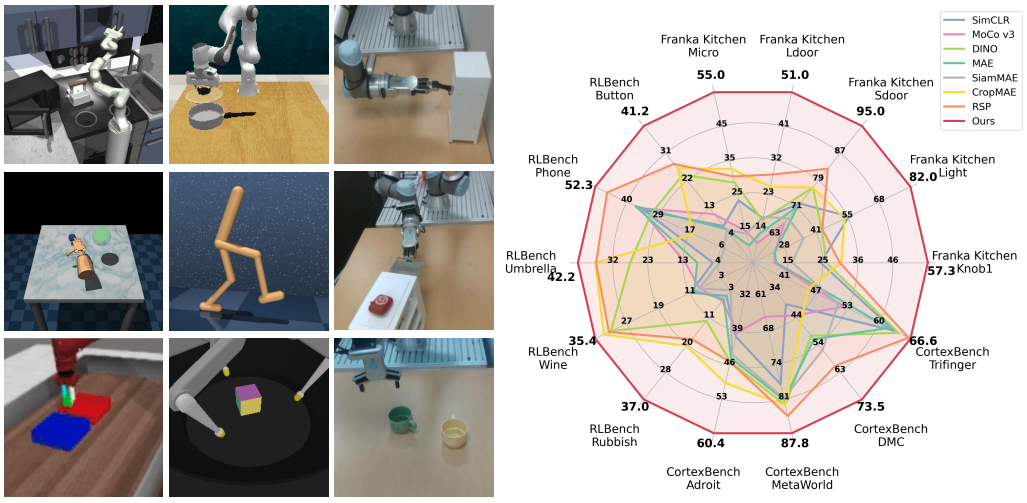
# SELF-SUPERVISED VISUAL STATE REPRESENTATION LEARNING FOR ROBOTICS FROM DYNAMIC SCENES

**Taekyung Kim**[1] **Dongyoon Han**[1] **Byeongho Heo**[1] **Jeongeun Park**[2] **Sangdoo Yun**[1]

[1]NAVER AI Lab [2]Korea University

## ABSTRACT

In robot policy learning, deriving informative state representations encompassing visual and proprioceptive representations is critical. While proprioceptions are acquired from internal sensors, visual state representations primarily rely on vision backbones. Therefore, leveraging a strong backbone generalized across diverse tasks and environments is essential for effective robotic perception. Self-supervised learning (SSL) has been a promising approach for pre-training such backbones. However, conventional SSL approaches for visual representation learning have predominantly focused on learning capability for a comprehensive understanding of a whole image or video, far from requisites for robotics such as seamless interactions. Bearing this in mind, we introduce a novel and intuitive self-supervised visual state representation learning pipeline designed to facilitate the acquisition of state representations through masked autoencoding. Our method implicitly dissolves the forming process of the state representations into the encoding process without any additional layers. Extensive experiments in diverse simulated environments demonstrate the superiority of our method in robot manipulation and locomotion tasks over previous baselines. Moreover, deploying our pre-trained model on physical robots confirms its robustness and effectiveness in real-world settings.




(a) Simulation (b) Real-world (c) Overall performance comparison

Figure 1: We validate the effectiveness of our method on (a) simulated environments and (b) real-world environments. (c) Our method significantly surpasses previous self-supervised visual representation learning methods designed for static (Chen et al., 2020a; 2021; Caron et al., 2021; He et al., 2022) and dynamic scenes (Gupta et al., 2023; Weinzaepfel et al., 2022; Jang et al., 2024) on various robot manipulation and locomotion tasks. More details are in §4.2 and §4.3.

## 1 INTRODUCTION

With the increasing interest in deploying robots in real-world environments, enabling seamless interactions with their surroundings has become a crucial challenge. Such interactions necessitate *state representations* that effectively capture both visual and proprioceptive information. While robots acquire their proprioceptions through internal sensors, visual state representations are predominantly obtained from the vision backbones. Therefore, leveraging a strong and robust backbone capable of generalizing across diverse tasks and environments has emerged as a key consideration in robotics.

