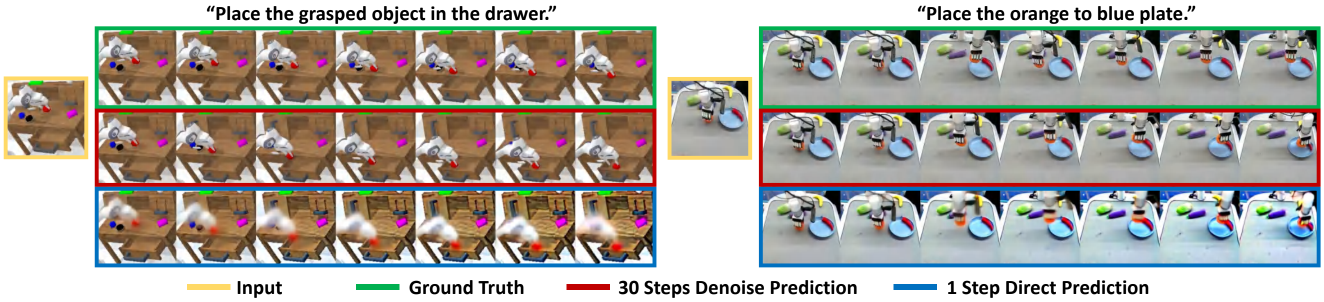


Figure 5. Visualization of the ground-truth video, the complete denoised video, and one-step forward video predictions. Although the textures and details are not precise in the one-step forward videos, they still provide valuable information on physical evolution.

Bridge	VideoFusion	Tune-A-Video	Seer	VPP
FVD↓	501.2	515.7	246.3	**41.4**

Table 3. Quantitative evaluation of prediction quality on bridge datasets. The results of VideoFusion [35], Tune-A-Video [52], Seer [22] are copied from [22].

Encoder	Pre-training Type	Avg. Length ↑
Video Prediction Diffusion Model	Video Generation	**4.29**
Stable-VAE	VAE Reconstruction	2.58
VC-1	MAE Reconstruction	1.23
Voltron	MAE Reconstruction+ Language Generation	1.54

Table 4. Ablation study on different visual representations.

Ablation Type	Average Length ↑
VPP	**4.29**
VPP w/o Internet data	3.97
VPP w/o Internet data w/o SVD Pretrain	1.63

Table 5. Ablation study on video pre-training and internet manipulation datasets.

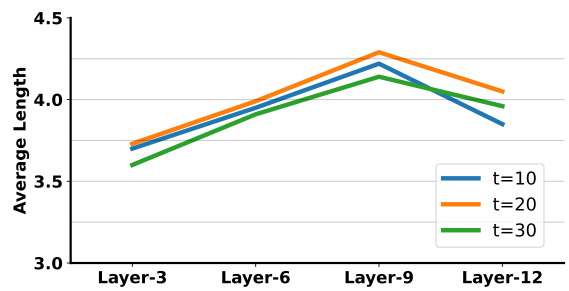


Figure 6. Influences of layer positions and initial noise scales.

Prediction Quality of Manipulation TVP Model. Additionally, we evaluate the quantitative FVD metric [50] on the bridge datasets [19] with complete 30 steps denoising as in [22]. The results are shown in Table 3. Surprisingly, our model easily outperforms the previous TVP model. We attribute this improvement to our use of the pre-trained video foundation model SVD [6], which the earlier TVP model did not leverage, giving us a significant advantage.

Comparisons with Other Visual Representations. To assess our predictive visual representations, we replaced them with alternative visual representations while maintaining other components of the Video Prediction Policy (VPP) unchanged. We considered visual representations pre-trained for different purposes: (1) Stable-VAE [6] pre-trained with VAE image reconstruction loss; (2) VC-1 [38] pre-trained with masked autoencoder loss, tailored for embodied tasks. According to the original study, we finetuned VC-1 on the Calvin datasets using MAE loss to better adapt to the new domain; (3) Voltron [29] pretrained with both MAE reconstruction and language generation tasks. The results, presented in Table 4, indicate that replacing our predictive visual representations leads to a clear decline in performance.

