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

CLAR: Learning 3D Representations for Robotic Manipulation by Fusing Masked Reconstruction with Multi-Level Contrastive Alignment

Wenbo Cui[*1,2,3], Chengyang Zhao[*4], Yuhui Chen[1,2], Haoran Li[1,2,3],
Zhizheng Zhang[3,5,6], Dongbin Zhao[1,2,3] He Wang[†3,5,6]

***Abstract*— The spatial information inherent in 3D point clouds is crucial for robotic manipulation. However, existing 3D pre-training methods face a fundamental trade-off: Masked Autoencoding (MAE) excels at capturing spatial-geometric features but lacks semantics, whereas contrastive learning, while able to distill semantics from 2D foundation models, is ill-suited for the fine-grained details required for manipulation tasks. To address these challenges, we propose *CLAR*, a novel 3D pre-training framework that synergizes global understanding with fine-grained local alignment. Our framework unifies MAE with global cross-modal contrastive learning to integrate robust spatial awareness with rich semantic understanding. To enhance its focus on fine-grained details, at the local level, we introduce an adaptive alignment mechanism that leverages deformable attention to force precise correspondences between local 3D geometry and 2D visual features, thereby overcoming the limitations of conventional global alignment in manipulation tasks. Extensive experiments in simulation and the real world demonstrate that *CLAR* achieves state-of-the-art performance, significantly outperforming existing methods in visuomotor policy learning. Our project page is https://cwb0106.github.io/CLAR/.**

I. INTRODUCTION

Building a robust representation module is essential for visuomotor policy learning, enabling robotic systems to perform various everyday tasks [1], [2], [3]. Many existing studies [4], [5], [6] highlight the advantages of integrating pre-trained 2D vision pre-training models into the perception module of robotic policies, leveraging them as powerful vision encoders. This approach allows policies to inherit the strong semantic understanding capabilities of foundation models while facilitating rapid adaptation to robotic tasks through fine-tuning on task-specific datasets.

Despite its benefits, 2D visual pre-training models still face fundamental bottlenecks in robotic manipulation. First, they lack 3D spatial awareness [4]. Their 2D backbones, engineered for planar images, are inherently ill-equipped to understand scene geometry, which prevents policies from effectively grounding themselves in the physical 3D world. Incorporating multi-view or depth information provides only marginal improvements. Second, the absence of a unified, robot-centric coordinate system creates severe multi-view ambiguity. This forces policies to learn brittle, camera-centric behaviors—for instance, an object appearing on the robot's "left" from one camera and its "right" from another (Fig. 1)—rather than forming a coherent understanding of spatial relationships. As a result, these learned skills do not generalize to different robot or camera setups, drastically limiting their autonomy and adaptability for real-world deployment.

To address the limitations of 2D foundation models in robotics tasks, recent works have shifted to using point clouds as observation for policy learning [7], [8]. While leveraging pre-training to circumvent the poor robustness of training from scratch [4] is a consensus, existing 3D pre-training methods face significant and distinct hurdles. On one hand, approaches that lift features from 2D models to 3D [4], [9] are constrained by the modality gap, which can cause information loss [4]. On the other hand, methods for pure 3D pre-training face their own inherent challenges. These methods exhibit a split between Masked Autoencoding (MAE) for spatial geometry [10] and contrastive learning for semantic understanding [11], yet a fusion of their strengths is crucial for robotic manipulation. Moreover, the application of contrastive learning to robotics is flawed: its global features loses local details, and point cloud cropping creates contextual mismatch that degrades representation quality. These combined challenges make it particularly difficult to design a pre-training framework that balances both geometry and semantics for robotics tasks.

