SPA: 3D SPATIAL-AWARENESS ENABLES EFFECTIVE EMBODIED REPRESENTATION

Haoyi Zhu[1,2], **Honghui Yang**[2,3], **Yating Wang**[2,4], **Jiange Yang**[2,5], **Limin Wang**[2,5], **Tong He**[2†]
[1]USTC, [2]Shanghai AI Lab, [3]ZJU, [4]Tongji, [5]NJU
[†] Corresponding Author

ABSTRACT

In this paper, we introduce SPA, a novel representation learning framework that emphasizes the importance of 3D spatial awareness in embodied AI. Our approach leverages differentiable neural rendering on multi-view images to endow a vanilla Vision Transformer (ViT) with intrinsic spatial understanding. We present the most comprehensive evaluation of embodied representation learning to date, covering 268 tasks across 8 simulators with diverse policies in both single-task and language-conditioned multi-task scenarios. The results are compelling: SPA consistently outperforms more than 10 state-of-the-art representation methods, including those specifically designed for embodied AI, vision-centric tasks, and multi-modal applications, while using less training data. Furthermore, we conduct a series of real-world experiments to confirm its effectiveness in practical scenarios. These results highlight the critical role of 3D spatial awareness for embodied representation learning. Our strongest model takes more than 6000 GPU hours to train and we are committed to open-sourcing all code and model weights to foster future research in embodied representation learning. Project Page: https://haoyizhu.github.io/spa/.

1 INTRODUCTION

Vision systems have made remarkable progress in understanding 2D images (He et al., 2020; Chen et al., 2020a; He et al., 2022; Feichtenhofer et al., 2022; Tong et al., 2022; Yang et al., 2023; Oquab et al., 2023; Radford et al., 2021; Fang et al., 2023b; Chen et al., 2024b). However, achieving true visual intelligence necessitates a comprehensive understanding of the 3D world. This is crucial for embodied AI, where agents must perceive, reason, and interact with complex 3D environments.

Existing visual representation learning methods for embodied AI (Nair et al., 2022; Radosavovic et al., 2023; Majumdar et al., 2023; Karamcheti et al., 2023; Shang et al., 2024; Yang et al., 2024b) largely rely on paradigms from 2D vision, predominantly employing contrastive-based or masked autoencoder (MAE)-based approaches. However, they often struggle to fully capture the spatial relationships and 3D structures inherent in the physical world. This limitation arises from their primary emphasis on 2D semantic understanding, which, though valuable, is still insufficient for the sophisticated spatial reasoning required in embodied AI tasks, where agents need to navigate environments, manipulate objects, and make decisions using their 3D spatial awareness.

In this paper, we introduce **SPA**, a general 3D spatial-aware representation learning framework for embodied AI. SPA leverages neural rendering (Mildenhall et al., 2021) as the pre-training pre-text task on multi-view images. Unlike explicit 3D representations like point clouds or meshes—which prior work (Wang et al., 2024b;a; Ze et al., 2024; Zhu et al., 2024) has shown to outperform pure 2D inputs in robot learning—multi-view images are easier to process and more readily available, making them ideal for large-scale training, such as from internet videos. Specifically, given a vanilla 2D image backbone, *e.g.* a Vision Transformer (ViT) (Dosovitskiy et al., 2021), we first extract multi-view feature maps from the input images. Using known camera poses, we then construct a feature volume from these feature maps and sample rays to apply differentiable neural rendering. This process generates multi-view RGB-D images and semantic maps for supervision without labels, enabling the pre-training of a 2D image backbone to enhance 3D spatial awareness.

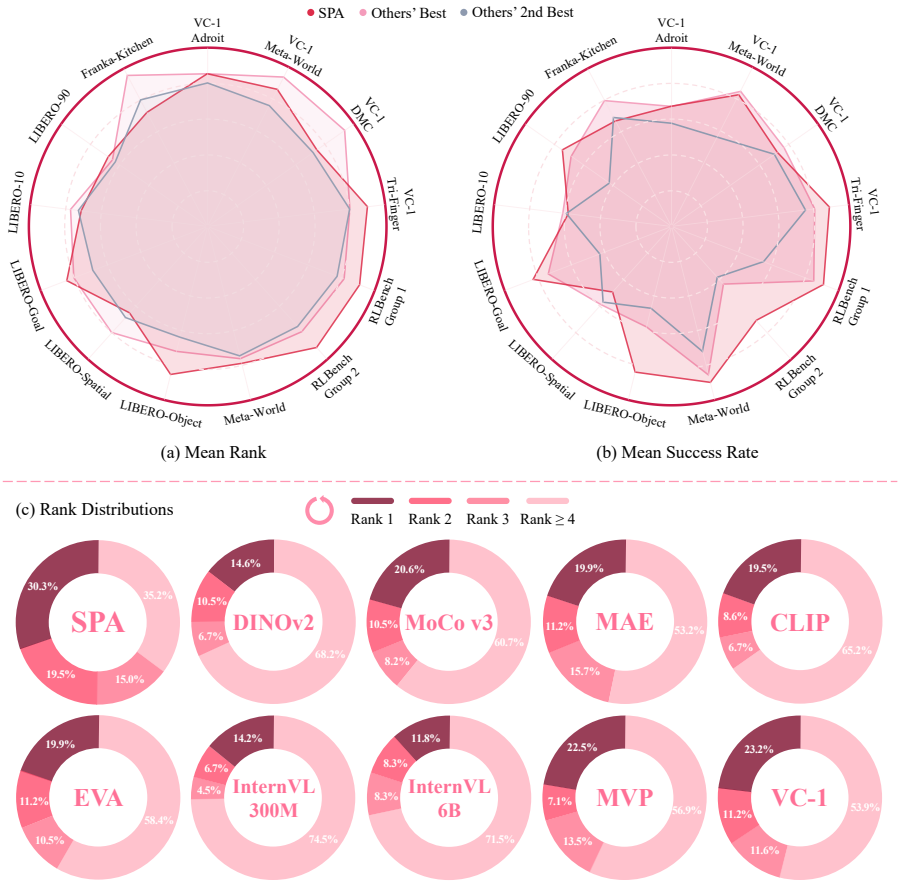

Figure 1: **Performance comparison across representations.** *Above*: (a) Mean rank and (b) mean success rate on benchmarks. Lines represent the performance of SPA, best, and second best performance on each benchmark. *Bottom*: Rank distributions for 268 individual tasks, showing proportions from rank 1 to rank ≥ 4 counterclockwise. Our model demonstrates superior overall performance.

To thoroughly validate our assumption and method, we collect 268 embodied tasks across 8 simulators using various policy methods. To our knowledge, this represents **the largest scale of embodied evaluation to date**. Previous work, such as R3M (Nair et al., 2022) and VC-1 (Majumdar et al., 2023), evaluated fewer than 20 tasks, potentially leading to incomplete or biased conclusions. Our evaluation spans both single-task and language-conditioned multi-task learning. We compare over 10 state-of-the-art representation learning methods, categorized as embodied-specific (Nair et al., 2022; Majumdar et al., 2023; Radosavovic et al., 2023), vision-centric (Oquab et al., 2023; Chen et al., 2021; He et al., 2022), and multi-modal (Radford et al., 2021; Fang et al., 2023b; Chen et al., 2024b). Our method consistently outperforms others, underscoring the importance of 3D spatial awareness for embodied AI. Notably, multi-modal models like CLIP (Radford et al., 2021), consistently perform poorly. This holds even the vision-language model scales the ViT to 6B parameters (Chen et al., 2024b). Through a camera pose estimation task and feature map visualization, we demonstrate that SPA has learned superior 3D spatial understanding. Further, we find that 3D awareness shows a positive correlation with embodied performance. Finally, we conduct several real-world tasks, where SPA also demonstrates superior performance. Our contribution can be summarized as follows.

- We propose a significant *spatial hypothesis*: 3D spatial awareness is crucial for embodied representation learning. Our experiments provide clear evidence for the hypothesis.
- We introduce SPA, a novel paradigm for representation learning in embodied AI. It enhances a vanilla Vision Transformer (ViT) with 3D awareness using differentiable neural rendering as the pre-text task on multi-view images.
- We conduct the largest evaluation benchmark for embodied representation learning, significantly larger than previous studies. It involves 268 tasks, 8 simulators, and over 10 SOTA methods with diverse downstream policies and task settings.

- Through extensive experiments in both simulators and real-world settings, SPA outperforms more than 10 SOTA representation learning methods, demonstrating its effectiveness.

2 METHODOLOGY

In this section, we first describe our process for handling multi-view image inputs and feature extraction in Sec. 2.1. Subsequently, we construct an explicit feature volume from these multi-view features, detailed in Sec. 2.2. Finally, we explain the image rendering from the feature volume and loss functions for network optimization in Sec. 2.3 and Sec. 2.4. Our pipeline is visualized in Fig. 2.

2.1 INPUT PROCESS AND FEATURE EXTRACTION

Given a set of multi-view images $\mathbf{I} = \{I_1, I_2, \ldots, I_N\}$, where each $I_i \in \mathbb{R}^{3\times H\times W}$ and $N \in \mathbb{Z}^+$, we utilize a 2D image backbone F, such as a ViT. The images are processed separately through F, yielding latent features $\mathbf{L} = \{l_1, l_2, \ldots, l_N\}$, where each $l_i = F(I_i) \in \mathbb{R}^{L\times C}$. Following MAE, we apply random masking to input images to enhance robustness, but without a ViT decoder and MAE's pixel reconstruction objective. For each l_i, masked positions are filled with a mask token, and we concatenate the global class token with other patch tokens as read-out tokens similar to DPT (Ranftl et al., 2020). We then unpatchify them to obtain a latent feature map of size $\frac{H}{P} \times \frac{W}{P}$, where P is the ViT patch size. Finally, two simple upsampling layers transform this into a feature map M_i matching the input resolution. Each upsampling layer includes a convolution, a GELU (Hendrycks & Gimpel, 2016) activation, and a pixel shuffle layer (Shi et al., 2016) with an upscale factor of $\sqrt{P}$.

2.2 DYNAMIC VOLUME CONSTRUCTION

To enable multi-view interaction, we construct a 3D feature volume from multi-view feature maps, $\mathbf{M}$. Unlike the bird's-eye view (BEV) construction in autonomous driving (Li et al., 2022), which usually relies on a fixed scene range around ego vehicle , our method dynamically adjusts the scene range based on the spatial extents of the environment to accommodate varying datasets. Specifically, the scene's bounds are first estimated using available depth data, sparse points, or pre-defined rules. We then partition the scene into a volume of size $X \times Y \times Z$, with voxel size dynamically adjusted to capture either fine object details or larger environments. Voxel features, $\tilde{\mathcal{V}}$, are initialized with learnable positional embeddings. Each voxel is projected onto the multi-view feature maps using the known transformation matrix $\mathbf{T}$. Deformable attention (Zhu et al., 2021) is then applied, where the multi-view features act as keys and values, and the voxel features as queries. Finally, a 3D convolution refines the output volume features to obtain $\mathcal{V}$. The process can be formulated as:

$$\mathcal{V} = \text{Conv3D}(\text{DeformAttn}(\tilde{\mathcal{V}}, \mathbf{M}, \mathbf{T})). \tag{1}$$

2.3 DIFFERENTIABLE VOLUMETRIC RENDERING

After constructing the feature volume, we employ differentiable neural rendering (Mildenhall et al., 2021) to connect 2D and 3D domains. For better geometry representation, we utilize the implicit signed distance function (SDF) field modeling as in NeuS (Wang et al., 2021). The SDF represents the 3D distance from a query point to the nearest surface, implicitly capturing the 3D geometry.

Given a feature volume $\mathcal{V}$, we apply a shallow 3D CNN ϕ to directly produce three outputs: an SDF feature volume $\mathcal{S} \in \mathbb{R}^{X\times Y\times Z}$, a spherical harmonic (SH) (Yu et al., 2021; Zhu et al., 2023a) coefficient field $\mathcal{K} \in \mathbb{R}^{D\times X\times Y\times Z}$ (where $D = 3 \cdot (l_{\max}+1)^2$) for color rendering, and a semantic feature volume $\mathcal{F} \in \mathbb{R}^{C_{\text{semantic}}\times X\times Y\times Z}$:

$$\mathcal{S} \in \mathbb{R}^{X\times Y\times Z}, \quad \mathcal{K} \in \mathbb{R}^{D\times X\times Y\times Z}, \quad \mathcal{F} \in \mathbb{R}^{C_{\text{semantic}}\times X\times Y\times Z} = \phi(\mathcal{V}). \tag{2}$$

Unlike prior work (Huang et al., 2023; Zhu et al., 2023b; Yang et al., 2024a), which employs an MLP to compute the attributes of each sampled point individually, we directly apply a 3D CNN to $\mathcal{V}$. This eliminates the need for pointwise MLP computations, reducing redundant processing and enabling more efficient execution. Consequently, our approach leads to substantial improvements in both time and memory efficiency, especially when sampling a large number of points during rendering.

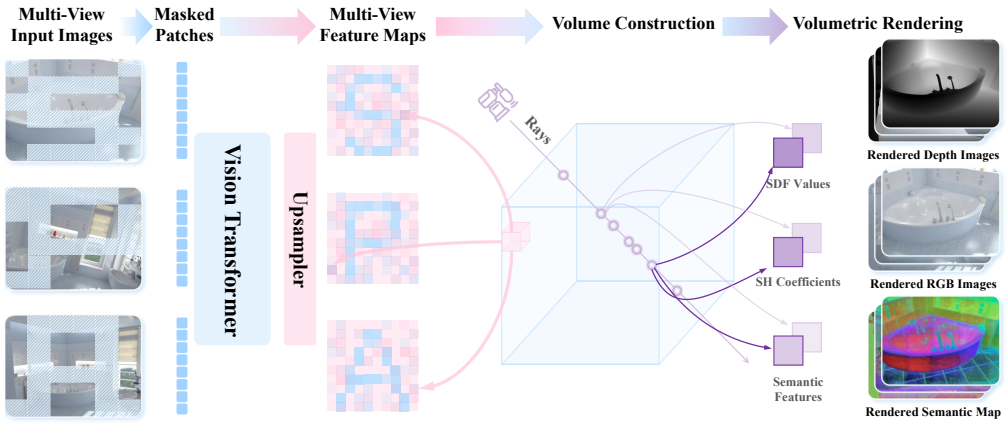

Figure 2: **Pipeline Overview.** Given multi-view images, we randomly mask patches and input the remaining into a Vision Transformer. The upsampled latent features generate multi-view feature maps, from which we construct a feature volume to derive SDF values, SH coefficients, and semantic features. We then render depth, RGB, and semantic maps for loss computation.

To render a 2D pixel i, we sample N ray points $\{\mathbf{p}_j = \mathbf{o} + t_j \mathbf{d}_i \mid j = 1, \ldots, N, \, t_j < t_{j+1}\}$ from ray $\mathbf{r}_i$, where $\mathbf{o}$ is the camera origin and $\mathbf{d}_i$ is the viewing direction. Attributes for each point are obtained via trilinear sampling:

$$s_j = \tau(\mathcal{S}, \mathbf{p}_j), \quad \mathbf{k}_j = \tau(\mathcal{K}, \mathbf{p}_j), \quad \mathbf{f}_j = \tau(\mathcal{F}, \mathbf{p}_j). \tag{3}$$

The SH vector $\mathbf{k}_j = (k_l^m)_{0 \le l \le l_{\text{max}}, -l \le m \le l}$, where $k_l^m \in \mathbb{R}^3$, is used to compute view-dependent colors $\hat{\mathbf{c}}_j$ by querying the SH basis functions $Y_l^m : \mathbb{S}^2 \to \mathbb{R}$ based on the viewing direction $\mathbf{d}_j$:

$$\hat{\mathbf{c}}_j = \text{Sigmoid}\left(\sum_{l=0}^{l_{\text{max}}} \sum_{m=-l}^{l} k_l^m Y_l^m(\mathbf{d}_j) \right). \tag{4}$$

Following the formulation in NeuS (Wang et al., 2021), the RGB color $\hat{\mathbf{C}}_i$, depth $\hat{\mathbf{D}}_i$, and semantic feature $\hat{\mathbf{F}}_i$ for pixel i are computed by integrating the predicted values along the ray:

$$\hat{\mathbf{C}}_i = \sum_{j=1}^{N} w_j \hat{\mathbf{c}}_j, \quad \hat{\mathbf{D}}_i = \sum_{j=1}^{N} w_j t_j, \quad \hat{\mathbf{F}}_i = \sum_{j=1}^{N} w_j \hat{\mathbf{f}}_j, \tag{5}$$

where $w_j = T_j \alpha_j$ is the occlusion-aware weight, with $T_j = \prod_{k=1}^{j-1}(1 - \alpha_k)$ representing the accumulated transmittance and α_j being the opacity value. Specifically, α_j is computed as:

$$\alpha_j = \max\left(\frac{\sigma_s(s_j) - \sigma_s(s_{j+1})}{\sigma_s(s_j)}, 0 \right), \tag{6}$$

where $\sigma_s(x) = (1 + e^{-sx})^{-1}$ is the sigmoid function modulated by a learnable parameter s.

2.4 Loss Functions

During pre-training, we randomly sample K pixels from multi-view inputs in each iteration. The rendering loss is calculated based on the differences between the input pixel values and the predicted values. For the semantic feature map, we use the feature map from AM-RADIO (Ranzinger et al., 2024) as supervision. Our framework has the capability to distill knowledge from multiple vision foundation models by adding multiple rendering heads. However, this paper does not explore that approach, as it is not the primary focus. The rendering loss is expressed as:

$$\mathcal{L}_{\text{render}} = \frac{1}{K} \sum_{i=1}^{K} \Big(\lambda_{\text{color}} \cdot \|\mathbf{C}_i - \hat{\mathbf{C}}_i\| + \lambda_{\text{depth}} \cdot \|\mathbf{D}_i - \hat{\mathbf{D}}_i\| + \lambda_{\text{semantic}} \cdot \|\mathbf{F}_i - \hat{\mathbf{F}}_i\| \Big). \tag{7}$$

Additionally, we incorporate the Eikonal regularization loss $\mathcal{L}_{\text{eikonal}}$, near-surface SDF supervision loss $\mathcal{L}_{\text{sdf}}$, and free space SDF loss $\mathcal{L}_{\text{free}}$, which are standard in neural surface reconstruction. Detailed definitions of these losses are provided in Appendix A. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{render}} + \lambda_{\text{eikonal}} \cdot \mathcal{L}_{\text{eikonal}} + \lambda_{\text{sdf}} \cdot \mathcal{L}_{\text{sdf}} + \lambda_{\text{free}} \cdot \mathcal{L}_{\text{free}}. \tag{8}$$

Figure 3: **Overview of our large-scale embodied evaluation.** We conduct the largest-scale evaluation of embodied representation learning to date. Our study encompasses 268 tasks across 8 simulators, including both single-task and language-conditioned multi-task settings. We evaluate diverse policy architectures and assess various state-of-the-art representation methods. This thorough evaluation allows us to provide a comprehensive and unbiased analysis of different representations.