Self-supervised learning (SSL) of visual representations has been highlighted as pivotal research in vision domains, with pre-trained models being widely adopted for effective backbone deployment. A series of studies have introduced promising recipes for learning image (Grill et al., 2020; Chen et al., 2020a; 2021; Caron et al., 2020; 2021; Chen & He, 2021; He et al., 2022) and video representations (Tong et al., 2022; Pan et al., 2021) without labeled data, enabling scene comprehension and a cohesive action understanding, respectively. However, robots require representations that encompass both a comprehensive understanding of the scene and an awareness of temporal evolution across consecutive observations, an aspect overlooked by previous approaches.

A sequence of studies has attempted to address the challenges. SiamMAE (Gupta et al., 2023) introduces a visual representation learning approach that leverages dynamic scenes. It address a key limitation of MAE (He et al., 2022), which lacks consideration on learning correspondence during pre-training, leading to poor similarity estimation despite strong localization capability. By incorporating implicit guidance for learning correspondence across different timestamps within the masked autoencoding process, SiamMAE enhances understanding of temporal evolution. However, it remains insufficient in forming effective state representations of timestamps beneficial for the instant action prediction.

In light of this, we introduce a simple and intuitive approach called State Representation Learning (SRL), which fully leverages the capabilities of backbone models to preserve the scene information in an efficient form. Specifically, we guide the model to rely heavily on the state representations during decoding, enforcing the decoder to predict unseen patches based on the state representations leads the model to store visual perception information effectively. In addition, our approach seamlessly integrates the formation of state representations into the encoding process, eliminating the need for additional layers.

Through extensive experiments across diverse simulated environments, we showcase the effectiveness of our method in robot manipulation and locomotion tasks, surpassing previous baselines (Chen et al., 2020a; 2021; Caron et al., 2021; He et al., 2022; Weinzaepfel et al., 2022; Gupta et al., 2023; Jang et al., 2024; Eymaël et al., 2025) with significant gaps (see Fig. 1). Furthermore, we extend our validation to real-world environments by deploying our pre-trained models on physical robots, demonstrating their robustness and superiority even in real-world scenarios.

## 2 RELATED WORK

**Self-supervised learning on a static scene** Self-supervised learning (SSL) approaches have been widely explored in the image domain. Contrastive learning approaches (Chen et al., 2020a; He et al., 2019; Chen et al., 2020b; 2021; Caron et al., 2020) aim to learn useful representations by maximizing the similarity between positive pairs derived from a static scene through strong augmentations. Although these methods excel in facilitating a cohesive understanding of images, they suffer from limited localization capabilities (Kim et al., 2024), essential for action prediction in robotics. On the other hand, masked image modeling (MIM) (Bao et al., 2021; He et al., 2022; Xie et al., 2022; Baevski et al., 2022; Kim et al., 2024) has recently gained attention for its promising capacity to learn visual representations through predictive learning. Inspired by masked language modeling (MLM) in transformers (Devlin et al., 2018), BEiT (Bao et al., 2021) extends MLM into the vision domain, adopting an external offline tokenizer. MAE (He et al., 2022) and SimMIM (Xie et al., 2022) showcase efficient MIM by directly reconstructing masked input pixels without any tokenizer. However, despite its strong capability in localization, models pre-trained by MIM approaches tend to show limited disriminatibility without additional post-training, showing lower linear probing performance compared to contrastive SSL approaches. Moreover, while the learning strategies of the

models determine the behaviors of representations, these models lacks capturing temporal evolution across scenes during training.

**Self-supervised learning on dynamic scenes** To enhance the capability to capture temporal evolution, SiamMAE (Gupta et al., 2023) proposes visual representation learning methods that utilize dynamic scenes. RSP employs stochastic frame prediction tasks along with masked autoencoding. On the other hand, CropMAE (Eymaël et al., 2025) introduce a simple augmentation strategy that enables the generation of dynamic scenes even from a single static scenes, While these approaches primarily focused on learning to capture temporal progression, we aim to encode visual state representations from the observations in a self-supervised manner.

# 3 METHOD

## 3.1 MOTIVATION

For precise action prediction in robotics, the visual state representation should comprehensively store both fine-grained and semantic information from observations while enabling the recognition of temporal evolution across consecutive observations. However, previous self-supervised learning (SSL) approaches have not fully addressed these aspects, suboptimal for robot backbones. In this section, we discuss the pros and cons of these approaches from a robot backbone perspectives.

