Visual Representation Learning with Stochastic Frame Prediction

Huiwon Jang [1] **Dongyoung Kim** [1] **Junsu Kim** [1] **Jinwoo Shin** [1] **Pieter Abbeel** [2] **Younggyo Seo** [1 3]

Abstract

Self-supervised learning of image representations by predicting future frames is a promising direction but still remains a challenge. This is because of the under-determined nature of frame prediction; multiple potential futures can arise from a single current frame. To tackle this challenge, in this paper, we revisit the idea of stochastic video generation that learns to capture uncertainty in frame prediction and explore its effectiveness for representation learning. Specifically, we design a framework that trains a stochastic frame prediction model to learn temporal information between frames. Moreover, to learn dense information within each frame, we introduce an auxiliary masked image modeling objective along with a shared decoder architecture. We find this architecture allows for combining both objectives in a synergistic and compute-efficient manner. We demonstrate the effectiveness of our framework on a variety of tasks from video label propagation and vision-based robot learning domains, such as video segmentation, pose tracking, vision-based robotic locomotion, and manipulation tasks. Code is available on the project webpage: https://sites.google.com/view/2024rsp.

1. Introduction

Recently, generative pre-training on sequential data has been extremely successful in learning models that can be easily fine-tuned (Oord et al., 2016; Yang et al., 2019; Dai et al., 2019; Radford et al., 2019) or achieve impressive performance with few adaptations or even without adaptation (Brown et al., 2020; Touvron et al., 2023). The core idea behind these successes is training the model to *predict the future*, *i.e.,* learning the distribution of future outputs conditioned on past data consisting of words (Bengio et al., 2000; Radford et al., 2019), audio signals (Oord et al., 2016; Dhariwal et al., 2020), or state of the world (Chen et al., 2021a), enabling the models to understand the temporal and causal relationships within the data.

There also have been efforts to learn rich representations in video domains by learning video prediction models (Srivastava et al., 2015; Vondrick et al., 2016; Finn et al., 2016; Yu et al., 2020b), for its promise of utilizing an abundance of videos for learning representations that understand how the world operates by predicting the future. However, it has been less successful when compared to its counterparts in image domains (Kingma & Welling, 2013; Donahue & Simonyan, 2019; Chen et al., 2020a; Li et al., 2023) or other self-supervised learning approaches that do not involve generative modeling of future frame (Wang & Gupta, 2015; Misra et al., 2016; Sermanet et al., 2018; Han et al., 2020b).

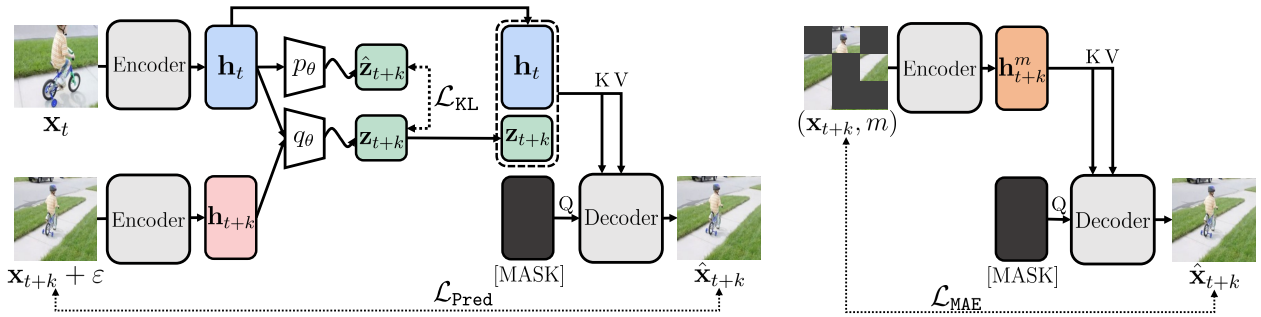
In this paper, we argue that this challenge can be attributed to the inherently under-determined nature of future frame prediction, where multiple potential futures can arise from a single current frame (Babaeizadeh et al., 2017; Denton & Fergus, 2018). This issue makes it difficult for deterministic models to learn useful representations from complex real-world videos because the model would struggle to approximate the multi-modal distribution of future frames. In contrast, recent video generation models have achieved remarkable successes in generating high-fidelity videos (Yan et al., 2021; Villegas et al., 2022; Ho et al., 2022; Blattmann et al., 2023a;b), where the core idea is to train a *stochastic* generative model[1] that can capture the uncertainty in generating or predicting the videos, such as denoising diffusion models (Ho et al., 2022; Yu et al., 2023) and autoregressive models (Yan et al., 2021; Villegas et al., 2022). Inspired by these successes, we aim to investigate how to adopt and utilize the idea of training a stochastic generative model for visual representation learning from videos.

Contribution We present visual Representation learning with Stochastic frame Prediction (RSP), a framework for visual representation learning from videos. Our key idea

[1]KAIST [2]UC Berkeley [3]Now at Dyson Robot Learning Lab. Correspondence to: Huiwon Jang <huiwoen0516@kaist.ac.kr>, Younggyo Seo <seo0gyo@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]While a deterministic prediction model learns a deterministic mapping from the current frame to the future frame, a stochastic prediction model aims to learn a distribution over the future frame conditioned on the current frame.



(a) Stochastic frame prediction

(b) Masked autoencoding with shared decoder

Figure 1: **Representation learning with stochastic frame prediction.** (a) We train a stochastic frame prediction model, which is built upon stochastic video generation model (Denton & Fergus, 2018), which consists of an encoder that extracts representations, a posterior model with access to both current and future frames, a prior model with only access to the current frame, and a decoder that generates frame conditioned on features from the current frame and a sample from either posterior or prior distributions. We train the model to accurately generate the future frame while enforcing the posterior and prior distributions to be close to each other, *i.e.,* encourage the posterior distribution to be more predictable and the prior distribution to predict the future. (b) We introduce an auxiliary masked autoencoding objective (He et al., 2022) with a shared decoder architecture. Our decoder makes the `[MASK]` tokens attend to different inputs via the cross-attention layer, enabling us to share the decoder parameters for different objectives.

is to learn *image* representations that capture temporal information between frames by learning a stochastic frame prediction model with videos. To this end, we revisit the idea of stochastic video generation (Denton & Fergus, 2018) that trains a time-dependent prior over future frames to capture uncertainty in frame prediction (see Figure 1a). Specifically, our key contribution lies in exploring various design choices and incorporating recent advances in self-supervised learning into the video generation model (Dosovitskiy et al., 2021; Hafner et al., 2021a; Gupta et al., 2023), to re-configure it for representation learning. We find that RSP allows for learning strong image representations from complex real-world videos when compared to deterministic prediction objectives. To learn dense information within each frame, we further introduce an auxiliary masked autoencoding objective (He et al., 2022), along with a shared decoder architecture that enables us to incorporate the auxiliary objective in a synergistic manner (see Figure 1b).