3 Large-Scale Embodied Evaluation

Unlike the CV or NLP communities, where large-scale benchmarks are common, embodied representations have not been thoroughly assessed. The largest previous evaluation, VC-1 (Majumdar et al., 2023), includes only 17 tasks. This may lead to randomness and bias. Therefore, we have created **the largest embodied evaluation to date**, encompassing **268 tasks** across 8 simulators—**over 15 times larger** than VC-1's evaluation. Additionally, unlike previous approaches (Majumdar et al., 2023; Nair et al., 2022; Radosavovic et al., 2023) that used a small MLP policy under single-task settings, our evaluation spans multiple policy types (*e.g.* MLP, diffusion, transformer) and includes both single-task and language-conditioned multi-task settings. This unprecedented scale and diversity ensure robust and convincing conclusions. During all evaluations, we adhere to standard practices by freezing the pre-trained representation model. Our detailed evaluation settings can be found in Appendix B. The overview of our evaluation is shown in Fig. 3.

We have included 3 *single-task benchmarks*:
1) VC-1 (Majumdar et al., 2023) involves 4 selected simulators with 14 tasks in total: Adroit (AD) (Kumar, 2016), Meta-World (MW) (Yu et al., 2020), DMControl (DMC) (Tunyasuvunakool et al., 2020), and TriFinger (TF) (Wüthrich et al., 2020). We use a 3-layer MLP as the policy network.
2) Franka Kitchen (Gupta et al., 2019) involves 5 selected tasks. Each task spans two camera viewpoints and three random seeds. We utilize 25 demonstrations to train a 2-layer MLP policy.
3) Meta-World (Yu et al., 2020) involves 48 selected tasks of varying difficulty. We implemented the Diffusion Policy (Chi et al., 2023) on this benchmark and adhered to the setup in Ze et al. (2024) to generate 10 demonstrations for each single-task training, followed by evaluation through 20 rollouts.

We have also included 2 *language-conditioned multi-task benchmarks*:
1) RLBench (James et al., 2020) features 71 selected tasks that can be successfully executed. We divide the tasks into two groups according to their category defined by PolarNet (Chen et al., 2023). We employ RVT-2 (Goyal et al., 2024), the SOTA method on this benchmark, as our policy.
2) LIBERO (Liu et al., 2024) comprises 130 tasks across 5 suites: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, LIBERO-10, and LIBERO-90. We train a language-conditioned transformer policy provided by the original LIBERO on each suite with only 20 demonstrations per task.

4 Training and Implementation Details

In this section, we present the implementation and training of our SPA model. We first compile several multi-view datasets, training ViT-B models on each to assess the impact of different datasets (Sec. 4.1). Finally, we integrate all factors and scale up both data and model size to train the strongest version of SPA using a ViT-large (ViT-L) backbone (Sec. 4.2). More details can be found in Appendix C.

4.1 DATASET INVESTIGATION

We collect several multi-view datasets. To investigate their effectiveness in SPA representation learning, we train a ViT-B model on one or two of the datasets, keeping the total training steps constant, and assess performance on the VC-1 benchmarks. For simplicity, semantic rendering is disabled. The datasets investigated are listed in the first column of Tab. 1. Most datasets provide ground-truth depth, which we use for supervision. As our findings above reveal that depth supervision is helpful, for datasets lacking ground-truth depth, we employ a depth estimation model. For instance, Droid (Khazatsky et al., 2024) only offers binocular images, so we apply CroCo-Stereo (Weinzaepfel et al., 2023) for dense depth estimation. Additionally, due to inaccurate camera poses in Droid, we treat its data as single-view inputs. The results are presented in Tab. 1, with further details in Appendix C. Our analysis reveals that some datasets can be detrimental. For example, although RH20T (Fang et al., 2023a) is a large-scale robotic dataset, its lack of visual diversity—stemming from data collected in the same lab—negatively impacts representation learning.

Table 1: **Influence of different datasets.** We present the performance results on the VC-1 benchmark. *Mean S.R.* refers to the mean success rate across all individual tasks.

Datasets	AD	MW	DMC	TF	Mean S.R.
ScanNet (Dai et al., 2017)	52.67±4.11	90.93±3.22	65.11±1.31	70.75±1.08	73.68
ScanNet++ (Yeshwanth et al., 2023)	56.00±2.83	89.87±4.20	62.24±4.51	71.28±0.38	72.51
Arkitscenes (Baruch et al., 2021)	50.67±5.73	89.87±4.59	60.51±2.55	66.54±0.13	70.45
Droid (Khazatsky et al., 2024)	53.33±5.25	90.40±4.90	60.99±3.72	73.28±0.61	72.16
Hypersim (Roberts et al., 2021)	52.67±4.11	88.80±3.27	60.84±2.06	72.29±0.47	71.29
Hypersim + ADT (Pan et al., 2023)	52.00±2.83	87.20±2.30	63.61±1.04	70.83±0.13	71.41
Hypersim + S3DIS (Armeni et al., 2017)	49.33±0.94	94.13±2.04	64.57±3.91	71.74±0.75	73.98
Hypersim + Structured3D (Zheng et al., 2020)	46.67±4.11	80.27±7.72	58.02± 2.34	65.05±0.40	65.35
Hypersim + RH20T (Fang et al., 2023a)	47.33±1.89	86.93±4.99	57.01±4.35	64.28±0.46	67.35
Hypersim + ASE (Avetisyan et al., 2024)	47.33±4.11	87.73±3.39	60.62±4.14	68.59±0.30	69.54

4.2 PUT ALL TOGETHER

Based on the previous analyses, we proceed to pre-train the final version of SPA. We use a mask ratio of 0.5 and enable all three rendering losses. Following Ponder (Huang et al., 2023), we set the weight for the RGB loss to 10, the weights for the depth and semantic losses to 1, and use $\lambda_{\text{eikonal}} = 0.01$, $\lambda_{\text{sdf}} = 10$, and $\lambda_{\text{free}} = 1$. The volume size is $128 \times 128 \times 32$. For stable training, we apply the Exponential Moving Average (EMA) technique with a decay of 0.999. We use AdamW (Loshchilov et al., 2017) as the optimizer with a weight decay of 0.04 and a learning rate of $8e^{-4}$. OneCycle (Smith & Topin, 2019) learning rate scheduler is adopted. We utilize 80 NVIDIA A100-SXM4-80GB GPUs, each with a batch size of 2, and accumulate gradients over 8 batches, resulting in a total effective batch size of $2 \times 8 \times 80 = 1280$. Training is conducted over 2000 epochs, sampling each dataset to match the size of ADT per epoch. The datasets used for the final version include ScanNet, ScanNet++, ADT, S3DIS, Hypersim, and Droid.

5 EXPERIMENT RESULTS

In this section, we present the results of our large-scale evaluation. Our experiments are designed to address the following research questions:
Q1: How does SPA compare to other methods in our large-scale embodied evaluation?
Q2: What insights do we gain about various representation learning approaches from our evaluation?
Q3: Does SPA really learn enhanced 3D awareness that results in improved embodied representation?
Q4: Can SPA facilitate robot learning in real-world environments in a zero-shot manner?

5.1 OVERALL COMPARISONS (Q1, Q2)

Evaluation Metrics. We follow prior work (Majumdar et al., 2023; Zhu et al., 2024) in reporting two metrics: *Mean Success Rate (Mean S.R.)* and *Mean Rank*. Mean S.R. is the average success rate across all tasks, indicating overall performance, while Mean Rank reflects the average ranking of each method's success rate across tasks, providing a measure of relative performance. Since RLBench has fixed train and test sets, we report a single result for this benchmark.

Table 2: **Summary of different representation learning methods.** '#Param.' is the total parameters of the encoder, while '#Frames' indicates the total number of image frames used during pre-training.

Method	*Vision-Centric* MoCoV3 (Chen et al., 2020b)	MAE (He et al., 2022)	DINOV2 (Oquab et al., 2023)	*Multi-Modal* CLIP (Radford et al., 2021)	EVA (Fang et al., 2023b)	InternViT-300M (Chen et al., 2024b)	InternViT-6B (Chen et al., 2024b)	*Embodied-Specific* MVP (Radosavovic et al., 2023)	VC-1 (Majumdar et al., 2023)	SPA (Ours)	*Distilled* AM-RADIO (Ranzinger et al., 2024)
Is Vanilla?	✓	✓	✗	✓	✓	✗	✗	✓	✓	✓	✗
Input Size	224	224	224	224	224	448	224	256	224	224	dynamic
Patch Size	16	16	14	14	14	14	14	16	16	16	16
#Param.	303M	303M	303M	303M	303M	303M	5.9B	303M	303M	303M	653M
#Frames	1.28M	1.28M	1.2B	400M	14M	5.0B	5.0B	4.5M	5.6M	3.8M	1.4B

Table 3: **Comparison of different representation learning methods.** 'OOM' indicates an out-of-memory error during evaluation. The best and second-best results are **bolded** and underlined respectively. The number in parentheses denotes the number of tasks. S.R. denotes 'Success Rate'.

Benchmark	Method	*Vision-Centric* MoCoV3	MAE	DINOV2	*Multi-Modal* CLIP	EVA	InternViT-300M	InternViT-6B	*Embodied-Specific* MVP	VC-1	SPA (Ours)
VC-1	AD (2)	58.7±7.0	58.0±2.0	47.3±3.1	48.7±3.1	58.0±6.0	53.3±3.1	60.0±9.2	53.3±4.2	54.0±4.0	**60.0±4.0**
	MW (5)	88.8±5.0	90.0±4.6	84.0±3.7	77.1±3.2	90.7±0.9	84.0±3.7	89.1±1.2	**93.6±5.2**	87.5±3.8	93.3±2.0
	DMC (5)	67.3±3.3	**74.4±1.8**	64.5±2.5	53.9±3.6	62.7±2.8	53.3±0.4	66.3±3.2	69.4±2.6	65.3±3.6	71.1±5.0
	TF (2)	67.9±0.2	73.0±0.5	68.5±0.4	56.1±1.6	67.2±0.2	65.2±1.6	70.7±0.9	73.2±0.8	70.9±1.1	**73.6±2.0**
RLBench	Group 1 (35)	73.7	78.3	78.2	76.8	75.2	74.1	OOM	76.2	80.1	**80.5**
	Group 2 (36)	54.2	57.7	56.1	55.7	57.0	54.9	OOM	56.3	55.7	**61.2**
Meta-World (48)		**69.3±1.5**	67.8±1.7	56.3±0.6	66.7±1.7	63.7±1.3	57.5±1.7	OOM	66.4±1.7	68.6±1.5	69.2±1.7
LIBERO	Object (10)	65.3±8.0	71.7±13.1	64.7±9.9	50.2±7.0	73.2±6.0	67.7±6.0	58.0±10.6	63.7±4.8	69.7±7.2	**76.7±5.3**
	Spatial (10)	40.5±0.9	57.2±2.9	36.3±11.8	32.2±0.6	**59.3±7.7**	48.3±6.4	42.0±10.3	58.0±6.2	50.5±7.5	50.0±3.8
	Goal (10)	49.2±8.1	54.3±6.0	22.2±2.3	30.3±3.2	56.8±2.9	58.8±4.5	33.2±2.0	63.8±2.8	57.5±6.6	**65.3±2.5**
	10 (10)	34.2±3.8	**41.2±4.5**	28.3±3.0	27.5±3.9	43.3±2.8	38.2±1.3	34.3±4.6	39.0±0.9	39.7±3.5	40.2±3.6
	90 (90)	30.0±1.4	29.9±2.0	27.5±2.2	29.4±2.0	31.3±2.3	23.8±1.8	27.1±2.1	32.1±3.5	30.6±3.3	**32.2±1.6**
Franka-Kitchen (5)		**48.3±4.7**	42.7±2.6	40.9±6.4	30.8±3.3	37.3±1.3	28.5±1.7	OOM	34.3±6.1	37.5±3.5	40.6±1.9
Mean S.R. ↑		81.67	85.13	75.18	77.10	83.84	75.41	30.65	84.85	84.69	**88.63**
Mean Rank ↓		4.51	4.07	5.61	5.17	4.37	5.92	7.57	4.24	4.13	**3.20**

Baselines. We evaluate 9 SOTA representation learning models, all using ViT-L backbone, categorized into vision-centric, multi-modal, and embodied-specific. This also includes a 6B multi-modal model (Chen et al., 2024b). The vision-centric methods are originally from the vision community; the multi-modal methods are typically CLIP-style language-image pre-trained models and are used specifically for VLMs; the embodied-specific methods are designed and pre-trained specifically for embodied AI tasks. Details are summarized in Tab. 2. The results on each benchmark are shown in Tab. 3. For detailed results on each task and each random seed, please refer to Appendix D. We also have visualized the performance radar chart and the per-task rank distributions in Fig. 1.

Finding 1: We observe that SPA demonstrates superior performance in both mean success rate and mean rank. While no method ranks first across all individual benchmarks, consistent with the findings by Majumdar et al. (2023), SPA achieves the best or second-best mean success rate **in 11 out of 13 benchmarks**. Additionally, it ranks in the top 3 for **over 65.5% of individual tasks**, surpassing the second and third highest percentages of 46.8% for MAE and 46.0% for VC-1, respectively. These trends demonstrate the robustness and superiority of SPA.

Finding 2: We observe that for vision-centric methods, superior performance on vision tasks does **not** necessarily translate to better embodied performance. Despite using 10 times more data, DINOV2 performs worse than MoCoV3 and MAE. Notably, MAE performs exceptionally well, likely due to its reconstruction objective, which enhances *2D spatial awareness*. Interestingly, methods like MVP and VC-1, which are MAE models pre-trained on human interaction data, show **no clear advantage** over ImageNet (Deng et al., 2009) pre-trained MAE. This suggests that while human activity data may seem more relevant, data diversity and thorough convergence are more critical.

Finding 3: Multimodal methods **generally perform poorly** in embodied evaluations, except EVA, which combines image-language contrastive techniques with MAE reconstruction. Furthermore, InternViT-6B, despite having significantly more model parameters, does not demonstrate superiority and even performs worse on some benchmarks compared to InternViT-300M. This indicates that current scaling properties of multimodal approaches do not effectively translate to embodied AI.

Finding 4: Focusing on a single benchmark can **lead to highly biased conclusions**. For instance, ImageNet pre-trained methods (*e.g.* MoCoV3 and MAE) perform exceptionally well on the Franka Kitchen benchmark, suggesting a minimal domain gap between ImageNet and Franka Kitchen observations. Moreover, despite being based on MAE, previous SOTA embodied representations like MVP and VC-1 do not consistently outperform the original ImageNet version. These observations underscore the importance of our large-scale embodied evaluation.

Table 4: **Additional comparisons of ViT-base models.** S.R. denotes 'Success Rate'.

Methods		DINOV2-B (Oquab et al., 2023)	MAE-B (He et al., 2022)	R3M-B (Nair et al., 2022)	VC-1-B (Majumdar et al., 2023)	STP-B (Yang et al., 2024b)	Voltron-B (Karamcheti et al., 2023)	Theia-B (Shang et al., 2024)	SPA-B (Ours)
Is Vanilla?		✗	✓	✓	✓	✓	✗	✓	✓
Embodied?		✗	✗	✓	✓	✓	✓	✓	✓
VC-1	AD	36.67±2.31	52.67±3.06	48.00±6.93	50.00±5.29	52.00±2.00	46.67±4.62	53.33±5.03	52.00±3.46
	MW	60.80±0.80	88.80±4.00	59.20±5.60	86.67±0.92	92.00±1.39	84.00±3.20	89.07±3.23	92.00±4.16
	DMC	35.19±4.87	62.39±4.97	49.57±4.85	60.92±0.70	61.40±2.86	56.36±2.01	64.98±3.42	64.21±3.52
	TF	54.50±1.16	70.78±0.17	56.18±7.00	72.33±0.69	67.96±0.95	74.26±1.57	69.41±0.60	73.06±0.51
Mean S.R.		47.31	71.63	54.37	70.19	71.92	69.50	72.55	**73.66**

Table 5: **Zero-shot camera pose estimation.** Trans. and Rot. denote 'translation' and 'rotation' errors respectively. The detailed metrics on the error calculation are listed in Appendix E.

Error	MoCoV3	MAE	DINOV2	CLIP	EVA	InternViT-300M	InternViT-6B	MVP	VC-1	SPA(Ours)
Trans. ($\times e^{-2}$)	2.29±0.07	2.15±0.07	6.55±0.07	4.21±0.37	5.49±0.24	4.62±0.14	5.39±0.41	2.15±0.12	2.02±0.07	**1.65±0.09**
Rot. ($\times e^{-1}$)	0.79±0.07	0.73±0.03	2.12±0.25	1.52±0.08	1.83±0.09	1.83±0.08	1.91±0.12	0.77±0.05	0.72±0.01	**0.61±0.01**

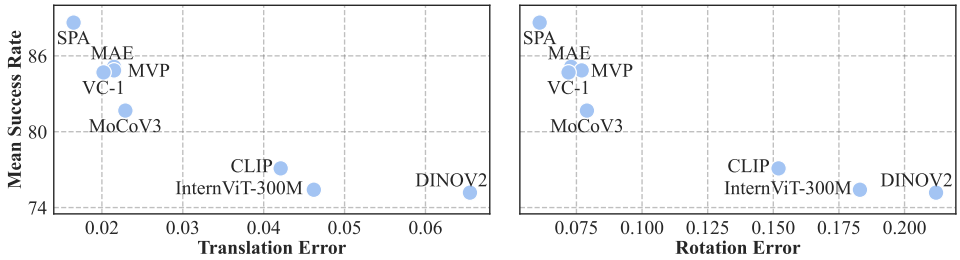

Figure 4: **Correlation between mean success rate and camera pose regression error.**

5.2 Additional Comparisons (Q1)

We primarily compare with SOTA methods using the ViT-L backbone, which is commonly available and pre-trained on large-scale datasets. However, some embodied-specific models are only offered in ViT-B variants. Therefore, we provide additional comparisons with several ViT-B models in Tab. 4. Our ViT-B version, SPA-B, also outperforms other baselines. Furthermore, when compared to SPA-L on VC-1 benchmarks, the mean success rate increases by 4.16 (73.66 → 77.82). This indicates that increasing the model size positively impacts SPA's performance.

5.3 Study on 3D Awareness of SPA (Q3)

Table 6: Additional ablations on VC-1.