Effectiveness of Video Pre-training and Internet Manipulation Datasets. A significant advantage of the VPP is its ability to leverage the physical knowledge encoded in pre-trained video generation models and Internet manipulation datasets. We conducted experiments to verify the effectiveness of these two components. As shown in Table 5, removing the co-trained Internet manipulation data resulted in a performance decrease from 4.29 to 3.97. Further removing the pre-trained SVD model and training the video prediction model on the Calvin data from scratch led to a substantial performance decline.

Influence of Layer Position and Initial Noise Scales. We are also interested in how different layers of representation and initial white noise scales influence the predictive representations. We experimented with representations from different upsample layers and various initial white noise by altering the total diffusion time-step t, following [53]. The results are shown in Figure 6. Our findings suggest that the

Figure 7. Two real-world hardware platforms and visualizations of sampled tasks. In the Panda arm platform, our experiments include 30+ tasks of 6 categories. In the Xhand dexterous platform, our experiments include 100+ tasks of 10 categories.

most effective predictive representations are located in the middle of the upsample blocks rather than the final prediction pixels. Additionally, the quality of representation is not sensitive to initial noise scales.

5.3. Real World Experiments

We further verified the Video Prediction Policy on two real-world hardware platforms:

- **Franka Panda Robot Arm.** On the Franka panda platform, we collected 2k trajectories for over 30+ tasks of 6 categories including picking, placing, pressing, routing, opening, and closing.
- **Xarm with 12-degree Xhand Dexterous Hand.** On the dexterous hand platform, we collected 2.5k trajectories over 100+ tasks of 10 categories, including picking, placing, cup-upright, relocating, stacking, passing, pressing, unplugging, opening, and closing.

We employ the same text-guided video prediction (TVP) model as in our simulated experiments, trained on both internet datasets and our self-collected real-world data. We train multi-task generalist policies for the Franka Panda and Xhand Dexterous hands, respectively, to solve all tasks in the domain. The hardware platform and visualizations of some selected tasks are shown in Figure 7.

Quantitative Results. Due to the complexity of deploying methods on real-world hardware, we select the strongest baseline models—GR-1, Susie, and the widely-used diffusion policy—as our baselines. We categorize the tasks into "seen" and "unseen" to assess the model's capabilities. The unseen tasks include new backgrounds and objects that do

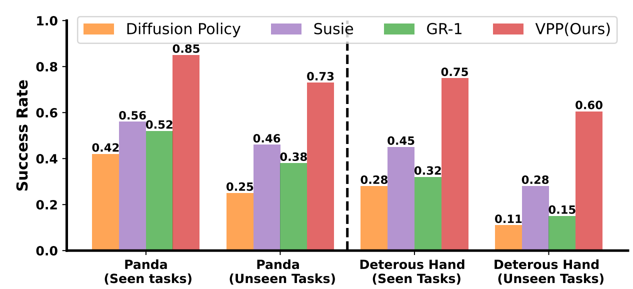


Figure 8. Evaluations on real-world seen/unseen tasks.

not appear in the dataset. For evaluation, we perform 200+ rollouts for Panda arm manipulation tasks and 500+ rollouts for dexterous hand manipulation tasks. Due to space constraints, we report only the average success rate in Figure 8. Detailed success rates can be found in Appendix 1, and videos of the roll-out trajectories are available in the supplementary.

6. Conclusion

We introduce Video Prediction Policy (VPP), a novel approach for learning a generalist robot policy by leveraging predictive visual representations from a video prediction model. Our results show that the representations generated by video prediction models are highly valuable for robot policy learning, yielding consistent improvements across both simulated and real-world tasks. We aim to highlight the potential of video generation models in embodied tasks and underscore the importance of visual representation learning in developing generalist robot policies.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 5

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. 2

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[4] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024. 3

[5] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023. 2, 3, 5, 6

[6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 2, 3, 7

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 1, 5, 6, 2

[8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 1

[9] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3, 5, 2

[10] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2

[12] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 2

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[14] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 2