Through extensive experiments, we show that RSP can effectively learn image representations from a large real-world video dataset. Pre-trained on Kinetics-400 dataset (Kay et al., 2017), RSP achieves competitive or superior performance to various self-supervised learning baselines on a variety of tasks from vision-based robot learning benchmarks (James et al., 2020; Majumdar et al., 2023) and video label propagation benchmarks (Pont-Tuset et al., 2017; Zhou et al., 2018; Jhuang et al., 2013). In particular, RSP achieves a 36.0% average success rate in challenging robotic manipulation tasks from RLBench (James et al., 2020), while MAE baseline only achieves a 13.5% success rate. We also provide extensive ablation studies and analyses on the importance of various design choices in our framework.

2. Related Work

Image self-supervised learning Self-supervised learning (SSL) from images has demonstrated remarkable success in visual representation learning by exploiting the rich, inherent structure of visual data without human labels (Chen et al., 2020b; He et al., 2020; Chen et al., 2021b; Caron et al., 2021; He et al., 2022). Pioneer works for SSL propose pretext tasks (Doersch et al., 2015; Pathak et al., 2016; Zhang et al., 2016; Noroozi & Favaro, 2016; Gidaris et al., 2018), and recently, contrastive learning (Chen et al., 2020b; He et al., 2020; Chen et al., 2021b; Caron et al., 2021) and masked image modeling (Bao et al., 2021; He et al., 2022; Xie et al., 2022; Li et al., 2023) have gained prominence. In this paper, we show that integrating an understanding of the temporal information between the frames can further enhance image representation.

Video self-supervised learning Most prior researches on SSL from videos aim to learn video representations capturing spatiotemporal information from videos that could be useful for video understanding tasks such as action recognition (Xu et al., 2019; Benaim et al., 2020; Han et al., 2020a;b; Feichtenhofer et al., 2021; Pan et al., 2021; Qian et al., 2021; Ge et al., 2021; Guo et al., 2022; Tong et al., 2022; Feichtenhofer et al., 2022). Our work differs in that we focus on learning useful image representations from videos. Similarly to our work, there have been approaches that focus on enhancing image representations, by designing pretext tasks for videos (Wang & Gupta, 2015; Misra et al., 2016), extending contrastive learning to video frames (Sermanet et al., 2018; Wang et al., 2019; Jabri et al., 2020;

Xu & Wang, 2021), and masked visual modeling (Feichtenhofer et al., 2022; Gupta et al., 2023). In particular, Gupta et al. (2023) learns visual correspondence by predicting the masked patches from the future frame. This is closely related to our work as it represents another approach to the future frame prediction objective. However, unlike Gupta et al. (2023), which resolves ambiguity about the future by conditioning on unmasked patches from the future frame, we aim to learn representations that capture the inherent stochasticity of future frame prediction.

3. Method

In this section, we present **R**epresentation learning with **S**tochastic frame **P**rediction (RSP), a framework that learns visual representations from videos via stochastic future frame prediction. We first describe how we revisit the idea of stochastic video generation (Denton & Fergus, 2018) for representation learning and improve it by incorporating a recent recipe for self-supervised learning (see Section 3.1). We then describe how we design a shared decoder architecture to effectively incorporate an auxiliary masked autoencoding objective (He et al., 2022) that learns dense information within the static parts of each frame (see Section 3.2). We provide the overview and pseudo-code of our framework in Figure 1 and Algorithm 1, respectively.

3.1. Representation Learning from Videos with Stochastic Frame Prediction

Our key idea is that learning a model that can predict multiple possible future frames can induce representations that capture temporal information between frames. To this end, we build our framework upon the stochastic video generation (SVG; Denton & Fergus, 2018) model that captures uncertainty in future prediction by learning a time-dependent prior distribution over future frames. Our key contribution lies in re-configuring SVG for representation learning by exploring multiple design choices and adopting recent advances in architectures and training techniques (Dosovitskiy et al., 2021; Hafner et al., 2021a; He et al., 2022; Gupta et al., 2023), which we describe in the rest of this section.

Inputs and encoder Given a video $\mathbf{x}$, we randomly sample two frames $\{\mathbf{x}_t, \mathbf{x}_{t+k}\}$ where k is randomly chosen from a fixed set of values by following Gupta et al. (2023). Then we use the same vision transformer (ViT; Dosovitskiy et al., 2021) encoder $f_\theta^{\texttt{enc}}$ that shares parameters for encoding frames $\mathbf{x}_t$ and $\mathbf{x}_{t+k}$. Specifically, we extract non-overlapping patches from a frame, add 2D fixed sin-cos positional embeddings (Chen et al., 2021b), and concatenate a `[CLS]` token to patches. We note that we separately process each frame and do not concatenate patches from both frames. We then process them through a series of Transformer layers (Vaswani et al., 2017) to obtain to obtain $\mathbf{h}_t$ and $\mathbf{h}_{t+k}$ consisting of `[CLS]` and patch representations.

Algorithm 1 RSP: PyTorch-like Pseudocode

```
# f, g: encoder, decoder
# q, p: posterior, learned prior

# input: x1 (current frame), x2 (future frame)
def rsp(x1, x2):
    h1, h2 = f(x1), f(perturb(x2))

    # Posterior distribution from both frames
    post_logits = q(cat(h1[:,0], h2[:,0]))
    post_dist   = make_dist(post_logits)
    post_z      = post_dist.rsample()

    # Prior distribution only from the current frame
    prior_logits = p(h1[:,0])
    prior_dist   = make_dist(prior_logits)

    pred_fut  = g(q=<mask>, kv=cat(h1, post_z))
    pred_loss = ((pred_fut - x2) ** 2).mean()
    kl_loss   = kl(post_dist, prior_dist)

    # Auxiliary MAE objective
    hm, mask, ids_restore = f(x2, mask=0.75)
    pred_mask = g(q=<mask>,
                  kv=restore(hm, ids_restore))
    mae_loss  = ((pred_mask - x2) ** 2).mean(dim=-1)
    mae_loss  = (mae_loss * mask).sum() / mask.sum()

    loss = pred_loss + kl_scale * kl_loss + mae_loss
    return loss
```

$$\text{Encoder:} \quad \begin{cases} \mathbf{h}_{t+k} = f_\theta^{\texttt{enc}}(\mathbf{x}_{t+k}) \\ \mathbf{h}_t = f_\theta^{\texttt{enc}}(\mathbf{x}_t) \end{cases} \tag{1}$$

Augmentations We apply the same augmentation *i.e.,* random resized crop and random horizontal flip, to both frames $\mathbf{x}_t$ and $\mathbf{x}_{t+k}$. This is because applying such a strong augmentation differently to frames can sometimes make the two frames be significantly different from each other (see Table 4a for supporting experiments). We then add a small Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma)$ to the future frame $\mathbf{x}_{t+k}$ to discourage the model from finding a shortcut that simply copies pixels from $\mathbf{x}_{t+k}$ for predicting $\hat{\mathbf{x}}_{t+k}$.