**Limited temporal evolution awareness of MAE.** MAE (He et al., 2022) has been highlighted with its strong capability in localization, which stems from its design that enforces the autoencoder to predict missing information based on available prior information (i.e., visible patches). This pipeline implicitly encourages the encoder to facilitate interactions among the remaining sparse tokens, thereby enhancing its localization capability. However, since MAE performs predictive learning on a single static scene, the encoder is not explicitly trained to handle dynamic transitions over time, leading to limited performance in sequential scene understanding tasks. Moreover, a recent study reveals that MAE falls short in learning broader contexts (Kim et al., 2024), leading to representations with a limited cohesive understanding of observed scenes. This limitation further constrains its potential to effectively understand sequential scenes.

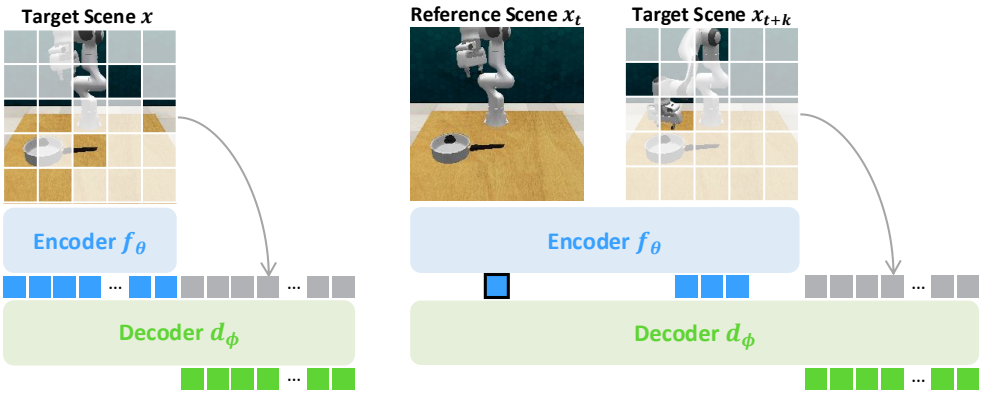
**Limited state representations formation of SiamMAE.** To alleviate the chronic limitation of SSL approaches on static scenes, SiamMAE leverages dynamic scenes to build non-trivial problems for explicit correspondence learning by sampling scenes from sequential data or collecting scenes from various camera perspectives. The core of these pipelines is the use of source and target scenes with additional cross-attention layers within decoders. The cross-attention layer guides the encoder to form discriminative similarity that enables capturing fine-grained token-wise similarity for accurate correspondence matching. However, while the SiamMAE pipeline enables capturing correspondences among consecutive scenes, it is fundamentally not capable of learning the model to form visual state representations, eventually storing insufficient information in its state representations and thus negatively affecting the execution of robotic tasks.

## 3.2 PRELIMINARY

**Masked autoencoding** Given a scene $x$, we patchify into $N$ non-overlapping patches $\{x\}_{i=1}^{N}$. We mask the patches by randomly selecting a set of masked patches $\mathcal{M} \subset \{1, 2, ..., N\}$ with a masking ratio $r \in (0, 1)$ and exclude them, where $\mathcal{M} \subset \{1, 2, ..., N\}$ and $|\mathcal{M}| = \lfloor rN \rfloor$. The remaining patches $\{x\}_{i \in \mathcal{M}^c}$ are concatenated with a learnable CLS token $e_{[CLS]}$ and fed into the encoder $f_\theta$, becoming spatial token representations $\{\mathbf{u}\}_{i \in \mathcal{M}^c}$ with an encoded CLS token. The encoded tokens are concatenated with mask tokens $\mathbf{m}_i$ corresponding to the position of $i$-th masked patches. The decoder $d_\phi$ predicts the excluded image patches $\{x_i\}_{i \in \mathcal{M}}$ by referring to the information from unmasked tokens.