Methods		SPA-B	SPA-MAE	RADIO	E-RADIO
VC-1	AD	52.00±3.46	55.33±3.06	55.33±3.06	56.67±2.31
	MW	92.00±4.16	90.67±6.00	72.00±9.23	83.47±4.11
	DMC	64.21±3.52	63.85±3.60	67.38±7.35	62.92±4.24
	TF	73.06±0.51	70.14±0.98	71.75±0.14	68.44±1.19
Mean S. R.		**73.66**	73.11	67.93	70.16

Firstly, we aim to provide clear evidence that the performance improvements of SPA are due to its 3D awareness. To demonstrate this, we conducted two additional ablation studies on the VC-1 benchmarks: 1) To determine whether the performance gain is due to SPA's pre-training objectives or the datasets used, we continue pre-training the ImageNet pre-trained MAE-B (the most competitive method besides SPA) on the same datasets used by SPA-B, referring to this model as SPA-MAE. Hyperparameters, including mask ratio and batch size, are kept at their default settings, and both the ImageNet pre-trained encoder and decoder weights are initially loaded. 2) Since SPA uses the feature map of RADIO for semantic rendering supervision, we also evaluate the original RADIO (653M parameters) and its efficient version, E-RADIO (391M parameters). Results are presented in Tab. 6.

Finding 5: The 3D-aware pre-training objective significantly enhances SPA's performance. It surpasses the single-image naive MAE with the same data. Notably, SPA learns superior representations compared to its semantic rendering teacher by a substantial margin.

Moreover, we provide both quantitative and qualitative evidence to demonstrate that SPA has acquired 3D awareness. For qualitative analysis, we visualize the zero-shot feature maps on multiview images of different encoder outputs, as shown in Fig. 5. The images are taken from the unseen Arkitscenes dataset. For quantitative analysis, we evaluate the zero-shot 3D awareness of various methods using a camera pose estimation task on the NAVI dataset (Jampani et al., 2023).

Specifically, given a pair of images from different viewpoints, we use a frozen encoder to extract features and concatenate them. A small MLP then regresses the relative camera pose and we report rotation and translation errors in Tab. 5. Details are in Appendix E. While El Banani et al. (2024) has explored 3D awareness of different vision models, their context differs. Their tasks can allow strong semantic models like DINOV2 to 'cheat'. For example, multiview correspondence can be achieved through semantic matching, and the relative depth estimation task involves transforming normalized values into discrete bins, resembling a per-pixel classification task. Additionally, they emphasize fine-grained dense local context, whereas, embodied AI focuses more on sparse, global information (Nair et al., 2022). Thus, we believe camera pose estimation, which predicts a global 'pose' from observations, is more relevant to embodied AI, where a policy must predict a global 'action'.

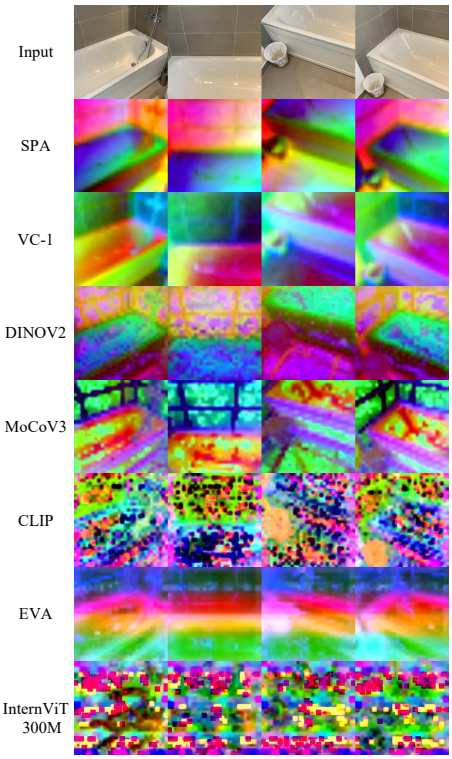

Figure 5: **Feature map visualization.**

Finding 6: We observe that SPA outperforms all other methods in zero-shot camera pose estimation. It achieved **an 18.3% improvement in translation and a 15.3% reduction in rotation error** compared to the second-best model. Additionally, we identify a **clear positive correlation** between camera pose estimation and embodied evaluation performance, as demonstrated in Fig. 4. This finding supports our spatial hypothesis and may offer valuable insights for future research on embodied representation.

Finding 7: The feature map visualization provides clear evidence that SPA has learned multi-view consistent knowledge, demonstrating its 3D awareness. Additionally, the features produced by SPA are **cleaner and more coherent**. Though VC-1 also generates smooth features, they are *not consistent across viewpoints*. The feature maps from the multi-modal approach are highly noisy and lack details.

5.4 Hyperparameter Investigation

We conduct hyperparameter tuning with a ViT-B model on ScanNet (Dai et al., 2017), and evaluate it on VC-1 benchmarks, as shown in Tab. 7. **1) Mask Ratio.** Our results indicate that a mask ratio of 0.5 is the most effective. **2) Loss Components.** As discussed in Sec. 2.4, our rendering loss consists of color, depth, and semantic components. We sequentially deactivate each and find that all three are valuable. However, deactivating the semantic loss has the least impact.

Table 7: **Mask ratio and loss components.** C., D., S. denote color, depth, and semantic.

Mask Ratio	Loss C.	Loss D.	Loss S.	VC-1 Benchmark AD	MW	DMC	TF	Mean S.R.
0.00	✓	✓	✓	53.3±4.6	88.5±5.7	57.5±2.6	74.1±0.6	70.36
0.25	✓	✓	✓	52.7±3.1	89.6±4.5	57.6±3.0	70.4±1.7	70.17
0.50	✓	✓	✓	53.3±4.2	88.8±1.6	60.1±3.1	72.6±0.7	**71.18**
0.75	✓	✓	✓	51.3±1.2	88.0±3.5	61.1±3.5	73.0±0.8	71.01
0.95	✓	✓	✓	51.3±1.2	85.6±4.0	62.5±5.3	73.1±0.2	70.67
0.50	✓	✗	✓	51.3±1.2	90.9±3.3	58.8±5.6	71.5±1.0	71.01
0.50	✗	✓	✓	52.0±2.0	89.3±3.3	53.9±4.3	70.9±1.3	68.71
0.50	✓	✓	✗	52.7±3.1	88.0±4.5	61.5±3.4	71.6±1.2	71.16

5.5 Real-World Experiments (Q4)

We conduct several real-world experiments to further investigate the generalization ability of different representations. Specifically, we utilize the open-sourced Low-Cost Robot Arm (Koch, 2024) to learn real-world tasks from pixels, with only 50 demonstrations per task using different frozen pre-trained representations. The robot performed two single-arm tasks: (1) picking a cube, and (2) stacking a yellow cube on a pink cube, as well as one dual-arm task: folding a cloth in half. Refer to Fig. 6 for illustrations and Appendix F for more details. We evaluate each task with 25 rollouts, with the results presented in Tab. 8. SPA consistently performs better on real-world tasks, suggesting that SPA's pre-trained representations can robustly adapt to real-world environments without finetuning.

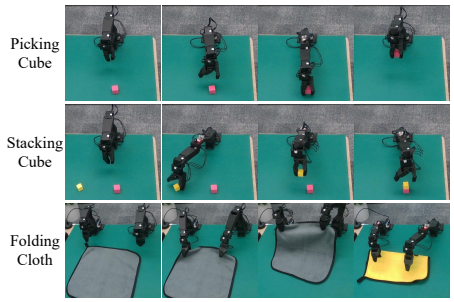

Figure 6: **Real-world task illustrations.**

Table 8: **Real-world experiment results.** S.R. denotes 'Success Rate'.

Methods	MoCoV3	MAE	DINOV2	CLIP	EVA	InternViT-300M	InternViT-6B	MVP	VC-1	SPA (Ours)
Picking Cube	28.00	64.00	20.00	28.00	56.00	32.00	52.00	36.00	40.00	64.00
Stacking Cube	16.00	32.00	4.00	16.00	8.00	8.00	36.00	20.00	16.00	48.00
Folding Cloth	48.00	64.00	32.00	24.00	28.00	48.00	44.00	64.00	60.00	84.00
Mean S.R.	30.67	53.33	18.67	22.67	30.67	29.33	44.00	40.00	38.67	**65.33**

6 Related Work

Representation Learning for Computer Vision. Recent advances in computer vision have increasingly focused on unsupervised and self-supervised learning to utilize large amounts of unlabeled data. Techniques like contrastive learning (Chen et al., 2020a; 2021; 2020b; He et al., 2020), masked autoencoders (He et al., 2022; Feichtenhofer et al., 2022; Bachmann et al., 2022; Tong et al., 2022; Wang et al., 2023), and self-distillation (Caron et al., 2021; Oquab et al., 2023; Ranzinger et al., 2024) have shown that effective representations can be learned without supervision. Moreover, multi-modal pre-training approaches (Radford et al., 2021; Fang et al., 2023b; Chen et al., 2024b) leverage language to learn more comprehensive representations. These developments have significantly improved transfer learning capabilities while also displaying zero-shot abilities.

Representation Learning for Embodied AI. Recent advances in embodied AI representation learning, inspired by computer vision, have applied techniques such as contrastive (Nair et al., 2022; Yang et al., 2023) and masked autoencoders (Radosavovic et al., 2023; Majumdar et al., 2023; Karamcheti et al., 2023; Yang et al., 2024b) to embodied AI. However, these approaches often emphasize semantic learning while overlooking the specific needs of embodied AI tasks. In this work, we propose a spatial hypothesis specifically for embodied AI representation learning, and we demonstrate how a standard 2D backbone can integrate 3D spatial awareness.

3D Robot Learning and 3D-Aware Computer Vision. Prior work in 3D robot learning has often relied on explicit 3D input (Zhu et al., 2024; Ze et al., 2024; Wang et al., 2024b;a; Shridhar et al., 2023; Chen et al., 2023), or lifting 2D features into 3D spaces (Ke et al., 2024; Goyal et al., 2024), providing a strong foundation for our spatial hypothesis. Given the scalability challenges of explicit 3D observations, some computer vision research has explored integrating 3D spatial awareness into 2D backbones (Yang et al., 2024a; Zhu et al., 2023b; Yue et al., 2025; Zhang et al., 2024). To the best of our knowledge, SPA is the first to systematically investigate this approach in embodied AI.

Neural Rendering. Recent advances in 3D vision, particularly in neural rendering (Mildenhall et al., 2021), have enabled the encoding of scenes using neural networks, which support differentiable rendering and reconstruction. Alongside improvements in neural rendering techniques themselves (Wang et al., 2021; Zhu et al., 2023a; Gropp et al., 2020; Ortiz et al., 2022; Wang et al., 2022), the Ponder series (Huang et al., 2023; Zhu et al., 2023b; Yang et al., 2024a) and subsequent works (Wang et al., 2024c; Irshad et al., 2024) have applied differentiable neural rendering for representation learning. However, they have primarily focused on 3D perception and autonomous driving scenarios. To the best of our knowledge, our work is the first to apply neural rendering for embodied AI representation learning using a standard 2D backbone, marking a novel contribution to this area of research.

7 Conclusion, Limitations, and Future Work

In this work, we propose that 3D spatial awareness is crucial for embodied AI and introduce SPA, a novel framework that pre-trains a standard ViT backbone with 3D spatial awareness. To validate our hypothesis, we conduct the largest-scale embodied evaluation to date, over 15 times larger than previous studies. Our experiments demonstrate the clear superiority of SPA and highlight the importance of 3D awareness. Despite strong results across simulated and real robotic tasks, limitations remain. Our evaluation is currently restricted to imitation learning (specifically behavior cloning), and exploring SPA's performance in other settings, such as reinforcement learning, presents an exciting future direction. Incorporating SPA into VLMs to enhance their performance on spatial aware tasks (Chen et al., 2024a; Majumdar et al., 2024) also represents an exciting research area. Additionally, SPA currently focuses on static multi-view scenes; extending it to dynamic, temporal scenarios could enhance its generality. Lastly, while we use the ViT encoder for fair comparison, the volume decoder's multi-view interaction knowledge could be leveraged in policy learning, offering further potential for improvement.

ACKNOWLEDGMENTS

This work is supported by the National Key R&D Program of China (NO.2022ZD0160102) and Shanghai Artificial Intelligence Laboratory.

REFERENCES

Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, et al. Scenescript: Reconstructing scenes with an autoregressive structured language model. *arXiv preprint arXiv:2403.13064*, 2024.

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pp. 348–367. Springer, 2022.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

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*, pp. 9650–9660, 2021.

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

Shizhe Chen, Ricardo Garcia, Cordelia Schmid, and Ivan Laptev. Polarnet: 3d point clouds for language-guided robotic manipulation. 2023.

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

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. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.

RealRobot Contributors. Realrobot: A project for open-sourced robot learning research. https://github.com/HaoyiZhu/RealRobot. 2024.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21795–21806, 2024.

Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11444–11453, 2020.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A robotic dataset for learning diverse skills in one-shot. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023a.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19358–19369, 2023b.

Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.

Minghao Gou, Hao-Shu Fang, Zhanda Zhu, Sheng Xu, Chenxi Wang, and Cewu Lu. Rgb matters: Learning 7-dof grasp poses on monocular rgbd images. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13459–13466. IEEE, 2021.

Ankit Goyal, Valts Blukis, Jie Xu, Yijie Guo, Yu-Wei Chao, and Dieter Fox. Rvt2: Learning precise manipulation from few demonstrations. *RSS*, 2024.

Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

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*, pp. 16000–16009, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Yingdong Hu, Renhao Wang, Li Erran Li, and Yang Gao. For pre-trained vision models in motor control, not all policy learning methods are created equal. In *International Conference on Machine Learning*, pp. 13628–13651. PMLR, 2023.

Di Huang, Sida Peng, Tong He, Honghui Yang, Xiaowei Zhou, and Wanli Ouyang. Ponder: Point cloud pre-training via neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16089–16098, 2023.

Muhammad Zubair Irshad, Sergey Zakharov, Vitor Guizilini, Adrien Gaidon, Zsolt Kira, and Rares Ambrus. Nerf-mae: Masked autoencoders for self-supervised 3d representation learning for neural radiance fields. In *European Conference on Computer Vision (ECCV)*, 2024.

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.

Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araujo, Ricardo Martin Brualla, Kaushal Patel, et al. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. *Advances in Neural Information Processing Systems*, 36:76061–76084, 2023.

Siddharth Karamcheti, Suraj Nair, Annie S. Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. In *Robotics: Science and Systems (RSS)*, 2023.

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

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Alexander Koch. Low-cost robot arm. https://github.com/AlexanderKoch-Koch/low_cost_robot, 2024. URL https://github.com/AlexanderKoch-Koch/low_cost_robot. GitHub repository.

Vikash Kumar. *Manipulators and Manipulation in high dimensional spaces*. PhD thesis, University of Washington, Seattle, 2016. URL https://digital.lib.washington.edu/researchworks/handle/1773/38104.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pp. 1–18. Springer, 2022.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017.

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

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16488–16498, 2024.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

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.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Joseph Ortiz, Alexander Clegg, Jing Dong, Edgar Sucar, David Novotny, Michael Zollhoefer, and Mustafa Mukadam. isdf: Real-time neural signed distance fields for robot perception. *arXiv preprint arXiv:2204.02296*, 2022.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20133–20143, 2023.

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

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

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12490–12500, 2024.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Jinghuan Shang, Karl Schmeckpeper, Brandon B. May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. *arXiv*, 2024.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pp. 785–799. PMLR, 2023.

Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pp. 369–386. SPIE, 2019.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020.

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024a.

Chenxi Wang, Hongjie Fang, Hao-Shu Fang, and Cewu Lu. Rise: 3d perception makes real-world robot imitation simple and effective. *arXiv preprint arXiv:2404.12281*, 2024b.

Jingwen Wang, Tymoteusz Bleja, and Lourdes Agapito. Go-surf: Neural feature grid optimization for fast, high-fidelity rgb-d surface reconstruction. In *2022 International Conference on 3D Vision (3DV)*, pp. 433–442. IEEE, 2022.

Letian Wang, Seung Wook Kim, Jiawei Yang, Cunjun Yu, Boris Ivanovic, Steven L Waslander, Yue Wang, Sanja Fidler, Marco Pavone, and Peter Karkus. Distillnerf: Perceiving 3d scenes from single-glance images by distilling neural fields and foundation model features. *arXiv preprint arXiv:2406.12095*, 2024c.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae V2: scaling video masked autoencoders with dual masking. In *CVPR*, pp. 14549–14560, 2023.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow. In *ICCV*, 2023.

Manuel Wüthrich, Felix Widmaier, Felix Grimminger, Joel Akpo, Shruti Joshi, Vaibhav Agrawal, Bilal Hammoud, Majid Khadiv, Miroslav Bogdanovic, Vincent Berenz, et al. Trifinger: An open-source robot for learning dexterity. *arXiv preprint arXiv:2008.03596*, 2020.

Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, Haoyi Zhu, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, et al. Unipad: A universal pre-training paradigm for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15238–15250, 2024a.

Jiange Yang, Sheng Guo, Gangshan Wu, and Limin Wang. Comae: single model hybrid pre-training on small-scale rgb-d datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3145–3154, 2023.

Jiange Yang, Bei Liu, Jianlong Fu, Bocheng Pan, Gangshan Wu, and Limin Wang. Spatiotemporal predictive pre-training for robotic motor control. *arXiv preprint arXiv:2403.05304*, 2024b.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. *arXiv preprint arXiv:2107.09645*, 2021.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5752–5761, 2021.

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*, pp. 1094–1100. PMLR, 2020.

Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning. In *European Conference on Computer Vision*, pp. 57–74. Springer, 2025.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy. *arXiv preprint arXiv:2403.03954*, 2024.

Sha Zhang, Jiajun Deng, Lei Bai, Houqiang Li, Wanli Ouyang, and Yanyong Zhang. Hvdistill: Transferring knowledge from images to point clouds via unsupervised hybrid-view distillation. *International Journal of Computer Vision*, pp. 1–15, 2024.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 519–535. Springer, 2020.

Haoyi Zhu, Hao-Shu Fang, and Cewu Lu. X-nerf: Explicit neural radiance field for multi-scene 360deg insufficient rgb-d views. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5766–5775, 2023a.

Haoyi Zhu, Honghui Yang, Xiaoyang Wu, Di Huang, Sha Zhang, Xianglong He, Tong He, Hengshuang Zhao, Chunhua Shen, Yu Qiao, et al. Ponderv2: Pave the way for 3d foundataion model with a universal pre-training paradigm. *arXiv preprint arXiv:2310.08586*, 2023b.

Haoyi Zhu, Yating Wang, Di Huang, Weicai Ye, Wanli Ouyang, and Tong He. Point cloud matters: Rethinking the impact of different observation spaces on robot learning. *arXiv preprint arXiv:2402.02500*, 2024.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

A Additional Rendering Losses

Here we detail the three additional rendering losses we have applied in Sec. 2.4.

Eikonal Regularization Loss. The Eikonal regularization loss, denoted as $\mathcal{L}_{\text{eikonal}}$, is a widely used loss function for the regularization of signed distance functions (SDFs) (Gropp et al., 2020). It is defined as:

$$\mathcal{L}_{\text{eikonal}} = \frac{1}{N_r N_p} \sum_{i=1}^{N_r} \sum_{j=1}^{N_p} (\|\nabla s(\mathbf{p}_{i,j})\| - 1)^2, \tag{9}$$

where $\nabla s(\mathbf{p}_{i,j})$ represents the gradient of the SDF s at the location $\mathbf{p}_{i,j}$. Since the SDF is a distance measure, $\mathcal{L}_{\text{eikonal}}$ encourages the gradients to have unit norm at the query point.