Posterior and learned prior Following Denton & Fergus (2018), our framework consists of two main components: (i) a *future frame prediction model* that predicts $\hat{\mathbf{x}}_{t+k}$ conditioned on $\mathbf{h}_t$ and a latent variable $\mathbf{z}_{t+k}$, which captures the uncertainty over future, from a posterior distribution $q_\theta(\mathbf{z}_{t+k} \mid \mathbf{h}_t, \mathbf{h}_{t+k})$ and (ii) a *prior network* that learns to approximate $p_\theta(\mathbf{z}_{t+k} \mid \mathbf{h}_t)$ without access to the future frame.

$$\begin{aligned} &\text{Posterior:} && \mathbf{z}_{t+k} \sim q_\theta(\mathbf{z}_{t+k} \mid \mathbf{h}_t, \mathbf{h}_{t+k}) \\ &\text{Learned prior:} && \hat{\mathbf{z}}_{t+k} \sim p_\theta(\hat{\mathbf{z}}_{t+k} \mid \mathbf{h}_t) \end{aligned} \tag{2}$$

In our implementation, we introduce two small 2-layer MLP models that take `[CLS]` representations from both $\mathbf{h}_t$ and $\mathbf{h}_{t+k}$ for the posterior network and `[CLS]` representation from $\mathbf{h}_t$ for the prior network. For the latent variable $\mathbf{z}_{t+k}$,

we use a set of categorical variables by following Hafner et al. (2021a) and use the straight-through estimator (Bengio et al., 2013) for updating the parameters, which we find to be more effective than using Gaussian distribution (see Table 3 for supporting experiments).

Decoder For decoding, we first project $\mathbf{h}_t$ and $\mathbf{z}_{t+k}$ with a linear layer and concatenate them to $[\mathbf{h}_t, \mathbf{z}_{t+k}]$. Our decoder block consists of a (i) cross-attention layer where `[MASK]` tokens attend to tokens from $[\mathbf{h}_t, \mathbf{z}_{t+k}]$ and (ii) self-attention layer where `[MASK]` tokens attend to each other. After processing the inputs through a series of decoder blocks, the final projection layer maps the token representations into normalized pixel patches $\hat{\mathbf{x}}_{t+k}$ (He et al., 2022).

$$\text{Decoder:} \quad \hat{\mathbf{x}}_{t+k} \sim p_\theta(\hat{\mathbf{x}}_{t+k} \,|\, \mathbf{h}_t, \mathbf{z}_{t+k}) \tag{3}$$

Here, we note that our architecture resembles the cross-self decoder (Gupta et al., 2023) where unmasked patches from $\mathbf{x}_{t+k}$ attend to $\mathbf{x}_t$ via cross-attention layers. But our design differs in that there is no interaction between $\mathbf{x}_t$ and $\mathbf{x}_{t+k}$ in our cross-attention layer. We adopt this design to be able to share the decoder parameters for multiple objectives by making `[MASK]` tokens attend to different types of inputs via cross-attention layers, which allows for effectively incorporating both frame prediction and MAE objectives into our framework, which we describe in Section 3.2.

Objective We train the future frame prediction model to provide accurate prediction $\hat{\mathbf{x}}_{t+k}$ while minimizing the KL divergence between the prior distribution $p_\theta(\mathbf{z}_{t+k} \,|\, \mathbf{x}_t)$ and the posterior distribution $q_\theta(\mathbf{z}_{t+k} \,|\, \mathbf{x}_t, \mathbf{x}_{t+k})$ as below:

$$\begin{aligned}\mathcal{L}(\theta) = \mathbb{E}_{q_\theta(\mathbf{z}_{t+k}|\mathbf{x}_t,\mathbf{x}_{t+k})}\Big[&-\ln p_\theta(\mathbf{x}_{t+k}|\mathbf{x}_t, \mathbf{z}_{t+k}) \\ &+\beta\, \mathrm{KL}\big[q_\theta(\mathbf{z}_{t+k}|\mathbf{x}_t, \mathbf{x}_{t+k}) \,\|\, p_\theta(\mathbf{z}_{t+k}|\mathbf{x}_t)\big]\Big],\end{aligned} \tag{4}$$

where β is a loss scale hyperparameter that adjusts the balance between decoding loss and KL loss. Intuitively, making the prior distribution to be closer to the posterior distribution corresponds to learning the prior network to *predict the future*. On the other hand, enabling the prediction model to generate better frames while making the posterior distribution closer to the prior distribution corresponds to *making the latent variable more predictable* by the prior network (Denton & Fergus, 2018). We find that our objective allows for learning strong representations from complex real-world videos when compared to the deterministic frame prediction model (see Table 3a for supporting experiments).

3.2. Auxiliary Representation Learning from Images

While stochastic future frame prediction can induce representations capturing temporal information, it might focus less on the static parts of frames as the model has full access to the previous frame $\mathbf{x}_t$ when predicting $\mathbf{x}_{t+k}$. To mitigate this issue, we introduce an auxiliary masked autoencoding (MAE; He et al., 2022) objective that focuses on learning the dense information within each frame. Moreover, we design our framework to share the decoder across the frame prediction and MAE objectives, which enables both objectives to be synergistic with a small computational overhead.

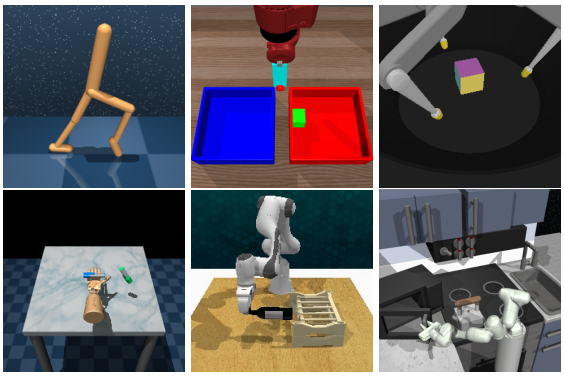


Figure 2: Examples of visual observations from CortexBench (Majumdar et al., 2023), RLBench (James et al., 2020), and FrankaKitchen (Gupta et al., 2019), which we used for training imitation learning agents that learn a mapping from observations to expert actions. Learning such agents requires representations that can understand both temporal and dense information.

Masked autoencoding with shared decoder We mask $m\%$ of the patches from $\mathbf{x}_{t+k}$ and process them through the encoder $f_\theta^{\texttt{enc}}$ to obtain $\mathbf{h}_{t+k}^m$ consisting of `[CLS]` and unmasked patch representations. We then project $\mathbf{h}_{t+k}^m$ with a linear layer, which is different from the linear layer used in the frame prediction, and process them through the *shared* decoder by making `[MASK]` tokens attend to $\mathbf{h}_{t+k}^m$ via cross-attention layers. Then the final projection layer maps the outputs into normalized pixel patches $\hat{\mathbf{x}}_{t+k}$.

$$\begin{aligned}&\text{Masking:} && \mathbf{x}_{t+k}^m \sim p^{\texttt{mask}}(\mathbf{x}_{t+k}, m) \\ &\text{Encoder:} && \mathbf{h}_{t+k}^m = f_\theta^{\texttt{enc}}(\mathbf{x}_{t+k}^m) \\ &\text{Decoder:} && \hat{\mathbf{x}}_{t+k} \sim p_\theta(\hat{\mathbf{x}}_{t+k} \,|\, \mathbf{h}_{t+k}^m)\end{aligned} \tag{5}$$

We note that this auxiliary objective effectively enhances performance by complementing the frame prediction objective, with a negligible increase in training time. We also empirically find that our shared decoder is crucial in making two objectives synergistic; training with a parallel decoder design achieves worse performance (see Table 3c for supporting experimental results).