## 3.3 STATE REPRESENTATION LEARNING (SRL)

**Claim** Regarding the perspectives in §3.1, our ultimate goal is to derive effective visual state representations capable of capturing temporal evolution from visual observations during the encoding




(a) Masked Autoencoder (b) State Representation Learning

Figure 2: **Overview of our proposed State Representation Learning (SRL).** (a) Masked Autoencoder (MAE) (He et al., 2022) learns visual representations by enforcing token interactions over scarce tokens via predictive learning on masked patches. (b) In contrast, our SRL reconstructs the masked patches from the state representation of the reference scene and a few patch tokens from the target scene. Such extreme scarcity of the target scene leads the decoder to rely heavily on the reference scene, facilitating the preservation of observed information in the state representation.

process. To accomplish this, our state representation learning (SRL) considers whether (1) visual observations can be reconstructed from the state representations and (2) temporal relationships are inherently preserved in the state representations. Ensuring both aspects simultaneously, we enforce the encoder to reconstruct the corrupted scenes from the state representation of a consecutive scene. We apply an extremely high masking ratio to the target scene, thereby restricting the decoder from precisely reconstructing the masked target patches solely based on the unmasked target tokens. In our pipeline, the quality of reconstruction directly linked to the ability of the encoder to effectively store visual information and map the representations into an embedding space capable of capturing temporal evolution. Therefore, minimizing the reconstruction objective is expected to facilitate the formation of more informative state representations.

**Overall pipeline** Suppose we sample a reference scene $x_t$ and a target scene $x_{t+k}$ with a gap in the temporal index $k$ from a given video. We patchify $x_t$ and $x_{t+k}$ into $N$ non-overlapping patches $\{x_{t,i}\}_{i=1}^N$ and $\{x_{t+k,i}\}_{i=1}^N$, respectively. We mask the target scene with an extremely high masking ratio $r \in (0,1)$ so that and remove them, where $\mathcal{M} \subset \{1,2,...,N\}$ and $|\mathcal{M}| = \lfloor rN \rfloor$. The remaining patches $\{x_{t+k,i}\}_{i\in\mathcal{M}^c}$ are concatenated with a learnable state token $e_{[state]}$ and fed into the encoder $f_\theta$, becoming a visual state representation $z_{t+k}$ and spatial token representations $\{\mathbf{u}_{t+k,i}\}_{i\in\mathcal{M}^c}$ for the target scene. The visual state representation $z_t$ and token representations $\{\mathbf{u}_{t,i}\}_{i=1}^N$ for the reference scene are derived in the same manner without masking. We concatenate the encoded state token $z_t$ of the reference scene $x_t$ with spatial token representations $\{\mathbf{u}_{t+k,i}\}_{i\in\mathcal{M}^c}$ of the target scene $x_{t+k}$ and mask tokens $\mathbf{m}_i$ corresponding to the position of $i$-th masked patches. The decoder $d_\phi$ predicts the excluded image patches $\{x_{t+k,i}\}_{i\in\mathcal{M}}$ by referring to the information from $z_t$ and $\{\mathbf{u}_{t+k,i}\}_{i\in\mathcal{M}^c}$, eventually obtaining the $i$-th decoded mask token $\mathbf{m}_i^d$, where $T_d \cup \{\mathbf{m}_i^d\}_{i\in\mathcal{M}} = d_\phi(z_t, \{\mathbf{u}_{t+k,i}\}_{i\in\mathcal{M}^c}, \{\mathbf{m}_i\}_{i\in\mathcal{M}})$. Here, $T_d$ is a set of decoded unmasked tokens. However, due to the excessive scarcity in $\{\mathbf{u}_{t+k,i}\}_{i\in\mathcal{M}^c}$, the decoder $d_\phi$ proactively rely on $z_t$, which enable the encoder $f_\theta$ to form richer state representations. We minimize the reconstruction loss $\mathcal{L}_{\text{SRL}} = \sum_{i\in\mathcal{M}} d(\mathbf{m}_i^d, x_{t+k,i})$ throughout the training procedure where $d(\cdot)$ is a distance function. Our proposed pipeline is illustrated in Fig. 2b.

**Decoder structure** Previous methods in dynamic SSL utilize cross-attention layers as a core component, placing them within the decoders to guide the encoder to learn representations that effectively capture correspondences. These approaches leverage a hybrid structure of cross-attention layers, self-attention layers, and multi-layer perceptrons (MLP) layers. In contrast, we employ self-attention layers that exclusively attend to the given information, with MLP layers for progressive transformation from representation embedding spaces into the pixel space.