Near-Surface and Free Space Loss for SDF. To improve SDF estimation, we incorporate additional approximate SDF supervision, similar to iSDF (Ortiz et al., 2022) and GO-Surf (Wang et al., 2022). Specifically, for near-surface points, the difference between rendered depth and ground-truth depth serves as pseudo-SDF ground-truth supervision. For points far from the surface, a free space loss is used to further regularize the SDF values.

To compute the approximate SDF supervision, we define an indicator $b(z)$ for each sampled ray point with ray length z and corresponding ground-truth depth D:

$$b(z) = D - z. \tag{10}$$

The value $b(z)$ can be considered a credible approximate SDF value when it is small. Let t be a user-defined threshold, set to 0.05 in our experiments. For sampled ray points satisfying $b(z) \leq t$, we apply the near-surface SDF loss to constrain the SDF prediction $s(z_{i,j})$:

$$\mathcal{L}_{\text{sdf}} = \frac{1}{N_r N_p} \sum_{i=1}^{N_r} \sum_{j=1}^{N_p} |s(z_{i,j}) - b(z_{i,j})|. \tag{11}$$

For the remaining sampled ray points, we utilize a free space loss:

$$\mathcal{L}_{\text{free}} = \frac{1}{N_r N_p} \sum_{i=1}^{N_r} \sum_{j=1}^{N_p} \max\Big(0, e^{-\alpha \cdot s(z_{i,j})} - 1, s(z_{i,j}) - b(z_{i,j})\Big), \tag{12}$$

where α is set to 5, following Ortiz et al. (2022); Wang et al. (2022). Due to the presence of noisy depth images, $\mathcal{L}_{\text{sdf}}$ and $\mathcal{L}_{\text{free}}$ are applied only to rays with valid depth values.

In our experiments, we adopt a similar weighting scheme to GO-Surf (Wang et al., 2022), setting $\lambda_C = 10.0$, $\lambda_D = 1.0$, $\lambda_{\text{sdf}} = 10.0$, and $\lambda_{\text{free}} = 1.0$. We observe that the Eikonal term can lead to overly smooth reconstructions, so we use a small weight of 0.01 for the Eikonal loss.

B Evaluation Setups

Here we detail the setups of our large-scale evaluation in Sec. 3. For the detailed visualizations of each task, we recommend the readers to read the original simulator's or benchmark's dataset.

B.1 Single-Task Benchmarks

VC-1 (Majumdar et al., 2023). This benchmark includes several simulators. We selected four: Adroit (Kumar, 2016), Meta-World (Yu et al., 2020), DMControl (Tunyasuvunakool et al., 2020), and TriFinger (Wüthrich et al., 2020). The Adroit subset focuses on dexterous manipulation with 2 tasks: `Relocate` and `Pen`. The Meta-World subset addresses two-finger gripper manipulation with 5 tasks: `Button Press Topdown`, `Drawer Open`, `Bin Picking`, `Hammer`, and `Assembly`. The DMControl subset is for locomotion control, also with 5 tasks: `Walker Stand`, `Walker Walk`,

Reacher Easy, Cheetah Run, and Finger Spin. The TriFinger subset targets three-finger manipulation with 2 tasks: Reach Cube and Move Cube. For all tasks, we use a 3-layer MLP as the policy network for each single-task training, following the original implementation. Each task is trained with 100 demonstrations, except for 25 on Meta-World, and evaluated 50 times using the specific seeds 100, 200, and 300. The [CLS] token of a frozen pre-trained ViT is used as the observation feature. All hyper-parameters are kept the same with the original implementation.

Franka Kitchen (Gupta et al., 2019). Franka-Kitchen is a MuJoCo-modeled simulation environment with a Franka robot in a kitchen scene. Its action space is the 9-dimensional joint velocity with 7 DoF for the arm and 2 DoF for the gripper. Following previous works (Nair et al., 2022; Karamcheti et al., 2023), we evaluate five tasks: Sliding Door, Turning Light On, Opening Door, Turning Knob, and Opening Microwave. Each task spans two camera viewpoints and three random seeds. Similar to the evaluation scheme in VC-1, we utilize 25 demonstrations to train a policy model, which is a 2-layer MLP with hidden sizes [256, 256] preceded by a BatchNorm.

Meta-World (Yu et al., 2020). This benchmark comprises a series of tasks in which an agent directs a Sawyer robot arm to manipulate objects in a tabletop environment. We selected 48 tasks, encompassing easy, medium, and hard levels. We implemented the Diffusion Policy (Chi et al., 2023) on this benchmark and adhered to the setup in Ze et al. (2024) to generate 10 demonstrations for each single-task training, followed by evaluation through 20 rollouts. The average results across three fixed seeds (100, 200, 300) are reported. The [CLS] token from a frozen pre-trained ViT serves as the observation feature. The 48 tasks include: Button Press Wall, Door Close, Door Unlock, Drawer Close, Drawer Open, Faucet Close, Plate Slide, Plate Slide Back, Plate Slide Side, Window Close, Basketball, Bin Picking, Box Close, Coffee Push, Assembly, Disassemble, Push Wall, Shelf Place, Door Open, Button Press, Sweep Into, Door Lock, Reach Wall, Hammer, Stick Push, Button Press Topdown, Handle Press Side, Plate Slide Back Side, Sweep, Button Press Topdown Wall, Handle Press, Push, Coffee Pull, Dial Turn, Reach, Coffee Button, Pick Place Wall, Stick Pull, Hand Insert, Peg Insert Side, Pick Place, Faucet Open, Push Back, Lever Pull, Handle Pull, Soccer, Window Open, and Pick Out Of Hole.

B.2 Language-Conditioned Multi-Task Benchmarks

RLBench (James et al., 2020). This benchmark is a prominent language-conditioned multi-task robot learning framework. PolarNet (Chen et al., 2023) has categorized all tasks into 9 groups. We selected 71 tasks from RLBench that can be successfully executed and split them into two groups uniformly on categories: Group 1 with 35 tasks and Group 2 with 36 tasks. Each task includes 100 training demonstrations and 25 testing rollouts. For each group, we train a language-conditioned multi-task agent. We employ RVT-2 (Goyal et al., 2024), the state-of-the-art (SOTA) method on this benchmark, as our policy. RVT-2 takes multiple images rendered from point clouds as inputs and uses a convolutional block to generate feature maps. We substitute the convolutional block with different pre-trained ViTs, unpatchifying the latent vectors concatenated with the global [CLS] token to obtain feature maps. All other architectures and hyperparameters remain consistent with the original RVT-2 implementation.

The 35 tasks in Group 1 include: Basketball In Hoop, Put Rubbish In Bin, Meat Off Grill, Meat On Grill, Slide Block To Target, Reach And Drag, Take Frame Off Hanger, Water Plants, Hang Frame On Hanger, Wipe Desk, Stack Blocks, Reach Target, Push Button, Lamp On, Toilet Seat Down, Close Laptop Lid, Open Box, Open Drawer, Pick Up Cup, Turn Tap, Take Usb Out Of Computer, Play Jenga, Insert Onto Square Peg, Take Umbrella Out Of Umbrella Stand, Insert Usb In Computer, Straighten Rope, Turn Oven On, Change Clock, Close Microwave, Close Fridge, Close Grill, Open Grill, Unplug Charger, Press Switch, and Take Money Out Safe.

The 36 tasks in Group 2 include: Change Channel, Tv On, Push Buttons, Stack Wine, Scoop With Spatula, Place Hanger On Rack, Move Hanger, Sweep To Dustpan, Take Plate Off Colored Dish Rack, Screw Nail, Take Shoes Out Of Box, Slide Cabinet Open And Place Cups, Lamp Off, Pick And Lift,

`Take Lid Off Saucepan`, `Close Drawer`, `Close Box`, `Phone On Base`, `Toilet Seat Up`, `Put Books On Bookshelf`, `Beat The Buzz`, `Stack Cups`, `Put Knife On Chopping Board`, `Place Shape In Shape Sorter`, `Take Toilet Roll Off Stand`, `Put Umbrella In Umbrella Stand`, `Setup Checkers`, `Open Window`, `Open Wine Bottle`, `Open Microwave`, `Put Money In Safe`, `Open Door`, `Close Door`, `Open Fridge`, `Open Oven`, and `Plug Charger In Power Supply`.

LIBERO (Liu et al., 2024). Built upon Robosuite (Zhu et al., 2020), LIBERO (Liu et al., 2024) generates a total of 130 language-conditioned tasks across five suites: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, LIBERO-10, and LIBERO-90. Each suite contains 10 tasks, except for LIBERO-90, which includes 90 tasks. We train a language-conditioned multi-task policy for each suite, adopting the transformer policy provided by LIBERO. The image encoders are modified from default CNNs to frozen pre-trained ViTs, utilizing the `[CLS]` token for feature extraction. To expedite policy training, we use only 20 demonstrations per task and forgo augmentations, allowing for pre-extraction of all image features during training. After training for 25 epochs, the checkpoints from the 20th and 25th are evaluated with 20 rollouts per task, and the best checkpoint's performance is taken. Finally, the results are averaged on 3 random seeds.

The 10 tasks in LIBERO-Spatial are:
1. Pick up the black bowl between the plate and the ramekin and place it on the plate.
2. Pick up the black bowl next to the ramekin and place it on the plate.
3. Pick up the black bowl from table center and place it on the plate.
4. Pick up the black bowl on the cookie box and place it on the plate.
5. Pick up the black bowl in the top drawer of the wooden cabinet and place it on the plate.
6. Pick up the black bowl on the ramekin and place it on the plate.
7. Pick up the black bowl next to the cookie box and place it on the plate.
8. Pick up the black bowl on the stove and place it on the plate.
9. Pick up the black bowl next to the plate and place it on the plate.
10. Pick up the black bowl on the wooden cabinet and place it on the plate.

The 10 tasks in LIBERO-Object are:
1. Pick up the alphabet soup and place it in the basket.
2. Pick up the cream cheese and place it in the basket.
3. Pick up the salad dressing and place it in the basket.
4. Pick up the BBQ sauce and place it in the basket.
5. Pick up the ketchup and place it in the basket.
6. Pick up the tomato sauce and place it in the basket.
7. Pick up the butter and place it in the basket.
8. Pick up the milk and place it in the basket.
9. Pick up the chocolate pudding and place it in the basket.
10. Pick up the orange juice and place it in the basket.

The 10 tasks in LIBERO-Goal are:
1. Open the middle drawer of the cabinet.
2. Put the bowl on the stove.
3. Put the wine bottle on top of the cabinet.
4. Open the top drawer and put the bowl inside.
5. Put the bowl on top of the cabinet.
6. Push the plate to the front of the stove.
7. Put the cream cheese in the bowl.
8. Turn on the stove.
9. Put the bowl on the plate.
10. Put the wine bottle on the rack.

The 10 tasks in LIBERO-10 are:
1. LIVING ROOM SCENE2: Put both the alphabet soup and the tomato sauce in the basket.
2. LIVING ROOM SCENE2: Put both the cream cheese box and the butter in the basket.
3. KITCHEN SCENE3: Turn on the stove and put the moka pot on it.
4. KITCHEN SCENE4: Put the black bowl in the bottom drawer of the cabinet and close it.
5. LIVING ROOM SCENE5: Put the white mug on the left plate and put the yellow and white mug on the right plate.

6. STUDY SCENE1: Pick up the book and place it in the back compartment of the caddy.
7. LIVING ROOM SCENE6: Put the white mug on the plate and put the chocolate pudding to the right of the plate.
8. LIVING ROOM SCENE1: Put both the alphabet soup and the cream cheese box in the basket.
9. KITCHEN SCENE8: Put both moka pots on the stove.
10. KITCHEN SCENE6: Put the yellow and white mug in the microwave and close it.

The 10 tasks in LIBERO-90 are:
1. KITCHEN SCENE10: Close the top drawer of the cabinet.
2. KITCHEN SCENE10: Close the top drawer of the cabinet and put the black bowl on top of it.
3. KITCHEN SCENE10: Put the black bowl in the top drawer of the cabinet.
4. KITCHEN SCENE10: Put the butter at the back in the top drawer of the cabinet and close it.
5. KITCHEN SCENE10: Put the butter at the front in the top drawer of the cabinet and close it.
6. KITCHEN SCENE10: Put the chocolate pudding in the top drawer of the cabinet and close it.
7. KITCHEN SCENE1: Open the bottom drawer of the cabinet.
8. KITCHEN SCENE1: Open the top drawer of the cabinet.
9. KITCHEN SCENE1: Open the top drawer of the cabinet and put the bowl in it.
10. KITCHEN SCENE1: Put the black bowl on the plate.
11. KITCHEN SCENE1: Put the black bowl on top of the cabinet.
12. KITCHEN SCENE2: Open the top drawer of the cabinet.
13. KITCHEN SCENE2: Put the black bowl at the back on the plate.
14. KITCHEN SCENE2: Put the black bowl at the front on the plate.
15. KITCHEN SCENE2: Put the middle black bowl on the plate.
16. KITCHEN SCENE2: Put the middle black bowl on top of the cabinet.
17. KITCHEN SCENE2: Stack the black bowl at the front on the black bowl in the middle.
18. KITCHEN SCENE2: Stack the middle black bowl on the back black bowl.
19. KITCHEN SCENE3: Put the frying pan on the stove.
20. KITCHEN SCENE3: Put the moka pot on the stove.
21. KITCHEN SCENE3: Turn on the stove.
22. KITCHEN SCENE3: Turn on the stove and put the frying pan on it.
23. KITCHEN SCENE4: Close the bottom drawer of the cabinet.
24. KITCHEN SCENE4: Close the bottom drawer of the cabinet and open the top drawer.
25. KITCHEN SCENE4: Put the black bowl in the bottom drawer of the cabinet.
26. KITCHEN SCENE4: Put the black bowl on top of the cabinet.
27. KITCHEN SCENE4: Put the wine bottle in the bottom drawer of the cabinet.
28. KITCHEN SCENE4: Put the wine bottle on the wine rack.
29. KITCHEN SCENE5: Close the top drawer of the cabinet.
30. KITCHEN SCENE5: Put the black bowl in the top drawer of the cabinet.
31. KITCHEN SCENE5: Put the black bowl on the plate.
32. KITCHEN SCENE5: Put the black bowl on top of the cabinet.
33. KITCHEN SCENE5: Put the ketchup in the top drawer of the cabinet.
34. KITCHEN SCENE6: Close the microwave.
35. KITCHEN SCENE6: Put the yellow and white mug to the front of the white mug.
36. KITCHEN SCENE7: Open the microwave.
37. KITCHEN SCENE7: Put the white bowl on the plate.
38. KITCHEN SCENE7: Put the white bowl to the right of the plate.
39. KITCHEN SCENE8: Put the right moka pot on the stove.
40. KITCHEN SCENE8: Turn off the stove.
41. KITCHEN SCENE9: Put the frying pan on the cabinet shelf.
42. KITCHEN SCENE9: Put the frying pan on top of the cabinet.
43. KITCHEN SCENE9: Put the frying pan under the cabinet shelf.
44. KITCHEN SCENE9: Put the white bowl on top of the cabinet.
45. KITCHEN SCENE9: Turn on the stove.
46. KITCHEN SCENE9: Turn on the stove and put the frying pan on it.
47. LIVING ROOM SCENE1: Pick up the alphabet soup and put it in the basket.
48. LIVING ROOM SCENE1: Pick up the cream cheese box and put it in the basket.
49. LIVING ROOM SCENE1: Pick up the ketchup and put it in the basket.
50. LIVING ROOM SCENE1: Pick up the tomato sauce and put it in the basket.
51. LIVING ROOM SCENE2: Pick up the alphabet soup and put it in the basket.

52. LIVING ROOM SCENE2: Pick up the butter and put it in the basket.
53. LIVING ROOM SCENE2: Pick up the milk and put it in the basket.
54. LIVING ROOM SCENE2: Pick up the orange juice and put it in the basket.
55. LIVING ROOM SCENE2: Pick up the tomato sauce and put it in the basket.
56. LIVING ROOM SCENE3: Pick up the alphabet soup and put it in the tray.
57. LIVING ROOM SCENE3: Pick up the butter and put it in the tray.
58. LIVING ROOM SCENE3: Pick up the cream cheese and put it in the tray.
59. LIVING ROOM SCENE3: Pick up the ketchup and put it in the tray.
60. LIVING ROOM SCENE3: Pick up the tomato sauce and put it in the tray.
61. LIVING ROOM SCENE4: Pick up the black bowl on the left and put it in the tray.
62. LIVING ROOM SCENE4: Pick up the chocolate pudding and put it in the tray.
63. LIVING ROOM SCENE4: Pick up the salad dressing and put it in the tray.
64. LIVING ROOM SCENE4: Stack the left bowl on the right bowl and place them in the tray.
65. LIVING ROOM SCENE4: Stack the right bowl on the left bowl and place them in the tray.
66. LIVING ROOM SCENE5: Put the red mug on the left plate.
67. LIVING ROOM SCENE5: Put the red mug on the right plate.
68. LIVING ROOM SCENE5: Put the white mug on the left plate.
69. LIVING ROOM SCENE5: Put the yellow and white mug on the right plate.
70. LIVING ROOM SCENE6: Put the chocolate pudding to the left of the plate.
71. LIVING ROOM SCENE6: Put the chocolate pudding to the right of the plate.
72. LIVING ROOM SCENE6: Put the red mug on the plate.
73. LIVING ROOM SCENE6: Put the white mug on the plate.
74. STUDY SCENE1: Pick up the book and place it in the front compartment of the caddy.
75. STUDY SCENE1: Pick up the book and place it in the left compartment of the caddy.
76. STUDY SCENE1: Pick up the book and place it in the right compartment of the caddy.
77. STUDY SCENE1: Pick up the yellow and white mug and place it to the right of the caddy.
78. STUDY SCENE2: Pick up the book and place it in the back compartment of the caddy.
79. STUDY SCENE2: Pick up the book and place it in the front compartment of the caddy.
80. STUDY SCENE2: Pick up the book and place it in the left compartment of the caddy.
81. STUDY SCENE2: Pick up the book and place it in the right compartment of the caddy.
82. STUDY SCENE3: Pick up the book and place it in the front compartment of the caddy.
83. STUDY SCENE3: Pick up the book and place it in the left compartment of the caddy.
84. STUDY SCENE3: Pick up the book and place it in the right compartment of the caddy.
85. STUDY SCENE3: Pick up the red mug and place it to the right of the caddy.
86. STUDY SCENE3: Pick up the white mug and place it to the right of the caddy.
87. STUDY SCENE4: Pick up the book in the middle and place it on the cabinet shelf.
88. STUDY SCENE4: Pick up the book on the left and place it on top of the shelf.
89. STUDY SCENE4: Pick up the book on the right and place it on the cabinet shelf.
90. STUDY SCENE4: Pick up the book on the right and place it under the cabinet shelf.