[15] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 2

[16] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023. 1, 2, 3, 5, 6

[17] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019. 5

[18] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 4, 5, 6

[19] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021. 5, 7

[20] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 1, 2, 5

[21] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2

[22] Xianfan Gu, Chuan Wen, Weirui Ye, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. *arXiv preprint arXiv:2303.14897*, 2023. 2, 7, 3

[23] Gunshi Gupta, Karmesh Yadav, Yarin Gal, Dhruv Batra, Zsolt Kira, Cong Lu, and Tim GJ Rudner. Pre-trained text-to-image diffusion models are versatile representation learners for control. *arXiv preprint arXiv:2405.05852*, 2024. 3

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 2

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[26] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[27] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[28] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022. 1, 5

[29] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023. 1, 2, 7

[30] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*, 2024. 2, 5, 6

[31] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 1

[32] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023. 5, 6

[33] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024. 1

[34] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[35] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023. 7

[36] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024. 1

[37] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022. 2

[38] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, 36:655–677, 2023. 2, 7

[39] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7327–7334, 2022. 2, 5

[40] Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. Grounding language with visual affordances over unstructured data. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11576–11582. IEEE, 2023. 5

[41] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022. 1, 2

[42] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 2, 5

[43] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. In *international conference on machine learning*, pages 17359–17371. PMLR, 2022. 2

[44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3

[46] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pages 416–426. PMLR, 2023. 2, 5

[47] Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals, 2024. 3, 5, 6, 2

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[49] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey

Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. 1, 5, 2

[50] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[51] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023. 3, 5, 6, 2

[52] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 7

[53] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15802–15812, 2023. 3, 7

[54] Karmesh Yadav, Arjun Majumdar, Ram Ramrakhya, Naoki Yokoyama, Alexei Baevski, Zsolt Kira, Oleksandr Maksymets, and Dhruv Batra. Ovrl-v2: A simple state-of-art baseline for imagenav and objectnav. *arXiv preprint arXiv:2303.07798*, 2023. 2

[55] Karmesh Yadav, Ram Ramrakhya, Arjun Majumdar, Vincent-Pierre Berges, Sachit Kuhar, Dhruv Batra, Alexei Baevski, and Oleksandr Maksymets. Offline visual representation learning for embodied navigation. In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2023. 2

[56] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2

[57] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020. 2, 5

[58] Jianke Zhang, Yanjiang Guo, Xiaoyu Chen, Yen-Jen Wang, Yucheng Hu, Chengming Shi, and Jianyu Chen. Hirt: Enhancing robotic control with hierarchical robot transformers. In *8th Annual Conference on Robot Learning*. 1

Video Prediction Policy: A Generalist Robot Policy with Predictive Visual Representations

Supplementary Material

For your convenience, a merged video of our rollouts is included in the supplementary zip file.

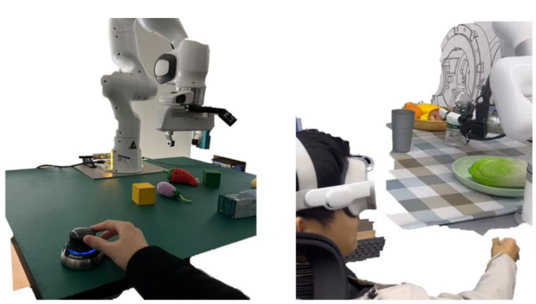
1. Real-world experiments

1.1. Panda Maniplation

On the Franka Panda platform, we gathered demonstrations by teleoperating the Panda robotic arm using a space mouse. we collected 2k trajectories for over 30+ tasks of 6 categories including picking, placing, pressing, routing, opening, and closing. Detailed success rates for each task in seen and unseen settings are shown in Table 6.