# 4 EXPERIMENTS

In this section, we evaluate the impacts of the proposed method in policy learning for robotic manipulation and locomotion across various simulated environments (Gupta et al., 2019; James et al., 2020; Majumdar et al., 2023). We extend our validation to real-world settings by deploying our pre-trained model on physical robots, showcasing its transferability.

## 4.1 EXPERIMENTAL SETUP

### 4.1.1 PRE-TRAINING

We pre-train ViT-S/16 (Dosovitskiy et al., 2021) on Kinetics-400 (Kay et al., 2017) for 400 epochs. We employ repeated sampling (Hoffer et al., 2020; Feichtenhofer et al., 2021) with a factor of 2 so that the models are indeed pre-trained for 200 epochs. We use AdamW optimizer (Loshchilov & Hutter, 2019) with a batch size of 1536, comprising dynamic scenes with a resolution of 224×224. These scenes are randomly sampled from videos at a rate of 30 FPS, with a temporal index gap ranging from 4 to 96. We simply apply random resized crop and horizontal flip to the scenes, aligning the cropping region across reference and target scenes. To drive the learning mechanism of our proposed method, we randomly mask the target scenes with an extremely high masking ratio of 0.9. Our decoder is composed of eight vision transformer blocks, i.e., each block contains self-attention layers and multi-layer perceptrons. We follow the default hyperparameters of the baselines for their pre-training on Kinetics-400 (Kay et al., 2017). More details are listed in the Appendix.

### 4.1.2 VISION-BASED ROBOT POLICY LEARNING

**Franka Kitchen.** We validate models pre-trained by our method and other baselines in five imitation learning tasks from the Franka Kitchen benchmark Gupta et al. (2019). Our experiments mainly follow the imitation learning evaluation setup in (Nair et al., 2022; Parisi et al., 2022). Specifically, we employ an agent comprising a frozen backbone initialized with pre-trained models and a policy network consisting of a two-layer MLP, with a batch normalization layer applied at the input stage. We define the state representation for the policy network as the combination of the visual representation and the robot proprioceptive. For the perception, we employ either a left or right camera with a 224×224 resolution while omitting depth. The policy network is trained with a standard behavior cloning loss. Training for each demonstration task progresses for 20,000 steps, with a periodic online evaluation in the simulated environment every 1,000 steps. We evaluate the highest success rates of each demonstration across four different seeds and report its average with a 95% confidence interval.

**RLBench.** We consider 5 manipulation tasks from RLBench (James et al., 2020). For each task, we generate 100 demonstrations and utilize them for training the agent. We employ a front camera with a 224×224 resolution. Point cloud information are excluded throughout all experiments. We employ the end-effector controller with path plannin. We evaluate the highest success rates of each demonstration across four different seeds.

**CortexBench.** We evaluate the models on four simulated environments from CortexBench (Majumdar et al., 2023). We consider two, five, five, and two demonstrations from Adroit, DeepMind Control (DMC) (Tassa et al., 2020), MetaWorld (Yu et al., 2020), and Trifinger, respectively. Proprioceptive data is utilized except the DMC benchmark. We mainly follow the experimental setups in (Majumdar et al., 2023). For each task, we train the agent for 100 epochs, with a periodic online evaluation in the simulated environment every 5 epochs. We report the normalized score for DMC and the highest

(a) Cabinet Opening (b) Drawer Closing (c) Cup Stacking

Figure 3: Task Description for Real-world Environments

Table 1: **Experimental results on vision-based robot policy learning on Franka Kitchen.** We report the performance of imitation learning agents on Franka Kitchen (Gupta et al., 2019), which are trained upon representations from the ViT-S/16 model pre-trained on Kinetics-400 (Kay et al., 2017) dataset. The success rates (%) are reported for all the tasks. We underline the second-best performance. We report the gains of our method over the second-best baseline.