C More Implementation Details

C.1 Dataset Details

The datasets used for SPA include ScanNet, ScanNet++, Hypersim, ADT, S3DIS, and Droid.

ScanNet consists of 1.89 million frames in total. Each epoch includes 1.5 times the dataset size. For each scene, a random starting frame is selected, followed by the sampling of 1 to 8 frames at random, with an interval of 8 frames between them.

ScanNet++ comprises 0.11 million frames. Each epoch includes 5 times the dataset size. For each scene, a random starting frame is selected, followed by the sampling of 1 to 8 frames at random, with an interval of 5 frames between them.

Hypersim contains 0.03 million frames. Each epoch includes 8 times the dataset size. For each scene, we randomly select 1 to 8 continuous frames.

ADT consists of 0.0015M frames in total. Each epoch includes 1 times the dataset size. For each scene, 1 to 8 continuous frames are randomly selected.

S3DIS consists of 0.015 million frames. Each epoch includes 5 times the dataset size. For each scene, a random starting frame is selected, followed by the sampling of 1 to 8 frames at random, with an interval of 5 frames between them.

Droid contains a large number of videos, but due to the high similarity between frames, the videos are first downsampled by a factor of 15 during pre-processing, resulting in 1.78 million frames. Since Droid does not provide depth data, we utilize Croco-Stereo Weinzaepfel et al. (2023) to estimate dense depth maps for rendering supervision. Additionally, due to the significant noise in the camera pose data, only a single frame is sampled at a time during training.

During pre-training, we first resize the multi-view input images to slightly larger than 224×224, and then randomly crop them to a final size of 224×224. Random photometric distortions with a probability of 0.5 are applied for augmentation, including brightness ranging from 0.875 to 1.125, contrast ranging from 0.5 to 1.5, saturation ranging from 0.5 to 1.5, and hue ranging from -0.05 to 0.05. Frames with very small valid depth areas or scene boxes are filtered out.

For semantic rendering supervision, we observe that using larger image sizes improves the quality of feature maps generated by RADIO. Consequently, we resize the images to 1024×1024 before feeding them into RADIO, which outputs a feature map of size 64×64. We then apply bilinear sampling to query the semantic feature labels for each pixel.

C.2 Pre-Training Details

For stability during pre-training, we apply the Exponential Moving Average (EMA) with a decay rate of 0.999. The model is trained for 2000 epochs on 80 NVIDIA A100-80G GPUs, using a gradient clipping threshold of 1.0. Each GPU processes a batch size of 2, with 8 gradient accumulation steps, resulting in a total effective batch size of $2 \times 80 \times 8 = 1280$. We employ the AdamW optimizer with a weight decay of 0.04. The base learning rate is set to 5×10^{-6}, and the actual learning rate is scaled by a factor of 8 times the effective batch size. A OneCycle learning rate scheduler is used, with a percentage start of 0.05, a divide factor of 100, and a final divide factor of 1000.

To facilitate faster convergence and improve stability, we initialize the encoder with ImageNet pre-trained weights from the Masked Autoencoder (MAE), applying a learning rate layer decay of 0.8. This initialization does not affect the validity of our conclusions, as demonstrated by the ablation study of SPA-MAE in Sec. 5.3. The ViT encoder and upsampling layers are trained with FP16 precision, while the volume decoder is trained with FP32 precision.

We set the loss weights to $\lambda_{\text{color}} = 10$, $\lambda_{\text{depth}} = 1$, $\lambda_{\text{semantic}} = 1$, $\lambda_{\text{eikonal}} = 0.01$, $\lambda_{\text{free}} = 1$, and $\lambda_{\text{sdf}} = 10$. For the NeuS sampler, the initial number of samples is set to 72, with 24 importance samples. In each iteration, we randomly sample 512 pixels per view for rendering and supervision.

D Detailed Results of Each Task

We present the results of all individual tasks in Tab. 12, Tab. 13, Tab. 14, Tab. 15, Tab. 16, and Tab. 17.

E Camera Pose Estimation Details

We adopt a setup similar to that of El Banani et al. (2024) for camera pose estimation using the NAVI dataset (Jampani et al., 2023). Given an image pair from different viewpoints, we first extract features from each image using a frozen, pre-trained Vision Transformer (ViT) encoder. Following standard protocols for embodied evaluation, we use the `[CLS]` token as the feature representation. The two `[CLS]` tokens are then concatenated and passed through a BatchNorm layer and a Multi-Layer Perceptron (MLP) to regress the camera pose. The MLP consists of four linear layers with three ReLU activations, using hidden sizes of 512, 256, and 128 units, and outputs a 7-dimensional pose vector. The first three dimensions represent the xyz translation, while the last four dimensions correspond to the rotation quaternions.

We employ the Mean Squared Error (MSE) loss function and optimize the model using the AdamW optimizer with a OneCycle learning rate scheduler. The model is trained for 100 epochs with a base learning rate of 1×10^{-3} and a starting percentage of 0.1. For evaluation, we use Euclidean distance

as the translation error metric and geodesic distance as the rotation error metric. The geodesic distance between two quaternions q_1 and q_2 is defined as:

$$\theta = 2 \cdot \arccos(|q_1 \cdot q_2|), \tag{13}$$

where q_1 and q_2 are normalized quaternions, and $\cdot$ denotes the quaternion dot product. The Euclidean distance d between two translation vectors $\mathbf{t}_1 = (x_1, y_1, z_1)$ and $\mathbf{t}_2 = (x_2, y_2, z_2)$ is given by:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}. \tag{14}$$

F Real-World Experiment Details

Our real-world hardware setup is based on the open-source Low-Cost-Robot project (Koch, 2024). We utilize two Intel RealSense D415 cameras for image capture. A visualization of our platform is provided in Fig. 8. For teleoperation, policy training, and evaluation, we leverage the open-source RealRobot project (Contributors, 2024). The policy used is the ACT policy (Zhao et al., 2023).

For each task, we collect 50 demonstrations, and during evaluation, we conduct 25 rollouts, each with randomized object locations and orientations. The model is trained for 10,000 epochs using four NVIDIA A100 GPUs. We employ the AdamW optimizer with a learning rate of 5×10^{-5} and a weight decay of 0.05. Additionally, a OneCycle learning rate scheduler is used, with a starting percentage of 0.1, a division factor of 10, and a final division factor of 100.

G Additional Reinforcement Learning Experiments

We conduct additional RL experiments following the settings in Hu et al. (2023) to use DrQ-v2 (Yarats et al., 2021), a state-of-the-art off-policy actor-critic approach for continuous vision-based control. We train some RL experiments with different pre-trained vision representations with ViT-Base architectures. The vision encoders are frozen during RL training. Five tasks in the Meta-World benchmark are chosen, as shown below. We train for a total of 1.1M frames and all other hyper-parameters including random seeds are kept as default and same. We run three seeds for each experiment. We report the evaluation success rate and episode reward below in Tab. 9. The reward curves are visualized in Fig. 7. From the results, it is evident that the reinforcement learning outcomes exhibit high variance. Nevertheless, overall, our 3D spatial-aware representation outperforms other representation learning methods.

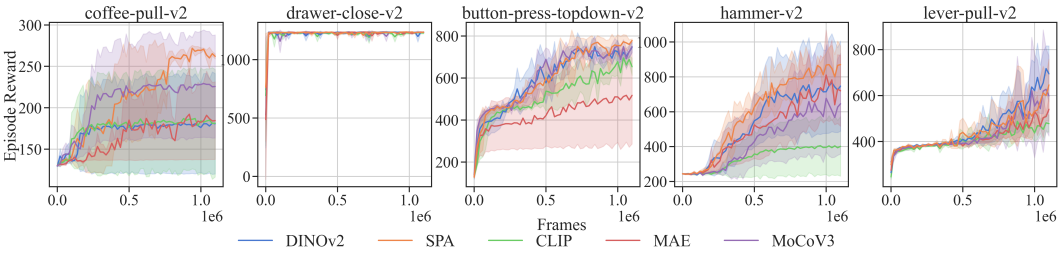

Figure 7: **Reinforcement learning reward curves visualization.**

H Additional Monocular Grasp Pose Detection Experiments

We conduct a monocular grasp pose detection experiment to further investigate more complex robotics learning paradigms. We follow similar settings in Gou et al. (2021), which train a neural network to detect the 7-DoF grasp poses on monocular image observations. The experiment is conducted on GraspNet-1Billion (Fang et al., 2020), a large-scale real-world object grasping benchmark. We follow the hyper-parameters and setups in the official implementation, except that we replace the default ResNet with different pre-trained ViT models for feature extraction. All pre-trained representations are with ViT-Base architecture and are frozen during training. We report the overall Top-K accuracy on the test set below. The results align well with our findings and indicate that SPA also outperforms other representation learning methods in the monocular grasp pose detection task.

Table 9: **Reinforcement learning comparison results.**

Meta-World RL Task	Method (ViT-B)	Success Rate	Episode Reward
button-press-topdown-v2	CLIP	0.93	653.97
	DINOv2	**1.00**	746.04
	MAE	0.46	517.54
	MoCoV3	0.99	749.93
	SPA (Ours)	**1.00**	**778.47**
hammer-v2	CLIP	0.00	401.41
	DINOv2	0.67	746.74
	MAE	0.66	720.19
	MoCoV3	0.59	645.46
	SPA (Ours)	**1.00**	**870.32**
lever-pull-v2	CLIP	0.00	478.18
	DINOv2	0.00	**694.73**
	MAE	0.00	540.44
	MoCoV3	**0.23**	598.54
	SPA (Ours)	0.15	646.33
coffee-pull-v2	CLIP	0.00	181.40
	DINOv2	0.00	180.72
	MAE	0.00	184.56
	MoCoV3	0.00	225.73
	SPA (Ours)	**0.00**	**262.11**
drawer-close-v2	CLIP	1.00	1228.90
	DINOv2	1.00	**1236.30**
	MAE	1.00	1233.91
	MoCoV3	1.00	1233.46
	SPA (Ours)	**1.00**	1235.81
Mean	CLIP	0.39	588.77
	DINOv2	0.53	720.91
	MAE	0.42	639.33
	MoCoV3	0.56	690.63
	SPA (Ours)	**0.63**	**758.61**

Table 10: **Overall top-k monocular grasp pose detection accuracy of various methods (ViT-Base).**

Method (ViT-Base)	CLIP	DINOv2	MoCoV3	MAE	SPA
Overall Accuracy	21.10	22.08	29.39	31.03	**31.20**

I Additional Ablation Study on Neural Rendering

To clarify the contribution of neural rendering to the overall performance of SPA, we conducted an additional ablation study. In this study, we maintained all settings identical—data loading, training techniques, hyperparameters, and the encoder—while replacing the volume neural rendering decoder with a multiview transformer-based decoder, similar to the MAE decoder. This alternative decoder receives masked patches filled with mask tokens corresponding to multiview images. Additional camera pose embeddings are added, and attention layers are used to fuse the multiview information and reconstruct RGB and depth images. We refer to this baseline as MV-MAE. It was trained on the ScanNet dataset without semantic supervision, ensuring a fair comparison with the result in the last line of Tab. 7. The results from this experiment demonstrate that neural rendering is crucial for incorporating explicit 3D spatial information. Simple multiview attention-based interaction, as used in MV-MAE, does not perform as effectively in learning 3D spatial awareness.

Table 11: **Additional Ablation Study on Neural Rendering.** The models are evaluated on two subsets of the VC-1 benchmark. The model architectures are both ViT-base.

Method	Meta-World	DMControl
MV-MAE	84.8±5.8	59.6±3.2
SPA	**88.0±4.5**	**61.5±3.4**

Table 12: **All results on Vc-1 benchmarks.**

Benchmark		AD		MW					DMC					TF	
Methods	Seed	Relocate	Pen	Button Press Topdown	Drawer Open	Bin Picking	Hammer	Assembly	Walker Stand	Walker Walk	Reacher Easy	Cheetah Run	Finger Spin	Reach Cube	Move Cube
ViT-L Methods															
MoCoV3	100	40.00	92.00	88.00	100.00	88.00	100.00	88.00	84.88	57.59	92.29	56.28	70.49	84.37	61.20
	200	36.00	80.00	88.00	100.00	80.00	100.00	84.00	82.95	55.02	92.08	43.17	69.49	84.20	61.26
	300	28.00	76.00	84.00	100.00	68.00	92.00	72.00	81.42	53.59	91.96	41.27	68.23	84.09	64.24
MAE	100	36.00	84.00	84.00	100.00	88.00	100.00	100.00	951.27	680.69	976.50	482.47	703.30	85.46	59.46
	200	36.00	80.00	84.00	100.00	76.00	96.00	96.00	933.67	676.92	952.20	49.22	695.00	86.88	59.45
	300	32.00	80.00	68.00	100.00	72.00	98.00	88.00	873.53	659.41	895.60	501.91	691.80	85.26	61.69
DINOV2	100	32.00	68.00	68.00	100.00	84.00	100.00	88.00	87.01	56.52	94.50	26.98	70.87	86.16	50.84
	200	28.00	68.00	60.00	100.00	80.00	96.00	80.00	86.00	53.97	89.97	21.84	68.78	86.87	50.78
	300	28.00	60.00	60.00	100.00	80.00	92.00	72.00	82.41	51.69	88.36	21.34	67.41	86.05	50.17
CLIP	100	24.00	80.00	28.00	100.00	88.00	100.00	84.00	66.16	43.94	90.71	18.40	68.30	73.28	41.09
	200	24.00	72.00	24.00	100.00	84.00	100.00	80.00	64.04	34.51	88.26	16.52	66.68	75.11	33.45
	300	24.00	68.00	16.00	100.00	84.00	96.00	72.00	52.70	31.60	85.17	14.51	67.49	74.73	38.95
EVA	100	44.00	84.00	72.00	100.00	76.00	100.00	100.00	77.92	51.64	98.17	31.04	70.17	82.56	52.04
	200	40.00	76.00	84.00	100.00	72.00	100.00	100.00	77.63	50.81	86.66	19.37	67.43	81.72	52.07
	300	32.00	72.00	96.00	100.00	68.00	96.00	96.00	77.21	47.71	88.41	29.19	67.43	82.13	52.70
InternViT-300M	100	40.00	72.00	80.00	100.00	72.00	100.00	84.00	70.04	30.44	80.80	16.70	67.05	78.59	53.67
	200	28.00	72.00	68.00	100.00	76.00	96.00	84.00	67.63	31.55	82.33	19.39	67.57	77.07	55.21
	300	28.00	80.00	60.00	100.00	72.00	96.00	72.00	66.95	29.28	81.87	18.80	68.55	78.27	48.58
InternViT-6B	100	32.00	72.00	88.00	100.00	80.00	100.00	84.00	88.53	70.02	93.09	26.54	70.62	85.96	57.52
	200	40.00	76.00	84.00	100.00	76.00	100.00	80.00	85.28	60.09	90.86	22.84	69.04	86.30	54.11
	300	60.00	80.00	80.00	100.00	88.00	100.00	76.00	81.88	59.17	87.87	21.53	67.20	85.68	54.86
MVP	100	32.00	84.00	96.00	100.00	96.00	100.00	100.00	84.88	57.59	92.29	56.28	70.49	84.37	61.20
	200	28.00	76.00	92.00	100.00	84.00	100.00	96.00	82.95	55.02	92.08	43.17	69.49	84.20	61.26
	300	24.00	76.00	84.00	100.00	68.00	100.00	88.00	81.42	53.59	91.96	41.27	68.23	84.09	64.24
VC-1	100	32.00	84.00	84.00	100.00	76.00	96.00	96.00	82.36	55.33	98.09	35.31	72.60	83.36	58.00
	200	28.00	80.00	68.00	100.00	72.00	92.00	84.00	80.21	53.90	89.83	34.10	70.15	83.17	61.00
	300	24.00	76.00	76.00	100.00	96.00	88.00	84.00	68.62	50.13	87.89	31.18	70.11	82.75	57.16
SPA-L	100	40.00	88.00	76.00	100.00	92.00	100.00	100.00	94.19	66.34	95.57	52.53	73.95	87.37	56.68
	200	44.00	76.00	84.00	100.00	88.00	100.00	84.00	92.28	60.60	81.43	44.99	71.83	87.26	64.35
	300	36.00	76.00	96.00	100.00	88.00	96.00	96.00	87.87	51.75	83.86	39.10	70.91	87.62	58.02
ViT-B Methods and Others															
STP-B	100	20.00	80.00	88.00	100.00	84.00	100.00	96.00	77.02	45.34	87.97	40.01	72.72	80.41	54.66
	200	28.00	76.00	92.00	100.00	84.00	100.00	80.00	71.50	33.60	84.08	34.30	72.18	80.13	57.97
	300	32.00	76.00	88.00	100.00	72.00	100.00	96.00	71.44	42.86	79.67	39.17	69.12	80.65	53.95
R3M-B	100	20.00	92.00	52.00	96.00	32.00	88.00	48.00	668.49	301.54	842.90	256.56	678.00	75.08	45.66
	200	12.00	76.00	48.00	96.00	32.00	88.00	44.00	634.62	256.82	661.40	198.63	660.90	75..62	48.09
	300	12.00	76.00	32.00	88.00	28.00	76.00	40.00	633.39	211.90	585.50	188.15	657.30	74.54	45.59
Theia-B	100	32.00	76.00	88.00	100.00	80.00	96.00	96.00	72.90	43.97	82.09	37.02	70.50	84.55	55.62
	200	36.00	80.00	60.00	100.00	84.00	100.00	84.00	79.05	56.99	94.36	39.59	70.22	83.27	54.59
	300	24.00	72.00	80.00	100.00	72.00	96.00	100.00	79.64	54.39	82.89	39.09	72.00	84.01	54.43
Voltron-B	100	16.00	72.00	76.00	100.00	64.00	100.00	96.00	74.31	42.05	68.88	36.94	70.91	86.28	65.11
	200	32.00	72.00	76.00	100.00	60.00	96.00	88.00	71.57	38.17	67.53	31.01	70.17	86.61	62.39
	300	20.00	68.00	72.00	100.00	52.00	96.00	84.00	71.25	36.50	66.14	30.11	69.88	86.16	59.02
MAE-B	100	24.00	88.00	88.00	100.00	84.00	96.00	96.00	88.28	42.55	95.18	44.08	69.26	85.63	55.68
	200	28.00	76.00	84.00	100.00	84.00	88.00	88.00	77.13	38.49	88.22	32.75	69.02	85.14	56.81
	300	28.00	72.00	76.00	100.00	80.00	88.00	80.00	75.60	36.93	78.35	31.03	69.01	84.11	57.30
DINOV2-B	100	8.00	60.00	40.00	100.00	44.00	96.00	20.00	45.95	16.61	63.57	13.38	60.11	74.07	36.29
	200	8.00	68.00	40.00	100.00	64.00	88.00	16.00	37.96	15.81	51.44	12.59	59.56	74.18	32.14
	300	12.00	64.00	48.00	100.00	64.00	88.00	4.00	32.43	14.31	36.01	11.67	56.54	73.77	36.53
VC-1-B	100	20.00	76.00	76.00	100.00	76.00	100.00	76.00	72.35	43.14	92.77	27.31	68.67	84.19	62.00
	200	32.00	80.00	68.00	100.00	76.00	100.00	92.00	81.83	44.05	83.62	27.80	70.98	83.88	59.63
	300	24.00	68.00	80.00	100.00	80.00	88.00	88.00	83.01	41.25	77.60	28.53	70.89	84.76	59.51
RADIO	100	28.00	76.00	48.00	100.00	72.00	100.00	84.00	87.84	62.72	96.53	15.71	67.88	85.70	57.52
	200	36.00	76.00	44.00	100.00	72.00	96.00	52.00	80.26	57.39	95.93	15.26	67.51	85.67	57.81
	300	44.00	72.00	32.00	100.00	40.00	92.00	48.00	79.62	53.51	89.16	14.80	66.57	85.64	58.14
E-RADIO	100	32.00	84.00	64.00	100.00	84.00	96.00	96.00	71.47	53.41	93.01	50.19	70.75	87.17	46.97
	200	32.00	84.00	60.00	100.00	68.00	88.00	96.00	68.80	44.56	88.54	33.19	70.46	87.39	50.72
	300	28.00	80.00	60.00	100.00	72.00	88.00	80.00	65.96	33.56	98.14	32.64	69.18	87.09	51.31
SPA-B	100	20.00	84.00	84.00	100.00	88.00	100.00	100.00	80.50	45.08	91.38	48.90	71.16	86.04	59.03
	200	28.00	80.00	68.00	100.00	84.00	100.00	84.00	79.71	46.65	85.75	40.84	71.01	86.16	60.05
	300	24.00	80.00	88.00	100.00	92.00	100.00	92.00	74.70	48.97	81.60	34.92	71.16	85.16	61.94