Seen Tasks	Diffusion Policy	Susie	GR-1	VPP
Pick	0.36	0.56	0.52	0.90
Place	0.40	0.42	0.38	0.86
Press	0.65	0.90	0.80	0.85
Route	0.40	0.55	0.50	0.75
Drawer	0.45	0.60	0.60	0.85
Average	0.425	0.563	0.519	**0.856**
Unseen Tasks	Diffusion Policy	Susie	GR-1	VPP
Pick	0.24	0.40	0.32	0.80
Place	0.12	0.44	0.32	0.72
Press	0.50	0.60	0.60	0.80
Route	0.20	0.50	0.50	0.70
Drawer	0.40	0.50	0.40	0.60
Average	0.250	0.463	0.388	**0.737**

Table 6. Specific success rate at category level. In seen tasks, We evaluate pick and place tasks 50 times and other tasks 20 times respectively. In unseen tasks, we evaluate pick and place tasks 25 times and other tasks 10 times respectively



(a) Data collection with Space Mouse (b) Data collection with Vision Pro

Figure 9. Data collection setups.

1.2. Dexterous Manipulation

To collect data for dexterous manipulation, we employ Vision-Pro to capture the finger joint movements of the human hand, which are then retargeted to our 12-degree-of-freedom dexterous hand. This setup enables a human operator to directly control the dexterous hand during various manipulation tasks. We collected 2.5k trajectories over 100+ tasks of 10 categories, including picking, placing, cup-upright, relocating, stacking, passing, pressing, unplugging, opening, and closing. A low-level PD controller is used to smooth the trajectories generated by VPP.

The detailed success rates for each task category in both seen and unseen settings are shown in Table 7.

Seen Tasks	Diffusion Policy	Susie	GR-1	VPP
Pick	0.38	0.61	0.48	0.83
Place	0.35	0.55	0.40	0.79
Cup-upright	0.00	0.00	0.00	0.64
Relocate	0.28	0.44	0.16	0.80
Stack	0.00	0.08	0.00	0.64
Pass	0.040	0.00	0.00	0.48
Press	0.68	0.96	0.64	0.96
Unplug	0.00	0.00	0.00	0.52
Drawer	0.40	0.64	0.48	0.72
Average	0.287	0.450	0.319	**0.749**
Unseen Tasks	Diffusion Policy	Susie	GR-1	VPP
Pick	0.12	0.42	0.26	0.75
Place	0.08	0.32	0.20	0.68
Cup-upright	0.00	0.00	0.00	0.40
Relocate	0.12	0.32	0.12	0.76
Stack	0.00	0.00	0.00	0.56
Pass	0.00	0.00	0.00	0.32
Press	0.44	0.76	0.40	0.88
Unplug	0.00	0.00	0.00	0.20
Drawer	0.28	0.44	0.24	0.56
Average	0.110	0.328	0.159	**0.605**

Table 7. Specific success rate at category level. In seen tasks, We evaluate pick and place tasks 100 times and other tasks 25 times respectively. In unseen tasks, we evaluate pick and place tasks 50 times and other tasks 20 times respectively

Method	Tasks completed in a row					
	1	2	3	4	5	Avg. Len ↑
VPP(Ours)	0.957	0.912	0.863	0.810	0.750	**4.29**
VPP(Single-view)	0.909	0.815	0.713	0.620	0.518	3.58
Ablation.1	0.949	0.900	0.839	0.780	0.714	4.18
Ablation.2	0.951	0.904	0.840	0.777	0.718	4.19

Table 8. More ablation studies.

2. Video Prediction Model

2.1. Datasets Sample Ratios

Given the varying quality and scale of these datasets, we have introduced different sample ratios to appropriately balance the influence of different datasets, similar to [49]. Detailed information is shown in Table 9.

2.2. More Visualization of Complete Prediction Results

We present additional visualizations of prediction results from our fine-tuned manipulation TVP model. Predictions on human manipulation datasets are displayed in Figure 10, and those on robotic manipulation datasets are illustrated in Figure 11. All trajectories are sampled from the validation datasets and are predicted using the same manipulation TVP model. Each sample was denoised in 30 steps using classifier-free guidance set at 7.5, as described in [22]. Our TVP model predicts a horizon of 16, and we visualize 8 frames at a skip step of 2 due to space constraints.