| Method | Knob1 on | Light on | Sdoor open | Ldoor open | Micro open |
|---|---|---|---|---|---|
| SimCLR (Chen et al., 2020a) | 25.3±2.1 | <u>55.8</u>±6.4 | 72.3±2.8 | 17.0±2.9 | 23.3±2.8 |
| MoCo v3 (Chen et al., 2021) | 11.5±3.9 | 24.3±5.0 | 66.5±3.2 | 10.3±2.1 | 14.3±2.5 |
| DINO (Caron et al., 2021) | 27.0±3.2 | 44.3±6.5 | 77.0±5.0 | 16.5±2.5 | 28.5±4.8 |
| MAE (He et al., 2022) | 12.0±3.3 | 24.3±4.2 | 71.5±4.3 | 12.8±3.9 | 10.0±2.8 |
| SiamMAE (Gupta et al., 2023) | 16.8±4.4 | 36.5±7.0 | 68.0±7.9 | 17.3±3.7 | 13.5±4.8 |
| RSP (Jang et al., 2024) | 31.0±2.4 | 44.5±5.6 | <u>82.5</u>±2.7 | <u>28.8</u>±4.8 | 30.3±5.6 |
| CropMAE (Eymaël et al., 2025) | 31.5±5.3 | 54.0±11.2 | 77.0±8.1 | 25.5±5.7 | 32.5±4.1 |
| Ours | **57.3**±2.3 | **82.0**±1.6 | **95.0**±7.1 | **51.0**±1.4 | **55.0**±1.4 |
| Gain | + 25.8 | + 26.2 | + 12.5 | + 26.2 | + 22.5 |

Table 2: **Experimental results on vision-based robot policy learning on RLBench.** We report the performance of imitation learning agents on RLBench (James et al., 2020), which are trained upon representations from the ViT-S/16 model pre-trained on Kinetics-400 (Kay et al., 2017) dataset. The success rates (%) are reported for all the tasks. We report the gains of our method over the second-best baseline.

| Method | Button | Phone | Umbrella | Wine | Rubbish |
|---|---|---|---|---|---|
| SimCLR (Chen et al., 2020a) | 7.4±2.6 | 34.6±6.6 | 5.8±3.3 | 11.0±2.1 | 5.2±1.2 |
| MoCo v3 (Chen et al., 2021) | 11.4±4.1 | 36.2±3.4 | 13.2±1.5 | 8.7±0.7 | 6.7±0.8 |
| DINO (Caron et al., 2021) | 24.7±1.5 | 32.0±5.5 | 28.1±1.4 | 31.4±1.5 | 12.9±1.5 |
| MAE (He et al., 2022) | 6.4±2.2 | 37.7±1.9 | 10.0±1.2 | 10.0±2.1 | 6.2±3.2 |
| SiamMAE (Gupta et al., 2023) | 6.1±2.3 | 5.4±0.5 | 4.0±0.0 | 8.7±0.8 | 3.5±0.9 |
| RSP (Jang et al., 2024) | <u>28.4</u>±3.0 | <u>48.0</u>±4.6 | 37.3±3.0 | 31.9±2.3 | 18.5±1.1 |
| CropMAE (Eymaël et al., 2025) | 26.9±6.7 | 16.6±3.8 | <u>37.5</u>±8.8 | <u>33.2</u>±0.2 | <u>20.6</u>±1.7 |
| Ours | **41.2**±7.4 | **52.3**±3.2 | **42.2**±6.9 | **35.4**±3.8 | **37.0**±6.1 |
| Gain | +12.8 | +4.3 | +4.7 | +2.2 | +18.5 |

success rates for other tasks. We conduct demonstration tasks for five different seeds and report its average with a 95% confidence interval.

**Real-world Environments.** We evaluate our proposed method in real-world robotic imitation learning tasks using a UR5e manipulator equipped with a parallel gripper. The policy operates at a control frequency of 5 Hz, executing actions defined as delta end-effector poses and gripper's state, with specific parameterizations for each task: (dx, dy) for drawer closing, (dx, dy, gripper open/close) for cabinet opening, and (dx, dy, dz, gripper) for cup stacking. The system employs joint position control at 50 Hz, with a numerical inverse kinematics (IK) solver running in the background to calculate the end-effector's pose to the joint position. Our training dataset consists of 50 demonstrations for cabinet opening and cup stacking and 30 demonstrations for drawer closing. We train the two-layer MLP policy for 100 epochs without incorporating proprioceptive states, using a top-front camera view with a resolution of 224×224. The final performance is evaluated based on the reported average success rate across tasks. Figure 3 provides visual examples of the three tasks under consideration.