Table 13: **All results on Franka Kitchen.**

Task	View	Seed	MoCoV3	MAE	DINOV2	CLIP	EVA	InternViT-300M	MVP	VC-1	SPA
Task 1	Left	100	86.00	76.00	84.00	72.00	78.00	74.00	66.00	74.00	84.00
		200	78.00	78.00	74.00	72.00	76.00	72.00	58.00	74.00	92.00
		300	80.00	80.00	78.00	62.00	82.00	70.00	64.00	74.00	80.00
	Right	100	82.00	80.00	86.00	78.00	78.00	72.00	82.00	78.00	86.00
		200	88.00	62.00	90.00	70.00	86.00	86.00	86.00	84.00	72.00
		300	86.00	82.00	92.00	82.00	86.00	76.00	92.00	78.00	86.00
Task 2	Left	100	60.00	56.00	48.00	26.00	40.00	22.00	40.00	32.00	48.00
		200	64.00	60.00	46.00	44.00	40.00	32.00	32.00	42.00	60.00
		300	58.00	54.00	40.00	26.00	32.00	34.00	30.00	50.00	66.00
	Right	100	62.00	54.00	56.00	26.00	44.00	26.00	32.00	54.00	48.00
		200	64.00	54.00	60.00	36.00	40.00	24.00	28.00	56.00	42.00
		300	64.00	52.00	50.00	38.00	40.00	30.00	34.00	44.00	42.00
Task 3	Left	100	16.00	24.00	18.00	18.00	22.00	24.00	6.00	24.00	28.00
		200	28.00	20.00	14.00	18.00	20.00	16.00	6.00	30.00	38.00
		300	22.00	16.00	14.00	10.00	26.00	22.00	8.00	26.00	30.00
	Right	100	46.00	26.00	38.00	22.00	14.00	8.00	32.00	12.00	10.00
		200	48.00	22.00	38.00	24.00	18.00	4.00	32.00	12.00	12.00
		300	54.00	34.00	52.00	14.00	12.00	6.00	26.00	14.00	16.00
Task 4	Left	100	32.00	36.00	26.00	22.00	34.00	12.00	16.00	36.00	22.00
		200	30.00	30.00	32.00	14.00	20.00	8.00	14.00	24.00	10.00
		300	24.00	46.00	28.00	14.00	32.00	4.00	20.00	36.00	16.00
	Right	100	38.00	24.00	28.00	22.00	32.00	12.00	26.00	12.00	30.00
		200	42.00	24.00	24.00	24.00	24.00	12.00	30.00	8.00	38.00
		300	46.00	16.00	32.00	28.00	36.00	16.00	26.00	12.00	30.00
Task 5	Left	100	36.00	18.00	8.00	16.00	24.00	22.00	26.00	28.00	20.00
		200	30.00	24.00	8.00	10.00	24.00	16.00	20.00	22.00	16.00
		300	22.00	22.00	10.00	12.00	16.00	14.00	28.00	30.00	18.00
	Right	100	24.00	46.00	20.00	4.00	14.00	10.00	26.00	22.00	30.00
		200	24.00	30.00	16.00	8.00	18.00	18.00	22.00	22.00	26.00
		300	14.00	36.00	16.00	12.00	10.00	12.00	20.00	16.00	22.00

Table 14: **All results on Meta-World.**

Method	MoCoV3			MAE			DINOV2			CLIP			EVA			InternViT-300M			MVP			VC-1			SPA		
Seed	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300
ButtonPressWall	100	100	100	100	100	100	95	90	95	100	100	100	95	95	100	95	100	95	100	100	100	100	100	100	100	100	100
DoorClose	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
DoorUnlock	70	65	80	70	65	85	35	35	30	60	50	60	85	85	90	80	65	75	75	70	85	80	75	90	80	75	80
DrawerClose	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
DrawerOpen	80	70	85	85	65	75	60	40	60	90	80	80	60	55	75	70	75	75	95	75	85	85	85	80	90	60	75
FaucetClose	70	80	55	60	80	60	55	65	35	70	80	50	65	75	60	50	65	50	70	75	60	65	75	100	70	80	65
PlateSlide	90	95	100	95	100	100	65	70	80	85	95	80	100	100	100	85	95	90	100	100	100	100	95	100	95	95	95
PlateSlideBack	80	65	85	85	65	85	90	75	90	85	75	90	80	65	90	85	80	90	85	70	90	80	70	90	80	70	85
PlateSlideSide	85	90	95	95	90	95	90	85	85	80	85	95	100	95	100	90	95	90	80	90	95	100	95	90	90	90	95
WindowClose	100	100	100	100	100	100	70	90	100	100	100	95	95	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Basketball	85	95	95	95	100	100	70	55	65	85	85	70	95	85	80	90	80	95	95	100	95	90	100	95	95	100	100
BinPicking	30	45	40	20	10	35	10	15	10	30	30	25	20	5	45	10	10	10	25	20	15	30	30	30	40	25	30
BoxClose	80	80	80	75	80	70	35	45	30	80	70	60	55	70	55	60	65	40	80	80	65	80	95	60	80	80	65
CoffeePush	45	50	40	40	55	30	30	25	15	45	40	55	45	35	40	45	30	25	25	45	25	30	45	55	35	40	30
Assembly	70	60	55	55	65	45	30	35	25	45	55	50	45	50	40	30	25	30	60	60	45	60	60	50	50	55	50
Disassemble	40	55	50	30	45	45	30	20	45	55	50	60	30	50	50	40	45	50	40	45	45	40	30	35	40	45	55
PushWall	25	35	30	20	30	40	40	30	45	35	30	35	25	35	30	15	15	25	30	35	35	55	55	60	30	40	45
ShelfPlace	35	35	20	25	45	20	30	30	35	30	35	30	25	20	15	15	15	15	25	25	15	25	15	15	15	35	15
DoorOpen	95	90	90	100	95	95	95	80	95	85	75	90	80	90	100	50	55	60	80	75	95	95	95	95	85	100	95
ButtonPress	75	85	85	85	100	100	80	95	85	55	70	75	80	90	90	85	90	80	85	90	95	80	90	85	100	95	100
SweepInto	45	45	40	55	55	45	50	50	40	45	45	40	45	25	30	35	25	30	45	50	40	55	50	50	50	50	45
DoorLock	100	85	85	90	100	100	95	90	85	85	85	75	95	100	95	80	90	95	85	100	95	100	100	90	80	95	85
ReachWall	70	70	80	75	85	70	85	80	85	90	90	80	65	75	85	60	55	55	75	65	75	75	80	80	75	80	70
Hammer	25	45	30	30	30	30	40	45	35	25	35	25	30	35	20	20	20	20	30	35	30	45	40	55	30	40	30
StickPush	95	95	100	80	90	95	90	90	90	75	85	85	90	85	100	60	80	85	90	95	90	85	85	95	90	90	85
ButtonPressTopdown	80	80	75	80	85	80	80	90	90	80	90	70	55	65	55	45	55	55	80	85	80	75	80	70	80	85	80
HandlePressSide	100	100	100	100	100	100	95	100	90	80	100	90	100	100	95	85	100	90	95	100	100	75	100	80	90	100	100
PlateSlideBackSide	100	100	100	100	95	95	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Sweep	50	80	70	35	60	60	65	85	95	60	85	75	35	50	55	15	60	35	35	70	65	35	65	65	30	65	55
ButtonPressTopdownWall	45	70	80	45	75	75	70	75	75	45	60	70	30	45	65	20	55	45	30	60	70	50	85	80	45	65	75
HandlePress	85	95	95	80	100	75	75	100	80	90	100	90	85	100	85	65	90	75	80	95	75	85	100	90	85	100	80
Push	25	30	30	25	30	30	30	25	40	25	15	30	30	25	20	25	20	20	25	20	25	40	30	25	30	15	35
CoffeePull	55	55	55	40	45	40	20	30	20	50	70	40	40	45	45	40	40	25	55	55	40	55	55	60	55	55	55
DialTurn	80	65	80	85	75	75	40	30	35	80	95	90	65	65	80	70	70	55	70	65	75	80	95	75	85	85	75
Reach	90	75	80	90	75	80	70	75	85	95	95	100	85	80	75	95	80	90	80	70	75	70	80	80	85	70	85
CoffeeButton	85	95	75	100	100	95	85	80	60	95	100	85	90	100	95	90	85	85	100	100	90	90	100	80	100	100	100
PickPlaceWall	45	35	65	40	25	45	15	10	20	35	40	50	20	25	35	30	25	25	25	35	40	35	20	45	40	35	55
StickPull	35	35	25	15	40	20	25	10	5	45	40	45	25	35	15	25	25	15	15	30	25	30	30	30	25	35	30
HandInsert	35	30	30	30	25	25	20	20	20	45	45	40	20	25	20	25	30	25	40	40	35	40	30	40	40	50	40
PegInsertSide	40	35	40	50	35	45	25	15	10	45	30	20	45	25	30	30	20	25	45	45	30	50	25	35	55	45	60
PickPlace	35	30	30	25	45	15	15	10	10	25	15	30	25	30	25	30	10	30	20	35	25	25	20	20	25	30	40
FaucetOpen	95	95	100	100	100	100	80	80	100	100	95	100	95	95	100	80	85	95	100	100	100	100	100	100	95	100	100
PushBack	65	70	60	40	55	40	15	15	25	30	45	25	35	35	45	15	35	25	40	45	45	45	55	25	35	55	50
LeverPull	70	80	80	80	70	75	15	30	35	55	85	80	70	65	70	60	55	55	65	80	65	65	80	70	85	70	80
HandlePull	85	85	80	85	80	85	45	55	40	75	75	80	80	60	65	45	70	65	80	90	80	70	85	75	100	90	90
Soccer	25	40	35	50	30	25	15	10	15	45	40	30	35	35	25	20	20	30	30	30	35	20	20	20	25	50	25
WindowOpen	65	80	80	55	80	85	60	50	65	50	65	60	65	80	75	55	75	75	60	70	70	60	70	75	55	85	65
PickOutOfHole	65	75	80	65	65	60	60	55	60	70	70	50	60	55	50	60	55	55	65	55	60	70	75	60	65	60	75

Table 15: **All results on RLBench.**

Method	MoCoV3	MAE	DINOV2	CLIP	EVA	InternViT 300M	MVP	VC-1	SPA
Group 1									
basketball in hoop	100	100	100	100	100	100	100	100	100
put rubbish in bin	100	100	96	96	96	100	96	100	100
meat off grill	100	100	100	100	100	100	100	100	100
meat on grill	80	76	76	68	80	72	68	76	80
slide block to target	0	84	96	24	4	0	100	100	4
reach and drag	100	96	88	100	96	100	96	100	100
take frame off hanger	88	88	92	88	84	84	88	88	96
water plants	64	60	28	64	60	44	52	60	68
hang frame on hanger	8	4	0	4	8	8	12	4	4
wipe desk	0	0	0	0	0	0	0	0	0
stack blocks	60	72	72	68	56	60	84	68	68
reach target	60	96	88	100	96	80	92	96	92
push button	100	100	100	100	100	100	100	100	100
lamp on	88	68	84	88	52	80	28	88	64
toilet seat down	100	100	100	100	100	100	96	96	100
close laptop lid	96	96	96	96	84	80	80	96	100
open box	12	12	20	4	16	4	0	12	16
open drawer	88	96	92	100	88	88	92	96	96
pick up cup	92	92	88	96	96	88	96	96	96
turn tap	88	84	84	96	88	92	96	100	100
take usb out of computer	100	100	100	100	100	100	100	88	100
play jenga	96	96	96	100	96	100	96	96	96
insert onto square peg	28	84	80	44	88	40	64	92	84
take umbrella out of umbrella stand	92	100	100	92	100	96	100	100	100
insert usb in computer	12	20	20	24	24	20	16	8	68
straighten rope	56	44	72	80	48	72	52	60	84
turn oven on	96	96	96	96	96	96	100	100	100
change clock	64	68	48	68	64	72	64	60	68
close microwave	100	100	100	100	100	100	100	100	100
close fridge	80	92	92	88	92	96	88	92	100
close grill	96	96	96	96	96	96	100	100	96
open grill	100	100	100	100	100	100	96	100	100
unplug charger	44	32	48	36	48	40	40	44	44
press switch	92	92	88	72	76	84	76	88	92
take money out safe	100	96	100	100	100	100	100	100	100
Group 2									
change channel	0	8	4	0	0	4	0	0	4
tv on	4	8	0	4	4	8	4	4	8
push buttons	12	4	4	0	0	0	0	12	4
stack wine	12	16	40	4	12	0	28	8	28
scoop with spatula	0	0	0	0	0	0	0	0	0
place hanger on rack	0	0	0	0	0	0	0	0	0
move hanger	0	0	0	0	0	0	0	0	0
sweep to dustpan	92	96	96	96	92	100	100	88	96
take plate off colored dish rack	96	100	96	92	84	96	88	92	96
screw nail	52	36	36	36	36	52	32	32	48
take shoes out of box	20	28	24	36	40	12	32	36	36
slide cabinet open and place cups	0	0	0	0	0	4	0	0	4
lamp off	100	96	96	100	96	96	100	100	100
pick and lift	88	96	92	96	92	80	96	96	96
take lid off saucepan	100	100	100	100	100	100	100	100	100
close drawer	100	100	100	100	96	100	100	100	100
close box	92	92	96	96	100	96	100	96	100
phone on base	100	100	100	100	100	96	100	100	100
toilet seat up	80	88	100	88	88	80	88	92	96
put books on bookshelf	12	24	24	28	28	20	20	28	16
beat the buzz	88	92	96	88	92	84	88	88	100
stack cups	40	56	52	52	48	56	64	68	64
put knife on chopping board	72	76	68	72	80	88	80	76	80
place shape in shape sorter	20	36	32	28	36	20	44	36	56
take toilet roll off stand	100	92	76	96	92	88	84	92	96
put umbrella in umbrella stand	8	0	12	12	0	4	12	8	12
setup checkers	76	80	68	68	88	92	92	80	80
open window	96	96	100	100	96	100	96	100	100
open wine bottle	80	100	88	92	92	88	96	88	88
open microwave	100	100	88	96	100	80	96	100	100
put money in safe	96	100	88	92	100	96	100	100	100
open door	100	96	96	96	96	96	84	96	96
close door	32	68	56	60	80	20	24	20	60
open fridge	44	52	48	44	36	64	52	32	64
open oven	8	4	12	8	4	20	4	4	16
plug charger in power supply	32	36	32	24	44	36	24	32	60

Table 16: **All results on LIBERO-OBJECT, LIBERO-SPATIAL, LIBERO-GOAL, LIBERO-10.**