4. Experiments

In this section, we demonstrate the effectiveness of the proposed framework through evaluations on a variety of vision-based robot learning tasks including robotic manipulation

Table 1: **Results on vision-based robot learning.** Performance of imitation learning agents on CortexBench (Majumdar et al., 2023) and RLBench (James et al., 2020), which are trained upon representations from ViT-S/16 model pre-trained on Kinetics-400 (Kay et al., 2017) dataset. We report the normalized score for DMC and success rates (%) for other tasks.

	CortexBench				RLBench					
Method	Adroit	MetaWorld	DMC	Trifinger	Button	Saucepan	Phone	Umbrella	Wine	Rubbish
SimCLR (Chen et al., 2020b)	40.4	78.4	39.7	63.3	7.4	39.5	34.6	5.8	11.0	5.2
MoCo v3 (Chen et al., 2021b)	39.6	65.4	43.7	53.3	11.4	45.8	36.2	13.2	8.7	6.7
Dino (Caron et al., 2021)	**45.6**	82.4	50.9	64.2	24.7	57.9	32.0	28.1	31.4	12.9
MAE (He et al., 2022)	44.8	81.4	52.1	62.2	6.4	36.8	37.7	10.0	10.0	6.2
SiamMAE (Gupta et al., 2023)	44.0	81.1	56.0	52.1	6.1	22.5	5.4	4.0	8.7	3.5
RSP (Ours)	**45.6**	**84.5**	**61.6**	**66.2**	**28.4**	**93.4**	**48.0**	**37.3**	**31.9**	**18.5**

and locomotion (see Section 4.2) and video label propagation tasks including video segmentation and pose tracking (see Section 4.3). We also provide extensive ablation studies and analysis on our design choices (see Section 4.4).

4.1. Experimental Setup

Pre-training For a fair comparison, we report all the experimental results using the ViT-S/16 model pre-trained on Kinetics-400 datasets (Kay et al., 2017) for 400 epochs. We use the repeated sampling of 2 and count the epochs as effective epochs (Hoffer et al., 2020; Feichtenhofer et al., 2022). For sampling frames $\mathbf{x}_t$ and $\mathbf{x}_{t+k}$, we follow Gupta et al. (2023) that randomly samples k from 4 to 48. We implement our decoder block to sequentially have self-attention, cross-attention, and feedforward layers. For the MAE objective, we use a 75% masking ratio (He et al., 2022). We use AdamW optimizer (Loshchilov & Hutter, 2019) with a batch size of 1536. For all baselines, we use the default hyperparameters. We provide more details in Appendix A.

Baselines We first consider image representation learning approaches, *i.e.,* SimCLR (Chen et al., 2020b), MoCo v3 (Chen et al., 2021b), Dino (Caron et al., 2021) and MAE (He et al., 2022), as our baselines to compare our framework against standard image representation learning methods. Moreover, we consider SiamMAE (Gupta et al., 2023) as our baseline for its superior performance over other masked visual modeling methods (Feichtenhofer et al., 2022; Tong et al., 2022) and its resemblance to our approach. With this comparison against SiamMAE, we evaluate the benefit of our stochastic frame prediction framework compared to the idea of predicting the masked patches of future frames conditioned on the unmasked patches.

4.2. Vision-Based Robot Learning

We evaluate our framework on vision-based robot learning benchmarks, where the goal is to train imitation learning agents that solve target tasks by learning the mapping from visual observations to expert actions via behavior cloning (Pomerleau, 1988). We consider this setup because training

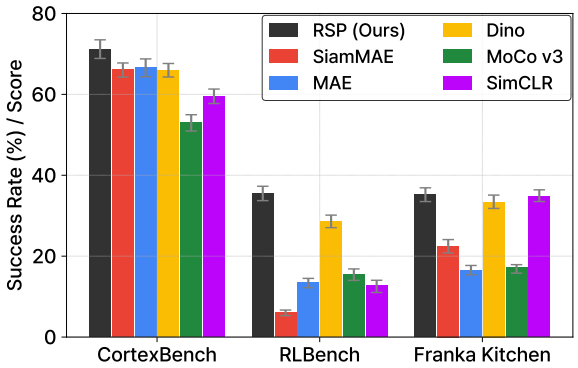


Figure 3: **Aggregate results on vision-based robot learning.** We report the interquartile mean (Agarwal et al., 2021) over 20 vision-based robot learning tasks from CortexBench (Majumdar et al., 2023), RLBench (James et al., 2020), and Frana Kitchen (Gupta et al., 2019).

such agents requires representations that capture both temporal and dense information from the visual observations (see Figure 2 for examples of tasks used in our experiments).

Experimental setup We first consider 4 domains from CortexBench (Majumdar et al., 2023) which includes locomotion and manipulation tasks from various benchmarks (Rajeswaran et al., 2018; Yu et al., 2020a; Tassa et al., 2020; Bauer et al., 2022). Moreover, we consider a more challenging setup by evaluating our framework on 6 manipulation tasks from RLBench (James et al., 2020) which has successfully served as a simulation for sim-to-real transfer (Seo et al., 2023) or a proxy for real-robot experiments (James et al., 2022; Shridhar et al., 2023). We train the imitation learning agents using 100 demos for each task, use keypoint augmentation (James & Davison, 2022) for demonstrations, and use the end-effector controller with path planning as an action mode. We use the front camera of 224×224 resolution without depth for the CortexBench and RLBench. Furthermore, we evaluate RSP on 5 tasks from Franka Kitchen (Gupta et al., 2019), following the setup in Nair et al. (2022) that uses a left or right camera

Table 2: **Results on video label propagation.** We report performances on video segmentation, video part segmentation, and pose tracking tasks from DAVIS (Pont-Tuset et al., 2017), VIP (Zhou et al., 2018), and JHMDB (Jhuang et al., 2013) benchmarks, respectively. For all methods, we report the performance with the representations pre-trained on the Kinetics-400 (Kay et al., 2017) dataset for 400 epochs. We further provide the performance of representations pre-trained on the ImageNet (Deng et al., 2009) dataset as a reference in Appendix D.

		DAVIS			VIP	JHMDB	
Method	Architecture	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$	mIoU	PCK@0.1	PCK@0.2
SimCLR (Chen et al., 2020b)	ViT-S/16	53.9	51.7	56.2	31.9	37.9	66.1
MoCo v3 (Chen et al., 2021b)	ViT-S/16	57.7	54.6	60.8	32.4	38.4	67.6
Dino (Caron et al., 2021)	ViT-S/16	59.5	56.5	62.5	33.4	41.1	70.3
MAE (He et al., 2022)	ViT-S/16	53.5	50.4	56.7	32.5	43.0	71.3
SiamMAE (Gupta et al., 2023)	ViT-S/16	58.1	56.6	59.6	33.3	**44.7**	73.0
RSP (Ours)	ViT-S/16	**60.1**	**57.4**	**62.8**	**33.8**	44.6	**73.4**
RSP (Ours)	ViT-B/16	60.5	57.8	63.2	34.0	46.0	74.6

of 224×224 resolution without depth. For all the tasks, we follow the setup in Majumdar et al. (2023) that trains the agents upon `[CLS]` representation to predict expert actions. We evaluate the model multiple times throughout training with a pre-defined interval and report the best performance.