2.3. More Visualizations of Predictive Representations

We visualize the intermediate predictive representations through one-step direct predictions. Additional visualizations can be found in Figure 12. As discussed in the experimental section, while the textures and details in the one-step forward videos are not precise, they still offer valuable insights into physical evolution. The movements of objects and robot arm itself already can be reflected in the visualized representations.

3. More Details for Experiments

3.1. Structure details

We provide the VPP architecture and hyperparameter setting details in four evaluate environments, as shown in Table 10. The transformer block in TVP follows the setting in [6], and the rest of the hyperparameter in Diffusion Transformer follows the work [47].

3.2. More ablation

In this section, we present additional ablation experiments conducted under the ABC→D setting of CALVIN [39].

Ablation 1 entails the removal of the Temporal-attn module from the Video Former while maintaining all other configurations same as VPP. The results, displayed in Table 8, demonstrate that the Temporal-attn module could enhance the temporal comprehension capabilities of the Video Former.

Ablation 2 introduces a 2-step denoising process in the TVP to derive the predictive visual representation. The outcomes are summarized in Table 8, revealing that the 2-step process did not yield superior performance. We hypothesize this is because a single denoising step suffices to generate an effective representation for trajectory prediction in our configuration. Additionally, the 2-step denoising process nearly doubles the inference time and reduces the control frequency by half. Due to these factors, we opted for a one-step direct encoder in our main experiments.

Single-view Ablation evaluate the Calvin ABC→D task using only a single observation viewpoint (static view) and find that the success rate for Task 5 reaches 3.58. This even surpasses the success rate achieved by the state-of-the-art 3D Diffuser Actor, which utilizes two viewpoints along with depth images.