#### 4.1.3 BASELINES.

We compare the performance of our method with conventional self-supervised learning (SSL) methods for visual representations including SimCLR (Chen et al., 2020a), MoCo v3 (Chen et al., 2021), DINO (Caron et al., 2021), and MAE (He et al., 2022) We also consider previous dynamic scene SSL methods, i.e., SiamMAE (Gupta et al., 2023), RSP (Jang et al., 2024), and CropMAE (Eymaël et al., 2025). We validate the impacts of explicitly learning state representations over these approaches.

Table 3: **Experimental results on vision-based robot policy learning on CortexBench.** The performance of imitation learning agents on CortexBench (Majumdar et al., 2023) is reported, where the agents are trained upon representations from the ViT-S/16 model pre-trained on the Kinetics-400 (Kay et al., 2017) dataset. We report the normalized score for DeepMind Control Suite (DMC) and success rates (%) for other tasks. We report the gains of our method over the second-best baseline.

| Method | Adroit | MetaWorld | DMC | Trifinger |
|---|---|---|---|---|
| SimCLR (Chen et al., 2020a) | 40.4±3.3 | 78.4±5.2 | 39.7±2.9 | 63.3±3.3 |
| MoCo v3 (Chen et al., 2021) | 39.6±4.3 | 65.4±8.0 | 43.7±3.2 | 53.3±1.6 |
| DINO (Caron et al., 2021) | 45.6±6.2 | 82.4±5.8 | 50.9±1.5 | 64.2±3.5 |
| MAE (He et al., 2022) | 39.6±4.3 | 65.4±8.0 | 43.7±3.2 | 53.3±1.6 |
| SiamMAE (Gupta et al., 2023) | 44.0±6.6 | 81.1±6.3 | 56.0±2.9 | 52.1±7.6 |
| RSP (Jang et al., 2024) | 45.6±4.6 | <u>84.5</u>±6.6 | <u>61.6</u>±3.4 | <u>66.2</u>±0.8 |
| CropMAE (Eymaël et al., 2025) | <u>50.0</u>±5.1 | 82.4±5.8 | 46.4±1.1 | 46.3±1.7 |
| Ours | **60.4**±2.2 | **87.8**±4.6 | **73.5**±8.6 | **66.5**±1.0 |
| Gain | +10.4 | +3.3 | +11.9 | +0.3 |

Table 4: **Experimental results on vision-based robot policy learning in real-world environments.** The performance of imitation learning agents on three demonstration tasks ('Cabinet Opening', 'Drawer Closing', and 'Cup Stacking') in real-world environments is reported. The agents are trained upon representations from the ViT-S/16 model pre-trained on the Kinetics-400 (Kay et al., 2017) dataset. We report the success rates (%) for the tasks.

| Method | Cabinet Opening | Drawer Closing | Cup Stacking |
|---|---|---|---|
| SiamMAE (Gupta et al., 2023) | 20.0 | 55.0 | 50.0 |
| RSP (Jang et al., 2024) | 25.0 | 65.0 | 55.0 |
| CropMAE (Eymaël et al., 2025) | 0.0 | 25.0 | 20.0 |
| SE-MAE (ours) | **65.0** | **75.0** | **80.0** |

### 4.2 Vision-based robot policy learning in simulated environments

We evaluated our method through imitation learning on robot manipulation and locomotion tasks across various simulated environments, including Franka Kitchen (Gupta et al., 2019), RLBench (James et al., 2020), and CortexBench (Majumdar et al., 2023).

**Franka Kitchen.** We present a comparison between our method and the baselines on vision-based robot policy learning in the Franka Kitchen environment in Table 1. The results demonstrate that our method significantly outperforms all the baselines across all tasks. Notably, our method achieves over 20% improvements in success rates on all tasks, except for the "Light on" task. This highlights the effectiveness of explicitly encoding visual state representation for vision-based robot policy learning.

**RLBench** Table 2 showcases the robot manipulation performance on five demonstration tasks in the RLBench environment. Notably, our method consistently exceeds all baselines across the five tasks. Morevoer, the degraded performance of MAE and SiamMAE further highlights the significance of state representation learning for the robot backbones.