Results We provide the main experimental results for each individual task (see Table 1) and aggregate performance (see Figure 3). We first find that our framework outperforms all the baselines by a significant margin, as shown in Figure 3 that reports interquartile mean (Agarwal et al., 2021) computed over 25 tasks from the benchmarks. This demonstrates that our framework indeed can induce representations that could be useful for solving complex robot learning tasks that require temporal understanding. We also observe that overall success rates are low in RLBench, as we consider a difficult setup of using only a single camera without depth information. Nevertheless, we find our method consistently achieves superior performance to all the baselines. In particular, RSP outperforms SiamMAE by a large margin in both benchmarks, *i.e.,* RSP achieves 35.6% while SiamMAE achieves 6.0% success rates in RLBench. This highlights the benefit of our approach that captures uncertainty over the future for representation learning.

4.3. Video Label Propagation

To evaluate how learned representations can capture temporal information between frames, we report the performance of three video label propagation tasks. The goal of these tasks is, given a first frame with ground-truth annotations, to predict the labels in each pixel from future frames.

Experimental setup We consider the video object segmentation, video part segmentation, and pose tracking tasks from DAVIS (Pont-Tuset et al., 2017), VIP (Zhou et al., 2018), and JHMDB (Jhuang et al., 2013) benchmarks, respectively. For evaluation, we follow the protocol of prior work (Wang et al., 2019; Li et al., 2019; Lai & Xie, 2019; Jabri et al., 2020) that uses a k-nearest neighbor inference, maintain a queue of length m to provide a temporal context and use a restricted set of source nodes with a spatial radius r. Due to computational constraints, we compare our framework against the baselines pre-trained under the same budget using the same architecture of ViT-S/16. We conduct a grid search on evaluation hyperparameters for each method and report the best performance.

Results We provide the quantitative evaluation in Table 2 and qualitative results in Figure 4. As shown in Table 2, we find that our framework achieves superior or competitive performance to all the baselines in video label propagation tasks. In particular, our framework, with both stochastic frame prediction and auxiliary MAE objectives, outperforms MAE by a large margin, *i.e.,* 6.6%p. This highlights the effectiveness of stochastic future frame prediction objectives for temporal understanding. Moreover, similar to the trend from robot learning experiments in Section 4.2, we find our framework outperforms SiamMAE. This again demonstrates the benefit of our approach over masked visual modeling approaches for image representation learning from videos.

4.4. Ablation Study and Analysis

We provide extensive ablation studies and analysis to investigate the importance of our design choices for building our framework upon prior work (Denton & Fergus, 2018). Due to computational constraints, we report the performance on the DAVIS benchmark.

Comparison with deterministic frame prediction To investigate the importance of *stochastic* future prediction,

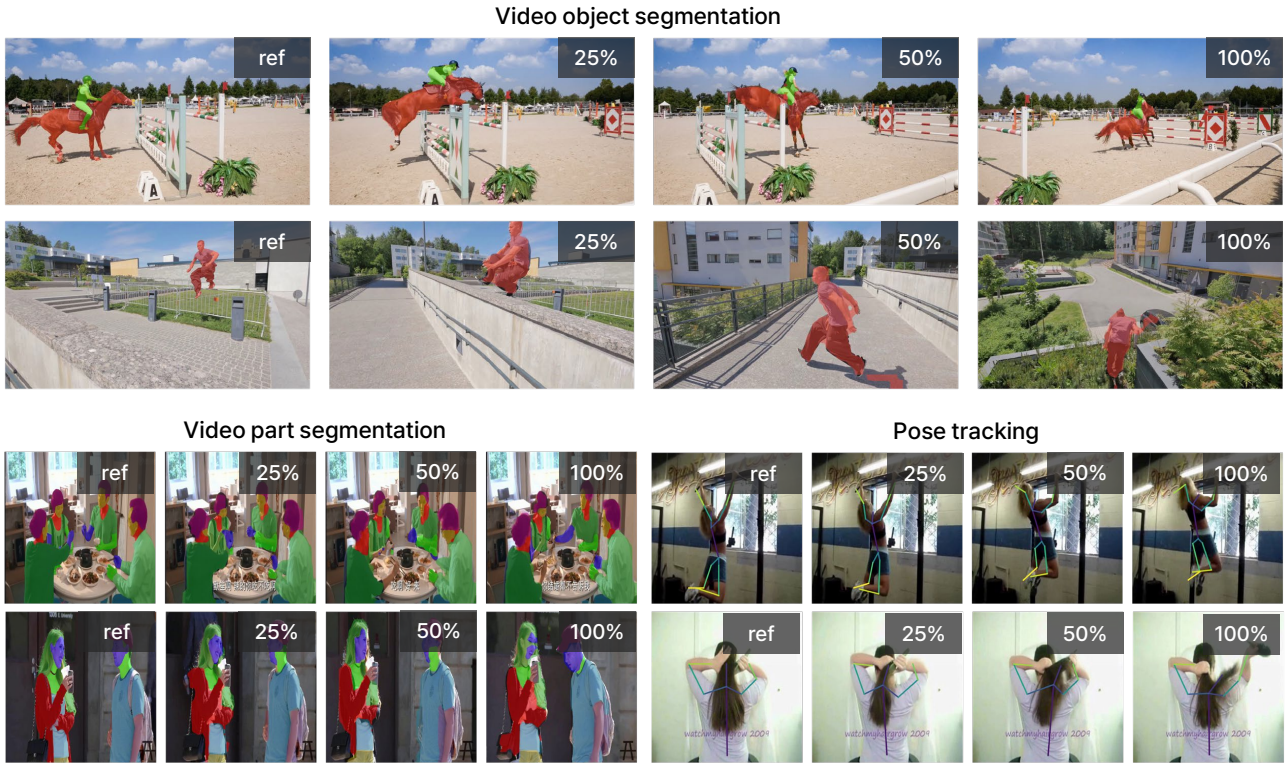


Figure 4: **Qualitative results.** We provide examples of predicted propagation from RSP on video object segmentation (Pont-Tuset et al., 2017), video part segmentation (Zhou et al., 2018), and pose tracking (Jhuang et al., 2013) benchmarks. "ref" indicates the ground-truth annotations, and 25, 50, and 100% refers to the propagated ratio of the videos. We provide additional qualitative results in Appendix E.

we compare our framework with *deterministic* frame prediction model. For a fair comparison, we also use the auxiliary MAE objective with the shared decoder for both methods. In Table 3a, we find that the deterministic frame prediction model significantly underperforms our framework, *i.e.,* the deterministic baseline achieves 54.4% while our stochastic framework achieves 60.1%. This shows that deterministic frame predictor struggles to learn useful representations from complex large video datasets like Kinetics-400 (Kay et al., 2017). On the other hand, our method can learn such representations by learning to predict possible multiple futures via stochastic frame prediction.

Latent variable design We explore two design choices on the stochastic latent variable $\mathbf{z}_{t+k}$. Specifically, we consider two variants that employ *Gaussian* distribution or a set of *Categorical* variables (Hafner et al., 2021a). Interestingly, in Table 3b, we find that utilizing the Categorical variable significantly outperforms the variant with Gaussian distribution. We hypothesize this is because it is easier to predict discrete labels compared to accurately approximating continuous Gaussian distribution. In addition, we would like to note that Meyer et al. (2023) demonstrated RL with discrete representations outperforms continuous representations when the environment dynamics gets more complex. This could also explain our observation because the Kinetics-400 (Kay et al., 2017) dataset consists of complex real-world videos. Given this result, it would be an interesting future direction to design models with a more expressive prior, *e.g.,* autoregressive prior.