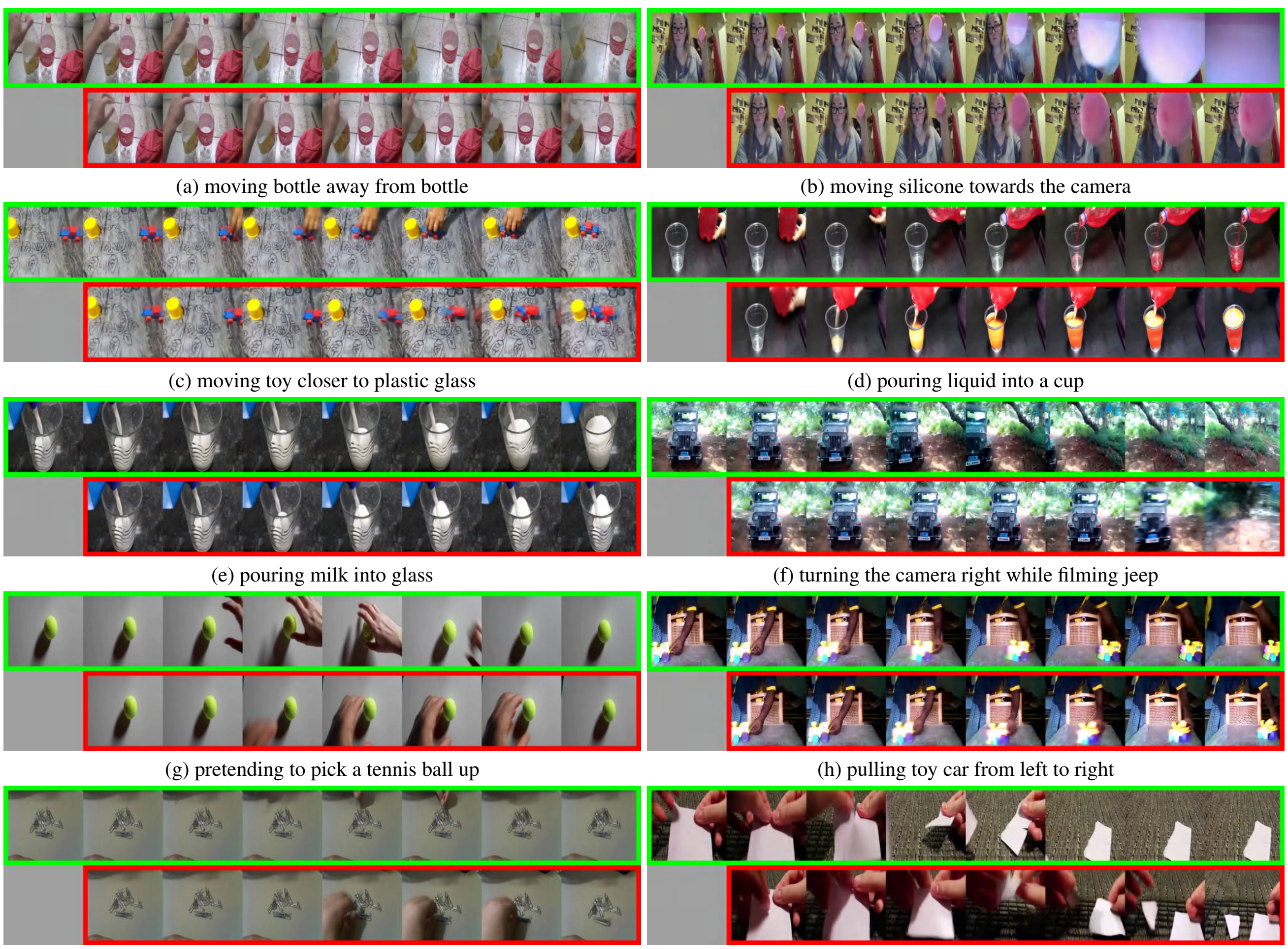
3.3. Baseline Implementations

The baseline methods, including RT-1 [7], GR-1 [51], and Diffusion Policy [16], are implemented based on their official repositories. For comparison with Susie [5] in both the Metaworld and real-world manipulation scenarios, we adopt InstructPix2Pix [9] as the future frame predictor and use an image-goal Diffusion Policy [16] to generate the state sequence.

CVPR
#9874

Dataset Type	Name	Trajectory Numbers	Smaple Ratio
Internet Human Maniplation	Something-something-v2	191,642	0.30
Internet Robot Datasets	RT-1	87,212	0.15
	Bridge	23,377	0.15
	BC-Z	43,264	0.08
	Taco-Play	3,603	0.01
	Jaco-Play	1,085	0.01
	Calvin-ABC	18,033	0.10
	Metaworld	2,500	0.05
Self-Collected Datasets	Panda Arm	2,000	0.05
	Dexterous Hand	2,476	0.10
Total	-	375,192	1.00

Table 9. We outline the dataset scales and sample ratios used for training our manipulation text-guided video prediction model. Following [22], we **exclude** 5,558 bridge trajectories and 2,048 something-something-v2 trajectories during training, reserving them for validation. For all other datasets, 3% of the trajectories are excluded and used as validation datasets.