**CortexBench.** We compare our method with the baselines for the vision-based robot manipulation and locomotion tasks in the Adroit, MetaWorld, DeepMind Control (DMC), and Trifinger environments in Table 3. The results show that our method achieves superior performance compared to the baselines across all tasks. In particular, our method surpasses the second-best performance with success rate gains of 11.9%p on DMC and 10.4%p on Adroit.

### 4.3 Vision-based Robot Policy Learning in Real-world Environments

To validate the robustness of our method in real-world environments, we further investigate SSL methods on real-world robotics robot tasks. Specifically, we design three demonstration tasks: *Door*

*Opening*, *Door Closing*, and *Cup Stacking*. For each task, We collect 50 demonstration episodes for training and 20 demonstration episodes for evaluation for imitation learning. Following the training protocol used in simulated environments, we train the policy network using a standard behavior cloning loss. The experimental results for each individual task are reported in Table 4. We first observe that our method exceeds SiamMAE (Gupta et al., 2023), RSP (Jang et al., 2024), and CropMAE (Eymaël et al., 2025) on all three tasks. Specifically, our method improves 40%p, 10%p, and 25%p over the baselines on the *Door Opening*, *Door Closing*, and *Cup Stacking* tasks, respectively. While previous SSL methods on dynamic scenes struggle with the Door Opening task, our method even successfully executes the task with a considerable success rate. This showcases that models pre-trained by our method can be robustly transferred to real-world environments.

## 5 Conclusion

We have introduced a state representation learning pipeline for robot backbones. Since deriving state representations from observed scenes is crucial for accurate robot action prediction, establishing a strong robot backbone is essential. Such backbone models should effectively encode both fine-grained and high-level semantic information from observations while facilitating the recognition of temporal progression across sequential scenes. However, though conventional self-supervised methods have provided promising recipes for visual representation learning, they have primarily focused on achieving a holistic understanding of static images or videos. Recent work on SSL has addressed this limitation by exploiting correspondence learning on dynamic scenes. However, the patch-wise representations of observations are not structured for the subsequent policy network, resulting in suboptimal performance for robot backbones. For an enhanced expression of the observations, we have introduced a simple and intuitive pipeline that explicitly learns state representation derivation during the encoding process. For a more effective derivation of representations, we have proposed a straightforward and intuitive pipeline that explicitly learns representation derivation during the encoding process. Specifically, we guide the masked autoencoding process to depend heavily on the state representation of the reference scene by applying extreme masking to the target scene. Our extensive experiments in robot policy learning on various simulated environments verified its superiority over conventional SSL methods and previous dynamic scene SSL methods. Extension to the real-world environment validated its robustness.

## References

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 Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/baevski22a.html.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of NeurIPS*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 2020a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pp. 15750–15758, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Alexandre Eymaël, Renaud Vandeghen, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Efficient image pre-training with siamese cropped masked autoencoders. In *European Conference on Computer Vision*, pp. 348–366. Springer, 2025.

Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. 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.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin (eds.), *NeurIPS*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020.

Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019.

Agrim Gupta, Jiajun Wu, Jia Deng, and Li Fei-Fei. Siamese masked autoencoders. In *Advances in Neural Information Processing Systems*, 2023.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

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*, 2022.

Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Huiwon Jang, Dongyoung Kim, Junsu Kim, Jinwoo Shin, Pieter Abbeel, and Younggyo Seo. Visual representation learning with stochastic frame prediction. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21289–21305. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/jang24c.html.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Taekyung Kim, Sanghyuk Chun, Byeongho Heo, and Dongyoon Han. Learning with unmasked tokens drives stronger vision learners. *European Conference on Computer Vision (ECCV)*, 2024.

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

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? In *Advances in neural information processing systems*, 2023.

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.

Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. 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.

Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17359–17371. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/parisi22a.html.

Yuval Tassa, Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, and Nicolas Heess. dm_control: Software and tasks for continuous control. *arXiv preprint arXiv:2006.12983*, 2020.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in neural information processing systems*, 2022.

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Revaud Jérôme. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. In *NeurIPS*, 2022.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *ICCV*, 2022.

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*, 2020.