Auxiliary MAE objective with shared decoder One important design in our framework is introducing the *auxiliary MAE objective* to learn dense representation within the frames, which might not be learned by the frame prediction objective. In Table 3c, we observe that our framework indeed outperforms a baseline that does not introduce the auxiliary objective by a large margin (+2.4%p). Moreover, to investigate the importance of having a *shared decoder*, we design a parallel decoder baseline that has an additional, separate decoder for the auxiliary MAE objective. We find that having a shared decoder is crucial for making both objectives synergistic, *i.e.,* our framework with the shared decoder achieves 60.1% while the parallel decoder baseline achieves 58.1%. This result is intriguing because our shared decoder design also has the benefit of being parameter-efficient compared to the parallel decoder.

Stochastic	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
✗	54.4	50.7	58.1
✓	**60.1**	**57.4**	**62.8**

(a) **Deterministic prediction.** We find our stochastic future frame prediction objective is crucial for representation learning from complex real-world videos.

Latent	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Gaussian	54.1	52.9	55.9
Categorical	**60.1**	**57.4**	**62.8**

(b) **Stochastic latent variable.** Using a set of categorical variables (Hafner et al., 2021a) for the latent variable outperforms a baseline that employs Gaussian distribution.

w/ MAE	Decoder	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
✗	-	57.7	54.9	60.5
✓	Separate	58.1	55.4	60.7
✓	Shared	**60.1**	**57.4**	**62.8**

(c) **Auxiliary MAE objective with a shared decoder.** We find that training with the auxiliary MAE objective works better, especially when combined with our shared decoder design.

KL scale	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
0.1	56.1	52.9	59.3
0.01	**60.1**	**57.4**	**62.8**
0.001	59.1	56.6	61.5

(d) **KL objective scale.** Training a model with too strong or weak KL objectives leads to worse performance.

Table 3: **Ablation studies.** We report the performance of various variants of RSP on DAVIS benchmark. For all experiments, we pre-train ViT-S/16 model on Kinetics-400 dataset for 400 epochs. Default settings are highlighted in gray.

Same aug	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
✗	53.7	52.2	55.2
✓	**60.1**	**57.4**	**62.8**

(a) **Applying the same augmentation.** Applying augmentations (*i.e.,* random resized crop and horizontal flip) differently to current and future frames significantly degrades performance.

Future frame aug	Scale	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
None	-	58.3	56.1	60.6
Masking	0.75	57.7	54.8	60.6
Masking	0.95	55.8	52.7	58.9
Noise	0.1	58.4	56.0	60.7
Noise	0.5	**60.1**	**57.4**	**62.8**
Noise	1.0	58.9	56.3	61.4

(b) **Future frame augmentation.** Applying a mild augmentation to a future frame can enhance performance. But strong augmentation such as masking degrades the performance.

Table 4: **Effect of data augmentation.** We investigate (a) the importance of applying the same augmentation to current and future frames and (b) the effect of applying mild augmentation to the future frame. Default settings are highlighted in gray.

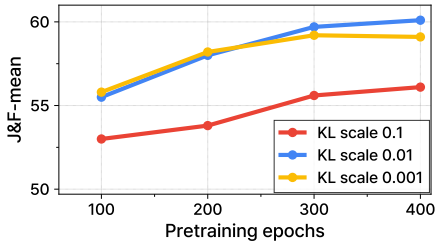


Figure 5: **Effect of KL loss scale.** We report the learning curves of models trained with different KL loss scales (β).

Effect of KL loss scale We also conduct analysis on the effect of the KL loss scale (β) to provide a deeper understanding of the learning dynamics of our framework. In Table 3d, we observe that too strong or weak KL loss scales lead to worse performance. This is because high β makes it difficult to learn good posterior by enforcing distributions to be close too early, *i.e.,* before the model starts to learn a good posterior distribution, which leads to overall worse performance as shown in Figure 5. On the other hand, low β makes the posterior distribution tend to ignore the prior distribution, and this consequently makes it difficult for the prior model to predict the posterior, which leads to lower asymptotic performance as shown in Figure 5.

Applying the same augmentation As we previously mentioned in Section 3.1, applying the same augmentation to both current and future frames is crucial for making the frame prediction objective valid. For instance, applying the random horizontal flipping augmentation differently to current and future frames would make it impossible to predict the future frame. In Table 4a, we indeed find that applying different augmentations to current and future frames significantly degrades the performance.

Additional future frame augmentation We study the effect of our design choice that augments the future frame by adding a small Gaussian noise in Table 4b. We also explore another augmentation scheme of applying masks to future frames, similar to Gupta et al. (2023). We find that applying masking augmentation degrades the performance, exhibiting a similar trend in Table 4a. This is because the prior have to also capture the stochasticity from aggressive masking augmentation, which makes it difficult to learn meaningful prior distribution. On the other hand, adding a small Gaussian noise can effectively improve the performance by delivering the benefit of augmentation, as it does not change the semantic meaning of frames.

5. Conclusion

In this work, we present RSP, a framework for visual representation learning from videos, that learns representations that capture temporal information between frames by training a stochastic future frame prediction model. Our key contribution lies in re-visiting the idea of stochastic video generation (Denton & Fergus, 2018) and re-designing it for representation learning by exploring and adopting various design choices. Our extensive experiments demonstrate that our framework consistently achieves competitive or superior performance to various baselines. We hope our work further facilitates research on representation learning from videos via future frame prediction.

Limitations and future directions One limitation of our work is that the quality of generated frames is not of high quality, though our focus is not on high-fidelity generation. Given this, it would be an interesting direction to incorporate recent video generative models based on diffusion models, similar to Hudson et al. (2023) that learns representations via image diffusion models. Moreover, due to computational constraints, our work does not include large-scale experiments with longer training budgets and larger models. Scaling up our approach would be an interesting future direction. Finally, an extension of our framework to multiple frames is a future direction we are keen to explore.

Impact Statement

This paper presents a framework for representation learning via generative modeling of videos. Thus there is a risk of potential misuse of our model for malicious purposes, *e.g.,* generating fake videos. However, unlike other high-fidelity generative models, our model generates outputs that are clearly distinguishable from real frames. This significantly reduces the risk of our model being used for generating fake videos. Nonetheless, it is still important to recognize and state such potential risk of misuse as the potential extension of our work is likely to have the capability to learn strong representations while generating high-quality videos.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST); No.RS-2021-II212068, Artificial Intelligence Innovation Hub and Samsung Electronics Co., Ltd (IO201211-08107-01).

References

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A. C., and Bellemare, M. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, 2021.

Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R. H., and Levine, S. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017.

Bao, H., Dong, L., Piao, S., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Bauer, S., Wüthrich, M., Widmaier, F., Buchholz, A., Stark, S., Goyal, A., Steinbrenner, T., Akpo, J., Joshi, S., Berenz, V., et al. Real robot challenge: A robotics competition in the cloud. In *NeurIPS 2021 Competitions and Demonstrations Track*, 2022.