(a) moving bottle away from bottle

(b) moving silicone towards the camera

(c) moving toy closer to plastic glass

(d) pouring liquid into a cup

(e) pouring milk into glass

(f) turning the camera right while filming jeep

(g) pretending to pick a tennis ball up

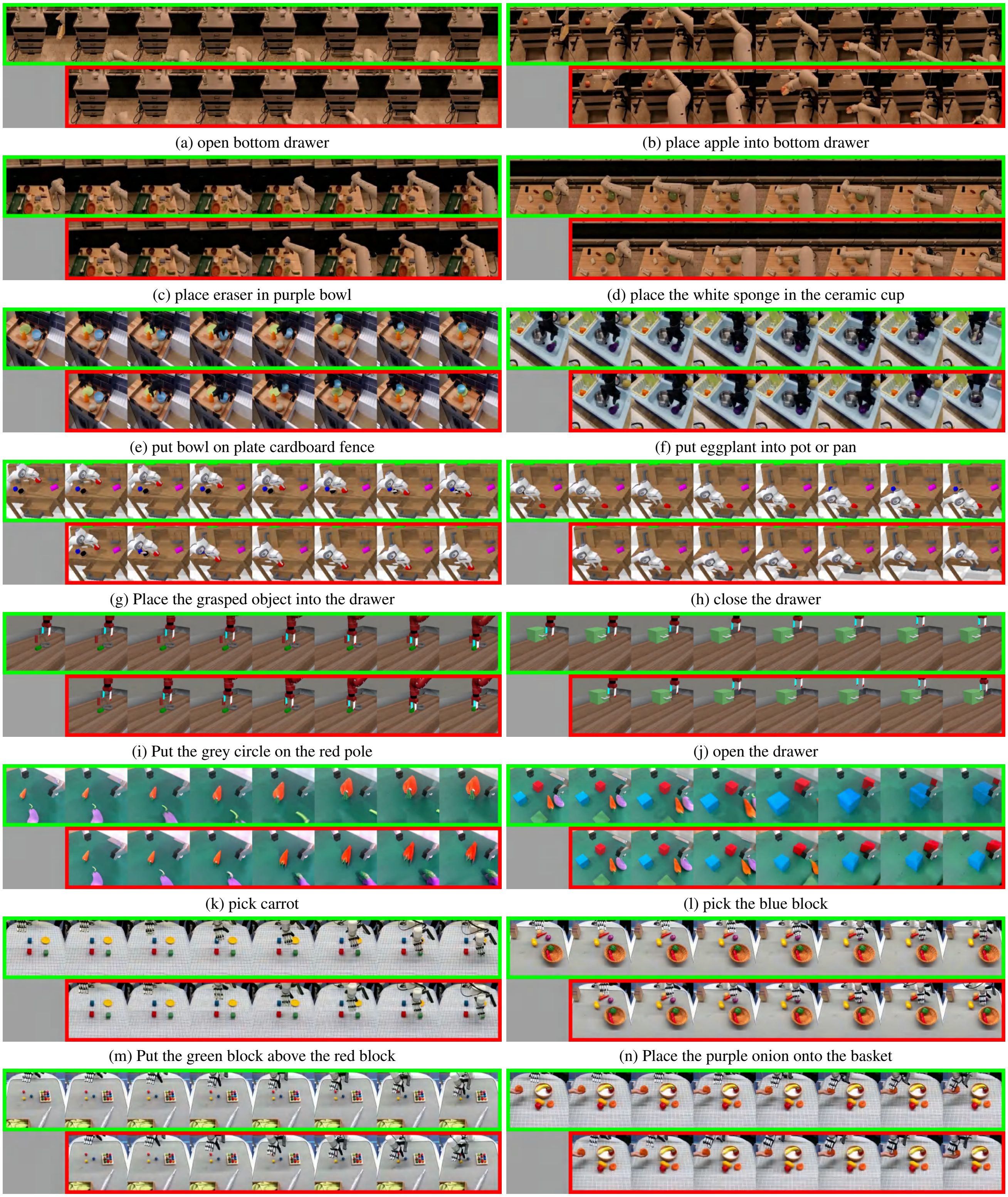
(h) pulling toy car from left to right

(i) putting something similar

(j) tearing paper into two pieces

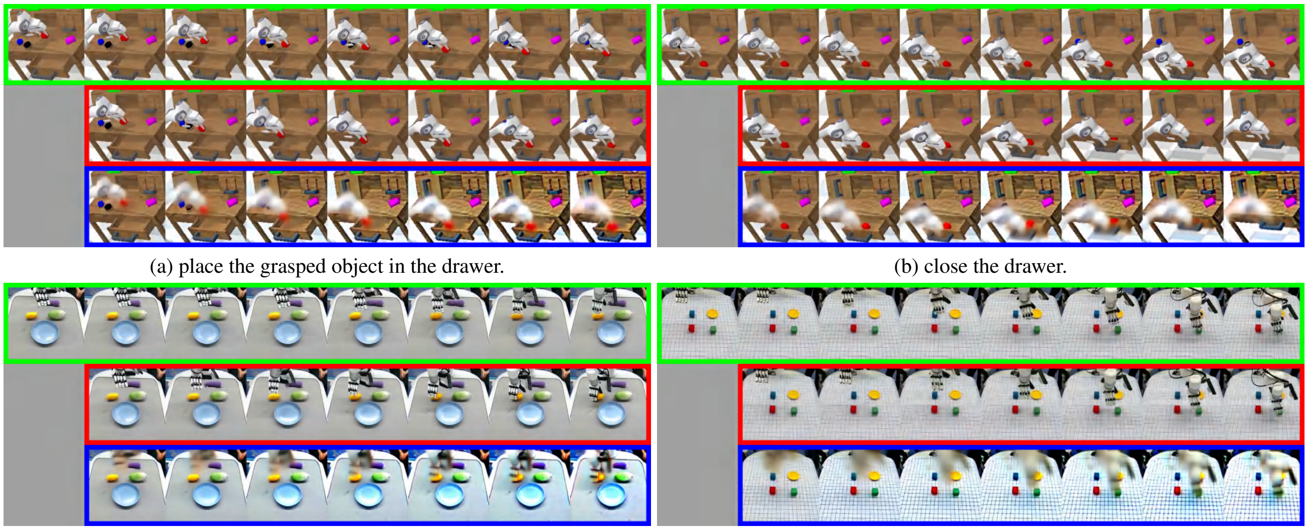
Figure 10. **Visualization of video prediction results on Internet human manipulation validation datasets with 30 steps de-noising**. The green frames indicate the ground truth while the red frames indicate the predicted futures. Zoom in for better comparisons.

CVPR
#9874



(a) open bottom drawer
(b) place apple into bottom drawer
(c) place eraser in purple bowl
(d) place the white sponge in the ceramic cup
(e) put bowl on plate cardboard fence
(f) put eggplant into pot or pan
(g) Place the grasped object into the drawer
(h) close the drawer
(i) Put the grey circle on the red pole
(j) open the drawer
(k) pick carrot
(l) pick the blue block
(m) Put the green block above the red block
(n) Place the purple onion onto the basket
(o) pick the yellow ball to the empty hole of the board
(p) Receive the orange from hand and place it into the basket