	MoCoV3			MAE			DINOV2			CLIP			EVA			InternViT-300M			InternViT-6B			MVP			VC-1			SPA		
Seed	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300
LIBERO-OBJECT																														
0	0.65	0.60	0.65	0.65	0.45	0.55	0.65	0.80	0.85	0.80	0.75	0.65	1.00	0.70	0.95	0.80	0.65	0.60	0.70	0.85	0.50	0.80	0.90	0.65	0.80	0.50	0.60	0.90	0.95	0.95
1	0.35	0.35	0.55	0.90	0.75	0.80	0.30	0.50	0.75	0.40	0.30	0.05	0.65	0.30	0.70	0.15	0.40	0.20	0.60	0.25	0.45	0.05	0.80	0.60	0.40	0.65	0.45	0.65	0.70	0.45
2	0.90	0.85	0.95	0.90	0.40	0.95	0.85	0.50	0.90	0.70	0.80	0.75	0.85	0.75	0.75	0.90	0.85	0.80	0.85	0.45	0.85	0.80	0.85	0.90	1.00	0.95	0.95	0.90	0.95	0.80
3	0.55	0.70	0.65	0.90	0.15	0.90	0.30	0.65	0.90	0.25	0.45	0.60	0.80	0.80	0.90	0.75	0.70	0.40	1.00	0.50	0.55	0.70	0.65	0.85	0.95	0.75	0.60	0.70	0.90	0.90
4	0.65	0.85	0.85	0.80	0.90	0.75	0.75	0.55	0.75	0.35	0.75	0.65	0.95	0.75	1.00	0.90	1.00	0.85	0.90	0.70	0.80	0.80	0.75	0.70	0.90	0.85	0.90	0.90	1.00	0.95
5	0.50	0.70	0.80	0.70	0.35	0.60	0.55	0.75	0.60	0.25	0.70	0.45	0.75	0.75	0.65	0.85	0.60	0.75	0.60	0.35	0.50	0.55	0.40	0.80	0.65	0.70	0.70	0.25	0.15	0.65
6	0.35	0.50	0.65	0.60	0.65	0.65	0.55	0.70	0.70	0.35	0.55	0.60	0.40	0.35	0.25	0.65	0.60	0.55	0.30	0.10	0.35	0.25	0.50	0.65	0.50	0.50	0.30	0.50	0.70	0.80
7	0.75	0.75	0.80	0.90	0.40	0.75	0.55	0.30	0.70	0.40	0.35	0.40	0.55	0.75	0.70	0.80	0.40	0.60	0.60	0.65	0.70	0.60	0.45	0.65	0.80	0.75	0.50	0.80	0.75	0.65
8	0.50	0.95	0.90	1.00	0.95	1.00	0.50	0.35	0.50	0.45	0.45	0.35	1.00	0.75	0.85	0.70	0.65	0.75	0.50	0.40	0.75	0.65	0.55	0.70	0.80	0.90	0.50	0.85	0.95	0.90
9	0.45	0.50	0.40	0.60	0.65	0.95	0.80	0.90	0.95	0.50	0.70	0.30	0.85	0.75	0.75	0.85	0.95	0.65	0.90	0.60	0.15	0.65	0.60	0.30	0.70	0.70	0.65	0.60	0.95	0.90
LIBERO-SPATIAL																														
0	0.35	0.55	0.45	0.45	0.40	0.70	0.65	0.50	0.60	0.25	0.20	0.35	0.70	0.75	0.65	0.55	0.65	0.55	0.55	0.50	0.30	0.75	0.75	0.60	0.35	0.55	0.60	0.45	0.50	0.35
1	0.65	0.70	0.70	0.80	0.80	0.50	0.55	0.30	0.35	0.75	0.75	0.70	0.55	0.70	0.25	0.35	0.50	0.50	1.00	1.00	0.90	0.60	0.40	0.60	0.45	0.65	0.80	0.65	0.65	0.85
2	0.55	0.50	0.50	0.35	0.60	0.40	0.20	0.05	0.55	0.10	0.00	0.40	0.70	0.80	0.50	0.70	0.75	0.60	0.75	0.60	0.20	0.85	0.55	0.75	0.45	0.45	0.70	0.50	0.50	0.40
3	0.50	0.70	0.75	0.55	0.60	0.75	0.80	0.70	0.95	0.15	0.40	0.30	0.85	0.90	0.85	0.35	0.50	0.40	0.40	0.30	0.15	0.95	0.55	0.60	0.50	0.70	0.65	0.55	0.85	0.60
4	0.15	0.15	0.20	0.55	0.70	0.80	0.50	0.05	0.45	0.35	0.30	0.20	0.45	0.55	0.40	0.35	0.25	0.40	0.25	0.15	0.15	0.60	0.50	0.70	0.60	0.60	0.80	0.70	0.70	0.50
5	0.45	0.10	0.10	0.65	0.40	0.30	0.30	0.20	0.35	0.55	0.45	0.45	0.65	0.50	0.45	0.40	0.30	0.70	0.55	0.60	0.60	0.55	0.30	0.25	0.05	0.05	0.15	0.35	0.35	0.30
6	0.30	0.35	0.45	0.55	0.25	0.95	0.40	0.30	0.40	0.20	0.25	0.10	0.75	0.70	0.85	0.45	0.55	0.55	0.40	0.35	0.05	0.45	0.75	0.65	0.60	0.70	0.35	0.35	0.45	0.30
7	0.10	0.20	0.25	0.50	0.35	0.45	0.05	0.00	0.10	0.30	0.15	0.25	0.30	0.30	0.20	0.30	0.60	0.65	0.15	0.05	0.05	0.60	0.65	0.60	0.10	0.20	0.40	0.40	0.15	0.45
8	0.55	0.70	0.50	0.35	0.65	0.70	0.40	0.15	0.55	0.30	0.55	0.30	0.85	0.70	0.60	0.30	0.55	0.60	0.65	0.40	0.35	0.70	0.40	0.55	0.70	0.60	0.70	0.75	0.70	0.40
9	0.55	0.05	0.10	0.85	0.75	0.50	0.20	0.05	0.25	0.20	0.20	0.20	0.55	0.50	0.30	0.35	0.50	0.30	0.45	0.40	0.35	0.45	0.45	0.30	0.50	0.55	0.65	0.35	0.50	0.45
LIBERO-GOAL																														
0	0.45	0.70	0.75	0.70	0.85	0.80	0.15	0.10	0.30	0.25	0.40	0.35	0.70	0.60	0.60	0.75	0.65	0.75	0.25	0.35	0.35	0.75	0.60	0.95	0.45	0.85	1.00	0.85	1.00	0.85
1	0.70	0.60	0.80	0.65	0.50	0.90	0.25	0.55	0.25	0.20	0.15	0.25	0.70	0.80	0.80	0.90	0.90	1.00	0.40	0.15	0.15	0.90	0.80	0.95	0.65	0.65	0.65	1.00	0.85	0.90
2	0.50	0.20	0.15	0.10	0.40	0.35	0.10	0.05	0.15	0.30	0.25	0.30	0.65	0.75	0.75	0.40	0.75	0.45	0.50	0.35	0.35	0.45	0.25	0.65	0.40	0.60	0.35	0.50	0.55	0.35
3	0.75	0.45	0.60	0.40	0.75	0.55	0.20	0.10	0.10	0.05	0.20	0.55	0.30	0.15	0.15	0.30	0.50	0.65	0.20	0.25	0.25	0.70	0.70	0.15	0.75	0.55	0.50	0.65	0.35	0.80
4	0.20	0.25	0.05	0.35	0.40	0.25	0.10	0.00	0.05	0.40	0.30	0.15	0.15	0.10	0.10	0.20	0.15	0.10	0.15	0.20	0.20	0.55	0.60	0.25	0.15	0.30	0.30	0.30	0.35	0.35
5	0.10	0.75	0.80	0.60	0.85	0.80	0.65	0.50	0.50	0.35	0.45	0.50	0.80	0.75	0.75	0.70	0.55	0.45	0.55	0.45	0.45	0.80	0.75	0.85	0.65	0.75	0.80	0.80	0.65	0.65
6	0.45	0.05	0.15	0.00	0.10	0.05	0.00	0.05	0.00	0.10	0.10	0.00	0.00	0.65	0.65	0.50	0.40	0.30	0.00	0.00	0.00	0.15	0.35	0.70	0.25	0.50	0.45	0.40	0.30	0.35
7	0.25	0.75	0.90	0.80	0.65	1.00	0.45	0.45	0.35	0.50	0.65	0.80	1.00	1.00	1.00	1.00	1.00	0.95	0.70	0.85	0.85	0.95	1.00	0.95	1.00	0.95	0.70	0.95	1.00	1.00
8	0.50	0.80	0.75	0.85	0.55	0.90	0.45	0.40	0.20	0.60	0.25	0.50	0.90	0.35	0.35	0.95	0.65	0.55	0.65	0.25	0.25	0.70	0.65	0.70	0.50	0.75	0.55	0.80	0.65	0.80
9	0.10	0.65	0.60	0.50	0.20	0.50	0.10	0.00	0.10	0.10	0.10	0.00	0.15	0.70	0.70	0.60	0.40	0.20	0.15	0.35	0.35	0.30	0.50	0.55	0.20	0.35	0.70	0.55	0.60	0.45
LIBERO-10																														
0	0.15	0.20	0.10	0.15	0.25	0.10	0.00	0.05	0.10	0.00	0.05	0.05	0.25	0.35	0.10	0.35	0.10	0.25	0.15	0.15	0.00	0.05	0.15	0.20	0.10	0.45	0.25	0.05	0.10	0.05
1	0.25	0.20	0.20	0.30	0.15	0.25	0.15	0.15	0.15	0.40	0.30	0.15	0.65	0.10	0.60	0.15	0.50	0.45	0.00	0.25	0.35	0.15	0.10	0.15	0.20	0.40	0.15	0.25	0.05	0.45
2	0.70	0.60	0.75	0.30	0.60	0.75	0.55	0.45	0.50	0.25	0.45	0.40	0.75	0.55	0.65	0.45	0.80	0.55	0.70	0.80	0.75	0.75	0.65	0.55	0.85	1.00	0.90	0.70	0.80	0.50
3	0.50	0.80	0.55	0.55	0.60	0.80	0.40	0.45	0.45	0.60	0.65	0.60	0.80	0.90	0.75	0.75	0.65	0.50	0.75	0.60	0.60	0.75	0.70	0.65	0.80	0.70	0.70	0.70	0.90	0.70
4	0.25	0.20	0.05	0.35	0.25	0.30	0.10	0.10	0.05	0.20	0.05	0.05	0.15	0.10	0.15	0.15	0.15	0.05	0.15	0.20	0.15	0.25	0.20	0.35	0.30	0.25	0.30	0.40	0.30	0.25
5	0.40	0.60	0.75	0.55	0.70	0.80	0.50	0.65	0.75	0.40	0.40	0.30	0.85	0.65	0.75	0.75	0.55	0.75	0.45	0.55	0.45	0.60	0.70	0.75	0.80	0.90	0.60	0.70	0.70	0.45
6	0.20	0.25	0.10	0.40	0.35	0.40	0.05	0.20	0.05	0.25	0.15	0.15	0.20	0.30	0.35	0.10	0.15	0.20	0.20	0.25	0.05	0.40	0.30	0.35	0.30	0.10	0.20	0.20	0.20	0.15
7	0.40	0.30	0.25	0.50	0.50	0.50	0.10	0.30	0.60	0.30	0.40	0.20	0.45	0.30	0.25	0.40	0.35	0.35	0.35	0.70	0.25	0.35	0.30	0.25	0.30	0.25	0.30	0.50	0.45	0.40
8	0.10	0.10	0.15	0.10	0.30	0.20	0.35	0.20	0.05	0.10	0.10	0.10	0.15	0.20	0.10	0.20	0.00	0.25	0.05	0.20	0.05	0.10	0.30	0.20	0.25	0.30	0.25	0.25	0.05	0.15
9	0.20	0.60	0.35	0.40	0.65	0.30	0.35	0.25	0.45	0.45	0.45	0.30	0.40	0.70	0.50	0.50	0.45	0.60	0.50	0.25	0.40	0.40	0.55	0.50	0.00	0.00	0.00	0.50	0.65	0.50