Benaim, S., Ephrat, A., Lang, O., Mosseri, I., Freeman, W. T., Rubinstein, M., Irani, M., and Dekel, T. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9922–9931, 2020.

Bengio, Y., Ducharme, R., and Vincent, P. A neural probabilistic language model. *Advances in neural information processing systems*, 2000.

Bengio, Y., Léonard, N., and Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 2020.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 2021a.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative pretraining from pixels. In *International conference on machine learning*, 2020a.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 2020b.

Chen, X., Xie, S., and He, K. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021b.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Denton, E. and Fergus, R. Stochastic video generation with a learned prior. In *International Conference on Machine Learning*, 2018.

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

Doersch, C., Gupta, A., and Efros, A. A. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, 2015.

Donahue, J. and Simonyan, K. Large scale adversarial representation learning. *Advances in neural information processing systems*, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R., and He, K. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3299–3309, 2021.

Feichtenhofer, C., Li, Y., He, K., et al. Masked autoencoders as spatiotemporal learners. In *Advances in neural information processing systems*, 2022.

Finn, C., Goodfellow, I., and Levine, S. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 2016.

Ge, C., Liang, Y., Song, Y., Jiao, J., Wang, J., and Luo, P. Revitalizing cnn attention via transformers in self-supervised visual representation learning. *Advances in Neural Information Processing Systems*, 34:4193–4206, 2021.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Guo, S., Xiong, Z., Zhong, Y., Wang, L., Guo, X., Han, B., and Huang, W. Cross-architecture self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19270–19279, 2022.

Gupta, A., Kumar, V., Lynch, C., Levine, S., and Hausman, K. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019.

Gupta, A., Wu, J., Deng, J., and Fei-Fei, L. Siamese masked autoencoders. In *Advances in Neural Information Processing Systems*, 2023.

Hafner, D., Lillicrap, T., Norouzi, M., and Ba, J. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021a.

Hafner, D., Lillicrap, T. P., 0002, M. N., and Ba, J. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021b.

Han, T., Xie, W., and Zisserman, A. Memory-augmented dense predictive coding for video representation learning. In *European conference on computer vision*, pp. 312–329. Springer, 2020a.

Han, T., Xie, W., and Zisserman, A. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020b.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Hoffer, E., Ben-Nun, T., Hubara, I., Giladi, N., Hoefler, T., and Soudry, D. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Hudson, D. A., Zoran, D., Malinowski, M., Lampinen, A. K., Jaegle, A., McClelland, J. L., Matthey, L., Hill, F., and Lerchner, A. Soda: Bottleneck diffusion models for representation learning. *arXiv preprint arXiv:2311.17901*, 2023.

Jabri, A., Owens, A., and Efros, A. Space-time correspondence as a contrastive random walk. In *Advances in neural information processing systems*, 2020.

James, S. and Davison, A. J. Q-attention: Enabling efficient learning for vision-based robotic manipulation. *IEEE Robotics and Automation Letters*, 2022.

James, S., Ma, Z., Arrojo, D. R., and Davison, A. J. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

James, S., Wada, K., Laidlow, T., and Davison, A. J. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Jhuang, H., Gall, J., Zuffi, S., Schmid, C., and Black, M. J. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, 2013.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Lai, Z. and Xie, W. Self-supervised learning for video correspondence flow. *arXiv preprint arXiv:1905.00875*, 2019.

Li, T., Chang, H., Mishra, S. K., Zhang, H., Katabi, D., and Krishnan, D. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Li, X., Liu, S., De Mello, S., Wang, X., Kautz, J., and Yang, M.-H. Joint-task self-supervised learning for temporal correspondence. In *Advances in Neural Information Processing Systems*, 2019.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Majumdar, A., Yadav, K., Arnaud, S., Ma, Y. J., Chen, C., Silwal, S., Jain, A., Berges, V.-P., Abbeel, P., Malik, J., et al. Where are we in the search for an artificial visual cortex for embodied intelligence? In *Advances in neural information processing systems*, 2023.

Meyer, E., White, A., and Machado, M. C. Harnessing discrete representations for continual reinforcement learning. *arXiv preprint arXiv:2312.01203*, 2023.

Misra, I., Zitnick, C. L., and Hebert, M. Shuffle and learn: unsupervised learning using temporal order verification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 527–544. Springer, 2016.

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

Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pp. 69–84. Springer, 2016.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Pan, T., Song, Y., Yang, T., Jiang, W., and Liu, W. Videomoco: Contrastive video representation learning with

temporally adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11205–11214, 2021.

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Pomerleau, D. A. Alvinn: An autonomous land vehicle in a neural network. In *Advances in neural information processing systems*, 1988.

Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., and Van Gool, L. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Qian, R., Meng, T., Gong, B., Yang, M.-H., Wang, H., Belongie, S., and Cui, Y. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6964–6974, 2021.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., and Levine, S. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In *Robotics: Science and Systems*, 2018.

Seo, Y., Kim, J., James, S., Lee, K., Shin, J., and Abbeel, P. Multi-view masked world models for visual robotic manipulation. In *International Conference on Machine Learning*, 2023.

Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., and Brain, G. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, 2018.

Shridhar, M., Manuelli, L., and Fox, D. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, 2023.

Srivastava, N., Mansimov, E., and Salakhudinov, R. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, 2015.

Tassa, Y., Tunyasuvunakool, S., Muldal, A., Doron, Y., Liu, S., Bohez, S., Merel, J., Erez, T., Lillicrap, T., and Heess, N. dm_control: Software and tasks for continuous control. *arXiv preprint arXiv:2006.12983*, 2020.

Tong, Z., Song, Y., Wang, J., and Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in neural information processing systems*, 2022.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Villegas, R., Babaeizadeh, M., Kindermans, P.-J., Moraldo, H., Zhang, H., Saffar, M. T., Castro, S., Kunze, J., and Erhan, D. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

Vondrick, C., Pirsiavash, H., and Torralba, A. Generating videos with scene dynamics. *Advances in neural information processing systems*, 2016.

Wang, X. and Gupta, A. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2015.

Wang, X., Jabri, A., and Efros, A. A. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., and Zhuang, Y. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10334–10343, 2019.

Xu, J. and Wang, X. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10075–10085, 2021.

Yan, W., Zhang, Y., Abbeel, P., and Srinivas, A. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 2019.

Yu, S., Sohn, K., Kim, S., and Shin, J. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., and Levine, S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, 2020a.

Yu, W., Lu, Y., Easterbrook, S., and Fidler, S. Efficient and information-preserving future frame prediction and beyond. In *International Conference on Learning Representations*, 2020b.

Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pp. 649–666. Springer, 2016.

Zhou, Q., Liang, X., Gong, K., and Lin, L. Adaptive temporal encoding network for video instance-level human parsing. In *Proceedings of the 26th ACM international conference on Multimedia*, 2018.