Figure 11. **Visualization of video prediction results on robotic datasets with 30 steps de-noising**. The green frames indicate the ground truth while the red frames indicate the predicted futures. (a)-(j) are sourced from internet robotic while (k)-(p) are from self-collected datasets. Zoom in for better comparisons.



(a) place the grasped object in the drawer.

(b) close the drawer.

(c) pick the mango and place in blue plate

(d) stack the green block on red block

Figure 12. **Visualization of Predictive representations**. Green frames represent the ground truth, red frames correspond to the predicted future states, and blue frames illustrate the visualized predictive representations. Zoom in for better comparisons.

Type	Name	Calvin	Metaworld	Franka Panda	Xhand
Prediction	Video lens	16	8	16	16
	Action shape	$10 * 7$	$4 * 4$	$10 * 7$	$10 * 18$
TVP	Language shape	$20 * 512$	$20 * 512$	$20 * 512$	$20 * 512$
	Image shape	$256 * 256$	$256 * 256$	$256 * 256$	$256 * 256$
Video Former	Token shape	$16 * 14 * 384$	$8 * 28 * 384$	$14 * 16 * 384$	$14 * 16 * 384$
	Input dim	1280	1280	1280	1280
	Latent dim	512	512	512	512
	Num heads	8	8	8	8
	num Layers	6	6	6	6
Diffusion Transformer	Latent dim	384	384	384	384
	Condition shape	$225 * 384$	$225 * 384$	$225 * 384$	$225 * 384$
	Num heads	8	8	8	8
	Encoder Layers	4	4	4	4
	Decoder Layers	4	4	4	4
	Sampling Steps	10	10	10	10
Hyperparameter	TVP batchsize	4	4	4	4
	Policy batchsize	76	64	128	128
	Epoch nums	12	30	30	40
	Learning rate	$1 * 10^{-4}$	$5 * 10^{-5}$	$1 * 10^{-4}$	$1 * 10^{-4}$

Table 10. Hyper-parameters in the Video Prediction Policy (VPP).