Table 17: **All results on LIBERO-90.**

	MoCoV3			MAE			DINOV2			CLIP			EVA			InternViT-300M			InternViT-6B			MVP			VC-1			SPA		
Seed	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300	100	200	300
LIBERO-90																														
0	0.95	0.85	0.90	1.00	0.90	0.80	0.80	1.00	0.60	0.90	0.80	0.80	1.00	1.00	1.00	0.90	0.80	0.85	0.75	0.80	0.95	0.95	0.95	1.00	0.95	1.00	0.95	1.00	1.00	0.95
1	0.60	0.35	0.60	0.35	0.50	0.15	0.50	0.55	0.30	0.40	0.65	0.35	0.70	0.50	0.25	0.30	0.45	0.40	0.25	0.35	0.55	0.80	0.55	0.30	0.40	0.50	0.05	0.65	0.40	0.50
2	0.85	0.50	0.80	0.55	0.55	0.20	0.65	0.60	0.30	0.45	0.30	0.50	0.35	0.50	0.70	0.85	0.65	0.80	0.25	0.35	0.30	0.45	0.70	0.70	0.75	0.55	0.35	0.70	0.85	0.60
3	0.10	0.10	0.00	0.05	0.00	0.00	0.05	0.00	0.00	0.15	0.00	0.00	0.00	0.10	0.15	0.00	0.05	0.05	0.00	0.00	0.05	0.10	0.00	0.05	0.05	0.10	0.00	0.10	0.10	0.00
4	0.40	0.05	0.20	0.30	0.25	0.30	0.15	0.40	0.55	0.40	0.40	0.35	0.10	0.25	0.15	0.40	0.05	0.25	0.20	0.45	0.40	0.30	0.40	0.15	0.25	0.05	0.15	0.15	0.15	0.35
5	0.05	0.05	0.05	0.10	0.05	0.20	0.00	0.20	0.00	0.05	0.05	0.20	0.25	0.25	0.10	0.10	0.05	0.30	0.20	0.10	0.25	0.05	0.15	0.10	0.10	0.05	0.05	0.35	0.00	0.15
6	0.10	0.00	0.00	0.00	0.00	0.05	0.05	0.05	0.10	0.05	0.10	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.05	0.10	0.10	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.05	0.05
7	0.35	0.30	0.65	0.20	0.60	0.30	0.35	0.25	0.40	0.50	0.60	0.35	0.60	0.10	0.20	0.15	0.25	0.10	0.40	0.25	0.45	0.50	0.20	0.40	0.20	0.30	0.25	0.30	0.30	0.65
8	0.10	0.15	0.00	0.05	0.20	0.10	0.15	0.25	0.10	0.20	0.10	0.10	0.10	0.05	0.00	0.05	0.00	0.10	0.20	0.15	0.20	0.10	0.00	0.20	0.05	0.15	0.20	0.05	0.05	0.15
9	0.30	0.25	0.35	0.50	0.25	0.30	0.35	0.60	0.70	0.25	0.20	0.50	0.25	0.10	0.50	0.10	0.10	0.25	0.60	0.25	0.30	0.25	0.15	0.45	0.25	0.05	0.35	0.25	0.20	0.25
10	0.50	0.75	0.50	0.50	0.60	0.55	0.65	0.60	0.60	0.90	0.45	0.55	0.40	0.85	0.35	0.35	0.05	0.25	0.45	0.45	0.65	0.40	0.50	0.55	0.45	0.75	0.40	0.40	0.35	0.35
11	0.45	0.35	0.75	0.45	0.70	0.65	0.35	0.20	0.15	0.40	0.70	0.55	0.80	0.25	0.70	0.50	0.50	0.10	0.35	0.25	0.45	0.80	0.60	0.95	0.70	0.75	0.60	0.60	0.60	0.65
12	0.15	0.15	0.10	0.15	0.15	0.05	0.20	0.20	0.15	0.10	0.05	0.05	0.10	0.25	0.05	0.05	0.00	0.00	0.25	0.30	0.10	0.15	0.10	0.10	0.20	0.25	0.10	0.05	0.10	0.15
13	0.20	0.35	0.30	0.15	0.30	0.20	0.30	0.35	0.10	0.30	0.40	0.35	0.30	0.10	0.45	0.20	0.35	0.40	0.25	0.15	0.55	0.30	0.30	0.15	0.45	0.10	0.10	0.10	0.20	0.10
14	0.05	0.10	0.00	0.30	0.30	0.20	0.10	0.10	0.15	0.15	0.40	0.20	0.25	0.35	0.10	0.15	0.15	0.05	0.20	0.15	0.10	0.20	0.35	0.10	0.20	0.10	0.20	0.15	0.15	0.10
15	0.60	0.75	0.45	0.70	0.50	0.65	0.35	0.50	0.55	0.45	0.65	0.40	0.70	0.75	0.40	0.40	0.65	0.50	0.35	0.55	0.45	0.70	0.60	0.55	0.80	0.80	0.70	0.65	0.80	0.55
16	0.05	0.20	0.00	0.30	0.15	0.05	0.10	0.10	0.05	0.10	0.00	0.10	0.20	0.20	0.15	0.15	0.15	0.20	0.05	0.00	0.10	0.15	0.05	0.10	0.00	0.15	0.15	0.05	0.10	0.15
17	0.05	0.15	0.15	0.10	0.25	0.05	0.05	0.10	0.05	0.05	0.00	0.05	0.05	0.20	0.15	0.10	0.10	0.15	0.00	0.10	0.00	0.20	0.10	0.20	0.15	0.10	0.00	0.25	0.10	0.10
18	0.45	0.40	0.60	0.40	0.75	0.65	0.30	0.35	0.40	0.45	0.25	0.35	0.25	0.35	0.60	0.40	0.05	0.70	0.60	0.50	0.35	0.35	0.25	0.45	0.30	0.60	0.35	0.60	0.35	0.55
19	0.30	0.30	0.25	0.35	0.40	0.20	0.20	0.05	0.35	0.45	0.45	0.30	0.30	0.35	0.25	0.15	0.25	0.20	0.35	0.30	0.15	0.55	0.30	0.40	0.40	0.45	0.35	0.45	0.20	0.35
20	0.85	0.75	0.80	1.00	1.00	0.95	0.95	1.00	1.00	0.75	0.85	0.30	1.00	1.00	1.00	0.95	0.95	0.90	1.00	0.50	0.80	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
21	0.40	0.20	0.40	0.35	0.25	0.30	0.25	0.40	0.20	0.25	0.10	0.45	0.30	0.70	0.05	0.00	0.05	0.10	0.35	0.10	0.30	0.40	0.15	0.30	0.30	0.70	0.60	0.65	0.40	0.60
22	0.90	0.95	0.95	1.00	0.85	0.95	0.25	0.60	0.40	0.75	0.75	0.75	0.95	1.00	0.95	0.85	0.95	0.60	0.45	0.25	0.25	0.90	1.00	1.00	0.90	0.90	0.95	1.00	0.95	1.00
23	0.15	0.05	0.15	0.05	0.10	0.00	0.05	0.10	0.05	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.25	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
24	0.80	0.30	0.85	0.85	0.50	0.80	0.60	0.50	0.65	0.70	0.45	0.60	0.70	0.70	0.80	0.40	0.65	0.60	0.55	0.80	0.45	0.65	0.60	0.90	0.90	0.80	0.80	0.90	0.80	0.75
25	1.00	0.80	0.85	1.00	1.00	0.90	0.75	0.90	0.90	0.80	0.95	0.90	0.90	1.00	0.95	0.70	0.70	0.85	0.95	0.60	0.65	1.00	0.85	1.00	1.00	1.00	0.90	1.00	1.00	1.00
26	0.15	0.20	0.25	0.25	0.40	0.40	0.05	0.30	0.40	0.45	0.05	0.15	0.05	0.30	0.15	0.25	0.40	0.20	0.25	0.15	0.20	0.20	0.30	0.60	0.25	0.45	0.25	0.25	0.20	0.20
27	0.30	0.15	0.20	0.35	0.35	0.10	0.05	0.10	0.00	0.35	0.05	0.10	0.05	0.20	0.05	0.05	0.15	0.00	0.10	0.10	0.05	0.35	0.20	0.30	0.10	0.45	0.40	0.10	0.40	0.20
28	0.90	0.90	1.00	0.95	0.70	0.70	0.80	0.50	0.80	0.90	0.75	0.90	0.95	1.00	0.85	0.90	0.85	1.00	0.50	0.60	0.45	0.85	0.75	0.90	0.75	0.95	0.60	0.90	0.65	0.90
29	0.15	0.50	0.35	0.60	0.55	0.20	0.50	0.50	0.40	0.50	0.50	0.30	0.65	0.30	0.30	0.15	0.30	0.40	0.25	0.35	0.25	0.35	0.50	0.25	0.60	0.60	0.10	0.30	0.35	0.70
30	0.15	0.25	0.15	0.60	0.35	0.35	0.35	0.10	0.50	0.25	0.20	0.45	0.30	0.70	0.20	0.10	0.15	0.20	0.50	0.20	0.25	0.00	0.05	0.20	0.10	0.25	0.10	0.40	0.25	0.40
31	0.70	0.60	0.80	0.70	0.75	0.60	0.45	0.70	0.75	0.95	0.65	0.95	0.80	0.75	0.45	0.50	0.55	0.35	0.95	0.80	0.85	0.80	0.70	0.75	0.75	0.75	0.45	0.70	0.90	0.80
32	0.30	0.05	0.20	0.05	0.10	0.05	0.00	0.00	0.05	0.35	0.10	0.10	0.20	0.10	0.15	0.10	0.15	0.10	0.05	0.15	0.05	0.20	0.25	0.10	0.05	0.10	0.05	0.20	0.10	0.05
33	0.50	0.55	0.40	0.15	0.30	0.30	0.10	0.20	0.35	0.30	0.25	0.30	0.00	0.50	0.40	0.35	0.20	0.25	0.25	0.25	0.30	0.20	0.35	0.40	0.35	0.45	0.65	0.15	0.15	0.15
34	0.30	0.35	0.30	0.40	0.40	0.25	0.35	0.40	0.15	0.40	0.40	0.50	0.10	0.40	0.10	0.15	0.05	0.25	0.35	0.30	0.50	0.25	0.30	0.30	0.55	0.25	0.05	0.05	0.30	0.10
35	0.65	0.40	0.60	0.85	0.95	0.75	0.80	0.80	0.60	0.65	0.55	0.90	0.90	1.00	0.80	0.10	0.20	0.55	0.85	0.75	0.70	0.85	0.80	0.85	0.25	0.80	0.55	1.00	0.70	1.00
36	0.05	0.10	0.15	0.05	0.00	0.00	0.00	0.25	0.05	0.05	0.05	0.00	0.15	0.20	0.10	0.15	0.10	0.25	0.00	0.00	0.10	0.20	0.00	0.15	0.30	0.05	0.20	0.05	0.20	0.10
37	0.35	0.30	0.40	0.55	0.20	0.25	0.65	0.50	0.35	0.30	0.60	0.60	0.70	0.80	0.60	0.45	0.55	0.45	0.50	0.55	0.55	0.45	0.45	0.50	0.45	0.75	0.35	0.75	0.50	0.55
38	0.50	0.30	0.45	0.55	0.50	0.35	0.25	0.15	0.30	0.50	0.45	0.30	0.50	0.20	0.35	0.35	0.55	0.45	0.35	0.50	0.35	0.70	0.40	0.55	0.70	0.65	0.30	0.70	0.45	0.45
39	0.80	0.80	0.75	0.45	0.70	0.60	0.60	0.55	0.65	0.60	0.65	0.70	0.65	0.85	0.55	0.60	0.15	0.60	0.65	0.60	0.60	0.45	0.40	0.70	0.30	0.65	0.35	0.60	0.55	0.60
40	0.40	0.50	0.20	0.40	0.30	0.40	0.30	0.45	0.55	0.50	0.25	0.30	0.70	0.65	0.30	0.25	0.60	0.25	0.55	0.40	0.25	0.65	0.70	0.80	0.35	0.45	0.25	0.55	0.35	0.30
41	0.20	0.60	0.40	0.50	0.45	0.60	0.35	0.25	0.55	0.20	0.45	0.50	0.85	0.45	0.50	0.45	0.65	0.45	0.20	0.30	0.35	0.55	0.80	0.20	0.35	0.50	0.65	0.65	0.80	0.40
42	0.20	0.40	0.20	0.35	0.15	0.05	0.30	0.30	0.55	0.35	0.05	0.20	0.70	0.45	0.20	0.40	0.65	0.60	0.25	0.10	0.05	0.30	0.50	0.65	0.35	0.60	0.50	0.60	0.60	0.40
43	0.50	0.50	0.50	0.35	0.40	0.30	0.35	0.35	0.15	0.35	0.25	0.45	0.20	0.60	0.50	0.35	0.35	0.05	0.30	0.30	0.15	0.30	0.60	0.40	0.55	0.40	0.20	0.55	0.70	0.40
44	0.90	0.90	0.95	1.00	1.00	0.85	0.80	0.90	0.95	0.75	0.85	0.85	0.95	0.80	1.00	1.00	0.95	0.85	0.85	0.65	0.95	0.85	0.95	0.90	0.95	0.95	0.95	0.95	1.00	0.95
45	0.55	0.40	0.60	0.80	0.75	0.60	0.45	0.55	0.60	0.70	0.70	0.55	0.45	0.40	0.30	0.55	0.15	0.55	0.40	0.65	0.30	0.65	0.50	0.35	0.55	0.65	0.45	0.60	0.75	0.60
46	0.00	0.00	0.00	0.05	0.35	0.10	0.00	0.15	0.00	0.10	0.10	0.05	0.05	0.10	0.00	0.00	0.00	0.10	0.10	0.00	0.05	0.05	0.05	0.20	0.00	0.05	0.10	0.10	0.00	0.10
47	0.45	0.35	0.25	0.25	0.20	0.20	0.10	0.20	0.15	0.60	0.25	0.45	0.45	0.30	0.30	0.15	0.30	0.30	0.15	0.35	0.20	0.40	0.40	0.30	0.20	0.25	0.10	0.55	0.45	0.30
48	0.00	0.15	0.30	0.10	0.00	0.15	0.20	0.05	0.20	0.10	0.10	0.25	0.05	0.10	0.25	0.15	0.00	0.25	0.20	0.10	0.20	0.05	0.10	0.10	0.15	0.20	0.10	0.10	0.05	0.20
49	0.05	0.05	0.25	0.25	0.30	0.15	0.10	0.20	0.25	0.25	0.40	0.30	0.15	0.25	0.05	0.05	0.00	0.15	0.25	0.30	0.00	0.00	0.00	0.15	0.15	0.05	0.25	0.15	0.05	0.10
50	0.25	0.05	0.25	0.00	0.05	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.10	0.05	0.05	0.00	0.00	0.05	0.00	0.05	0.05	0.05	0.10	0.00	0.00	0.05	0.15	0.05
51	0.05	0.15	0.05	0.10	0.15	0.05	0.15	0.15	0.20	0.15	0.20	0.15	0.15	0.10	0.05	0.05	0.20	0.05	0.35	0.15	0.35	0.15	0.20	0.30	0.15	0.00	0.10	0.05	0.10	0.10
52	0.00	0.00	0.00	0.00	0.00	0.05	0.05	0.10	0.00	0.00	0.05	0.05	0.10	0.00	0.15	0.10	0.00	0.00	0.05	0.05	0.00	0.05	0.15	0.10	0.00	0.00	0.10	0.15	0.00	0.05
53	0.05	0.05	0.10	0.15	0.00	0.05	0.20	0.20	0.05	0.00	0.10	0.20	0.05	0.25	0.05	0.05	0.00	0.00	0.25	0.15	0.25	0.00	0.00	0.00	0.05	0.15	0.10	0.10	0.05	0.15
54	0.05	0.20	0.05	0.05	0.05	0.00	0.05	0.05	0.05	0.25	0.05	0.05	0.05	0.05	0.10	0.20	0.10	0.00	0.20	0.20	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.05	0.15	0.05
55	0.10	0.05	0.05	0.00	0.00	0.00	0.10	0.05	0.00	0.05	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.05	0.05	0.05	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.10	0.00
56	0.05	0.15	0.10	0.10	0.10	0.05	0.15	0.15	0.30	0.20	0.10	0.25	0.10	0.10	0.05	0.10	0.20	0.00	0.25	0.15	0.10	0.05	0.00	0.05	0.20	0.15	0.10	0.10	0.15	0.20
57	0.10	0.00	0.05	0.10	0.10	0.05	0.05	0.20	0.00	0.15	0.10	0.05	0.10	0.15	0.00	0.05	0.05	0.10	0.20	0.25	0.10	0.00	0.00	0.20	0.10	0.00	0.20	0.20	0.05	0.05
58	0.05	0.10	0.20	0.00	0.05	0.10	0.10	0.25	0.10	0.25	0.20	0.15	0.20	0.15	0.15	0.05	0.05	0.15	0.05	0.15	0.05	0.15	0.10	0.10	0.05	0.10	0.20	0.15	0.05	0.10
59	0.05	0.10	0.10	0.05	0.10	0.10	0.05	0.15	0.00	0.05	0.10	0.05	0.05	0.10	0.25	0.05	0.15	0.25	0.05	0.10	0.05	0.10	0.00	0.10	0.15	0.15	0.00	0.00	0.20	0.05
60	0.55	0.65	0.70	0.30	0.60	0.45	0.50	0.50	0.40	0.40	0.35	0.65	0.20	0.75	0.55	0.35	0.35	0.50	0.40	0.40	0.50	0.45	0.50	0.75	0.20	0.50	0.55	0.45	0.65	0.90
61	0.05	0.15	0.05	0.00	0.10	0.00	0.00	0.10	0.05	0.05	0.10	0.10	0.05	0.10	0.05	0.00	0.00	0.10	0.15	0.05	0.20	0.10	0.15	0.20	0.05	0.15	0.05	0.15	0.05	0.15
62	0.45	0.40	0.40	0.45	0.20	0.20	0.25	0.30	0.25	0.55	0.15	0.70	0.40	0.50	0.45	0.30	0.10	0.50	0.40	0.15	0.05	0.20	0.15	0.35	0.25	0.35	0.50	0.40	0.10	0.35
63	0.10	0.00	0.05	0.15	0.05	0.00	0.15	0.20	0.15	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00
64	0.10	0.05	0.05	0.00	0.05	0.00	0.00	0.05	0.05	0.00	0.05	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.00
65	0.05	0.00	0.15	0.20	0.05	0.10	0.00	0.05	0.05	0.10	0.05	0.05	0.05	0.00	0.05	0.00	0.05	0.05	0.05	0.10	0.10	0.10	0.00	0.10	0.15	0.05	0.05	0.05	0.05	0.05
66	0.20	0.35	0.05	0.05	0.30	0.15	0.40	0.40	0.50	0.20	0.15	0.25	0.10	0.20	0.05	0.25	0.10	0.25	0.15	0.20	0.30	0.20	0.20	0.20	0.10	0.10	0.20	0.35	0.40	0.25
67	0.00	0.00	0.05	0.05	0.00	0.15	0.05	0.05	0.05	0.05	0.00	0.05	0.05	0.00	0.00	0.10	0.05	0.00	0.00	0.05	0.00	0.00	0.10	0.00	0.05	0.10	0.10	0.00	0.00	0.00
68	0.05	0.40	0.20	0.30	0.25	0.15	0.30	0.40	0.60	0.25	0.20	0.25	0.20	0.25	0.15	0.20	0.25	0.15	0.40	0.50	0.40	0.35	0.30	0.35	0.15	0.15	0.35	0.30	0.15	0.30
69	0.15	0.00	0.20	0.10	0.25	0.05	0.00	0.00	0.00	0.10	0.05	0.10	0.30	0.05	0.50	0.20	0.20	0.00	0.10	0.20	0.10	0.30	0.15	0.25	0.30	0.30	0.10	0.25	0.35	0.10
70	0.20	0.45	0.05	0.20	0.20	0.20	0.30	0.35	0.20	0.20	0.30	0.25	0.20	0.20	0.30	0.20	0.30	0.25	0.20	0.20	0.20	0.25	0.55	0.40	0.25	0.15	0.50	0.40	0.20	0.40
71	0.35	0.25	0.10	0.20	0.20	0.35	0.35	0.20	0.15	0.15	0.30	0.05	0.15	0.15	0.30	0.25	0.10	0.30	0.35	0.05	0.45	0.35	0.10	0.15	0.05	0.10	0.25	0.20	0.45	0.30
72	0.10	0.35	0.25	0.05	0.10	0.40	0.05	0.05	0.25	0.05	0.20	0.10	0.20	0.35	0.35	0.10	0.10	0.30	0.05	0.20	0.05	0.25	0.25	0.50	0.15	0.40	0.30	0.05	0.30	0.25
73	0.20	0.05	0.25	0.10	0.05	0.15	0.05	0.00	0.10	0.20	0.15	0.15	0.00	0.15	0.00	0.25	0.10	0.25	0.15	0.05	0.15	0.05	0.05	0.25	0.20	0.25	0.05	0.10	0.10	0.30
74	0.30	0.40	0.25	0.20	0.25	0.40	0.20	0.30	0.15	0.40	0.15	0.25	0.35	0.30	0.25	0.15	0.10	0.25	0.45	0.35	0.15	0.40	0.25	0.45	0.40	0.50	0.20	0.50	0.15	0.25
75	0.20	0.30	0.15	0.20	0.10	0.25	0.10	0.20	0.10	0.05	0.20	0.15	0.20	0.20	0.20	0.25	0.15	0.20	0.10	0.10	0.10	0.20	0.35	0.55	0.20	0.20	0.15	0.10	0.20	0.25
76	0.15	0.15	0.40	0.75	0.50	0.20	0.40	0.40	0.35	0.45	0.25	0.30	0.00	0.30	0.05	0.00	0.00	0.00	0.60	0.20	0.30	0.05	0.25	0.40	0.10	0.05	0.10	0.10	0.05	0.05
77	0.35	0.35	0.10	0.05	0.05	0.30	0.25	0.50	0.35	0.15	0.00	0.40	0.15	0.15	0.50	0.30	0.00	0.10	0.15	0.20	0.20	0.15	0.25	0.30	0.55	0.45	0.55	0.35	0.15	0.40
78	0.10	0.20	0.25	0.00	0.20	0.15	0.10	0.00	0.00	0.10	0.10	0.05	0.10	0.40	0.20	0.05	0.00	0.10	0.05	0.05	0.00	0.20	0.05	0.05	0.20	0.30	0.10	0.15	0.00	0.25
79	0.15	0.40	0.35	0.25	0.20	0.40	0.15	0.45	0.40	0.35	0.35	0.20	0.35	0.25	0.35	0.35	0.00	0.20	0.30	0.15	0.30	0.25	0.15	0.60	0.50	0.65	0.25	0.35	0.05	0.25
80	0.15	0.30	0.40	0.20	0.40	0.45	0.25	0.20	0.35	0.10	0.05	0.15	0.40	0.20	0.20	0.20	0.00	0.15	0.30	0.25	0.50	0.25	0.40	0.45	0.60	0.20	0.30	0.40	0.00	0.30
81	0.15	0.10	0.15	0.00	0.05	0.20	0.25	0.15	0.05	0.10	0.05	0.10	0.05	0.20	0.05	0.15	0.00	0.05	0.00	0.00	0.05	0.25	0.25	0.25	0.00	0.15	0.15	0.20	0.25	0.20
82	0.20	0.35	0.45	0.55	0.30	0.45	0.15	0.45	0.25	0.40	0.60	0.65	0.70	0.45	0.60	0.20	0.00	0.15	0.35	0.45	0.50	0.25	0.45	0.65	0.45	0.65	0.15	0.40	0.30	0.45
83	0.10	0.35	0.40	0.10	0.15	0.20	0.10	0.15	0.10	0.30	0.00	0.10	0.20	0.15	0.20	0.05	0.10	0.15	0.40	0.20	0.15	0.15	0.00	0.25	0.15	0.25	0.20	0.15	0.10	0.20
84	0.05	0.20	0.20	0.15	0.25	0.05	0.20	0.15	0.40	0.35	0.20	0.10	0.30	0.20	0.05	0.05	0.05	0.10	0.50	0.25	0.20	0.20	0.05	0.20	0.20	0.05	0.20	0.05	0.15	0.00
85	0.05	0.00	0.05	0.00	0.10	0.05	0.05	0.20	0.00	0.20	0.10	0.15	0.00	0.05	0.05	0.00	0.00	0.00	0.20	0.10	0.05	0.00	0.00	0.05	0.10	0.05	0.00	0.05	0.05	0.00
86	0.05	0.25	0.15	0.20	0.60	0.25	0.00	0.05	0.00	0.15	0.25	0.30	0.25	0.25	0.30	0.20	0.05	0.15	0.15	0.20	0.10	0.10	0.00	0.45	0.05	0.20	0.05	0.40	0.30	0.30
87	0.40	0.65	0.65	0.25	0.55	0.40	0.30	0.20	0.20	0.40	0.30	0.35	0.55	0.65	0.55	0.25	0.50	0.40	0.45	0.30	0.30	0.30	0.45	0.90	0.35	0.65	0.70	0.50	0.45	0.55
88	0.40	0.20	0.55	0.20	0.55	0.35	0.45	0.65	0.65	0.55	0.60	0.55	0.45	0.50	0.50	0.40	0.20	0.50	0.55	0.25	0.20	0.75	0.45	0.40	0.35	0.65	0.30	0.45	0.55	0.50
89	0.10	0.00	0.05	0.05	0.10	0.10	0.05	0.20	0.05	0.20	0.05	0.05	0.00	0.10	0.00	0.10	0.25	0.05	0.10	0.15	0.05	0.00	0.25	0.30	0.30	0.05	0.10	0.20	0.20	0.10

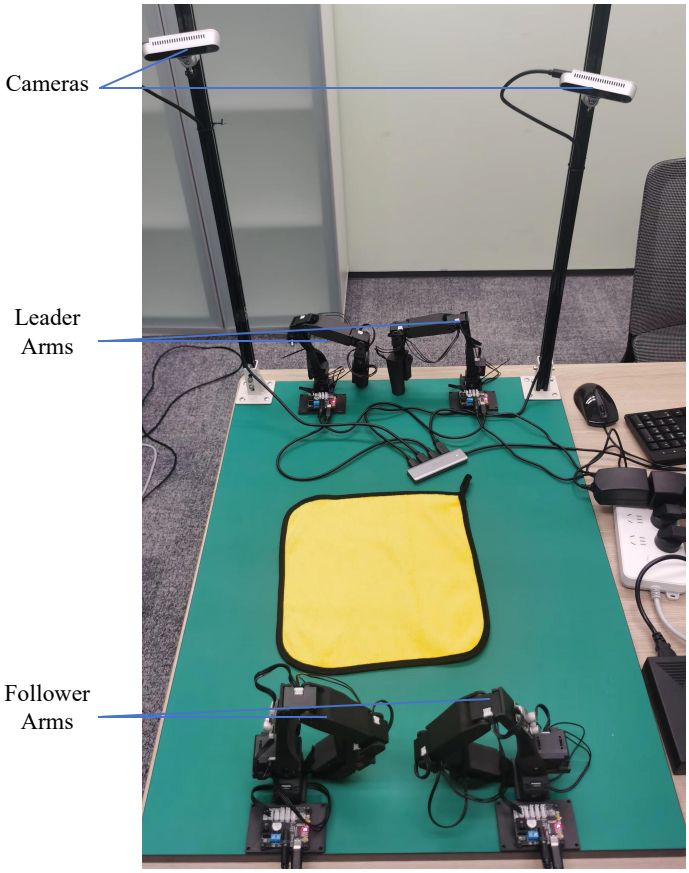

Figure 8: **Real-world hardware platform.**