A. Implementation Details

We build our framework upon the official implementation of MAE (He et al., 2022).[2] We summarize our hyperparameters of pre-training and video label propagation in Table 5. We follow Hafner et al. (2021b) for various design choices with regard to stochastic latent variable. Specifically, we employ a set of 32 Categorical variables with 32 classes for the posterior and prior distributions. Furthermore, to prevent over-regularizing the representations towards an inadequately trained prior, we incorporate KL balancing with a ratio of $\alpha = 0.8$, as introduced in Hafner et al. (2021b).

config	value
optimizer	AdamW (Loshchilov & Hutter, 2019)
optimizer momentum	β_1, β_2 = 0.9, 0.95 (Chen et al., 2020a)
optimizer weight decay	0.05
learning rate	1.5e-4
learning rate scheduler	Cosine decay (Loshchilov & Hutter, 2017)
warmup epochs (Goyal et al., 2017)	40
pre-train epochs	400
repeated sampling (Hoffer et al., 2020)	2
batch size	1536
frame sampling gap	[4, 48]
augmentation	hflip, crop [0.5, 1.0]
Discrete latent dimensions	32
Discrete latent classes	32
KL balancing ratio	0.8

(a) **Pre-training hyperparameters**

config	DAVIS	VIP	JHMDB
top-k	7	7	10
neighborhood size	30	5	5
queue length	30	3	30

(b) **Evaluation hyperparameters**

Table 5: **Hyperparameter details of pre-training and evaluation**

Architectural details We use standard ViT-S/16 (Dosovitskiy et al., 2021) as our encoder. For the decoder, each block is composed of cross-attention, self-attention, and feed-forward MLP layers. The hyperparameters for the decoder, including embedding dimension, depth, and the number of heads, are aligned with those specified in He et al. (2022).

B. Additional Ablation Study and Analysis

We provide additional ablation studies and analysis to investigate the importance of our design choices. We report the performance on the DAVIS benchmark in Table 6.

Projection	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Same	56.6	54.3	58.9
Distinct	**60.1**	**57.4**	**62.8**

(a) **Encoder-decoder projection.** We find that distinct projections for stochastic frame prediction and auxiliary MAE objective is crucial for learning representation.

Concat	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Channel dim	54.1	52.9	55.9
Tokens	**60.1**	**57.4**	**62.8**

(b) **Concatenating latent variable and patch representations.** We find that concatenating the latent variable and patch representations along the channel dimension works better than concatenating them along the channel dimension.

Table 6: **Ablation studies.** We report the performance of various variants of RSP on DAVIS benchmark. For all experiments, we pre-train ViT-S/16 model on Kinetics-400 dataset for 400 epochs. Default settings are highlighted in gray.

[2] https://github.com/facebookresearch/mae

C. Experimental Results with 95% Confidence Interval

We here provide the experimental results of Table 1 with 95% confidence intervals in Table 7.

Table 7: **Results on vision-based robot learning.** Performance of imitation learning agents on CortexBench (Majumdar et al., 2023), RLBench (James et al., 2020), and Franka Kitchen (Gupta et al., 2019) with a 95% confidence interval. We have 5, 4, and 4 runs for CortexBench, RLBench, and Franka Kitchen respectively.

(a) CortexBench

Method	Adroit	MetaWorld	DMC	Trifinger
SimCLR	40.4±3.3	78.4±5.2	39.7±2.9	63.3±3.3
MoCo v3	39.6±4.3	65.4±8.0	43.7±3.2	53.3±1.6
Dino	**45.6**±6.2	82.4±5.8	50.9±1.5	64.2±3.5
MAE	44.8±4.3	81.4±6.3	52.1±3.7	62.2±5.0
SiamMAE	44.0±6.6	81.1±6.3	56.0±2.9	52.1±7.6
RSP (Ours)	**45.6**±4.6	**84.5**±6.6	**61.6**±3.4	**66.2**±0.8

(b) RLBench

Method	Button	Saucepan	Phone	Umbrella	Wine	Rubbish
SimCLR	7.4±2.6	39.5±2.2	34.6±6.6	5.8±3.3	11.0±2.1	5.2±1.2
MoCo v3	11.4±4.1	45.8±3.9	36.2±3.4	13.2±1.5	8.7±0.7	6.7±0.8
Dino	24.7±1.5	57.9±5.9	32.0±5.5	28.1±1.4	31.4±1.5	12.9±1.5
MAE	6.4±2.2	36.8±6.4	37.7±1.9	10.0±1.2	10.0±2.1	6.2±3.2
SiamMAE	6.1±2.3	22.5±0.8	5.4±0.5	4.0±0.0	8.7±0.8	3.5±0.9
RSP (Ours)	**28.4**±3.0	**93.4**±1.8	**48.0**±4.6	**37.3**±3.0	**31.9**±2.3	**18.5**±1.1

(c) Franka Kitchen

Method	Knob1 on	Light on	Sdoor open	Ldoor open	Micro open
SimCLR	25.3±2.1	**55.8**±6.4	72.3±2.8	17.0±2.9	23.3±2.8
MoCo v3	11.5±3.9	24.3±5.0	66.5±3.2	10.3±2.1	14.3±2.5
Dino	27.0±3.2	44.3±6.5	77.0±5.0	16.5±2.5	28.5±4.8
MAE	12.0±3.3	24.3±4.2	71.5±4.3	12.8±3.9	10.0±2.8
SiamMAE	16.8±4.4	36.5±7.0	68.0±7.9	17.3±3.7	13.5±4.8
RSP (Ours)	**31.0**±2.4	44.5±5.6	**82.5**±2.7	**28.8**±4.8	**30.3**±5.6

D. Comparison with ImageNet Pre-trained SSLs

Table 8: **Results on video label propagation.** We report performances on video segmentation, video part segmentation, and pose tracking tasks from DAVIS (Pont-Tuset et al., 2017), VIP (Zhou et al., 2018), and JHMDB (Jhuang et al., 2013) benchmarks, respectively. We compare the Kinetics-400 pre-trained approaches to the ImageNet pre-trained approaches as a reference.

		DAVIS			VIP	JHMDB	
Method	Architecture	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$	mIoU	PCK@0.1	PCK@0.2
Kinetics-400 pre-trained							
SimCLR (Chen et al., 2020b)	ViT-S/16	53.9	51.7	56.2	31.9	37.9	66.1
MoCo v3 (Chen et al., 2021b)	ViT-S/16	57.7	54.6	60.8	32.4	38.4	67.6
Dino (Caron et al., 2021)	ViT-S/16	59.5	56.5	62.5	33.4	41.1	70.3
MAE (He et al., 2022)	ViT-S/16	53.5	50.4	56.7	32.5	43.0	71.3
SiamMAE (Gupta et al., 2023)	ViT-S/16	58.1	56.6	59.6	33.3	**44.7**	73.0
RSP (Ours)	ViT-S/16	**60.1**	**57.4**	**62.8**	**33.8**	44.6	**73.4**
RSP (Ours)	ViT-B/16	60.5	57.8	63.2	34.0	46.0	74.6
ImageNet pre-trained							
Dino (Caron et al., 2021)	ViT-S/16	61.8	60.2	63.4	36.2	45.6	75.0
MAE (He et al., 2022)	ViT-B/16	53.5	52.1	55.0	28.1	44.6	73.4

E. Additional Qualitative Results

Figure 6: **Additional qualitative results.** We provide more qualitative results of predicted propagation from RSP on DAVIS video object segmentation (Pont-Tuset et al., 2017) benchmarks. ”ref” indicates the ground-truth annotations, and 25, 50, and 100% refers to the propagated ratio of the videos.