In this paper, we introduce ***CLAR***, a novel pre-training method that enhances robotic representations by unifying **C**ontrastive **L**earning for semantic understanding, local **A**lignment for fine-grained details, and 3D **R**econstruction for spatial perception. The key insight behind *CLAR* is that both spatial awareness and advanced semantic understanding are equally crucial for developing a powerful and robust representation model in manipulation tasks. To enhance spatial understanding, we employ a point cloud MAE within an encoder-decoder architecture. This approach enables the model to learn high-level latent features from unmasked point patches while reconstructing masked ones, thereby strengthening its spatial reasoning capabilities [10]. For robust semantic understanding, we leverage pre-trained 2D foundation models and use a contrastive learning-based

* Equal Contribution.
† Corresponding author

This work was supported by the Suzhou Innovation and Entrepreneurship Leading Talents Programme - Innovation Leading Talent in Universities and Research Institutes with Grant No. ZXL2025310

[1]SKL-MAIS, Institute of Automation, Chinese Academy of Sciences, [2]School of Artificial Intelligence, University of Chinese Academy of Sciences, [3]Beijing Academy of Artificial Intelligence, Beijing, China, [4]Carnegie Mellon University, [5]Galbot, [6]CFCS, School of Computer Science, Peking University,

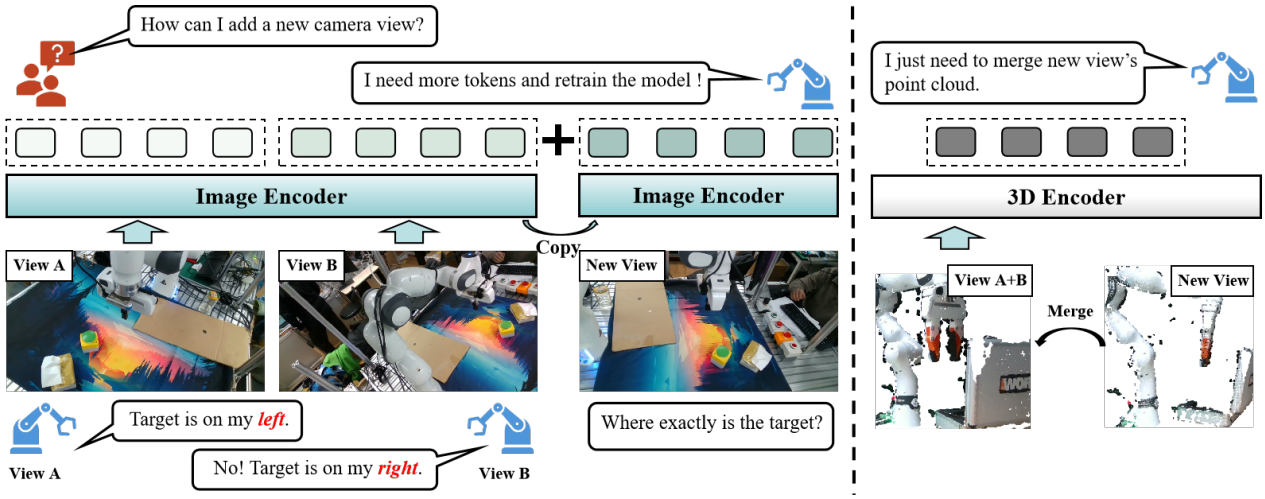

Fig. 1. **A Comparison of 2D and 3D Modalities for Robotic Pre-training. Left:** 2D-based methods lack a unified spatial coordinate system, leading to multi-view ambiguity where an object's relative position (e.g., "left" vs. "right") is inconsistent across views. This forces the policy to learn complex, view-dependent mappings, making it difficult to incorporate new camera views without costly retraining. **Right:** 3D-based methods using point clouds operate in a unified, robot-centric coordinate frame. This resolves view ambiguity and allows new visual information to be seamlessly integrated via simple geometric fusion, enhancing the policy's adaptability and spatial awareness.

mechanism to align the point cloud feature space with the powerful and expressive feature space of foundation models (such as CLIP). This alignment facilitates rapid semantic knowledge transfer from the 2D foundation model to our 3D representation model without requiring extensive 3D pre-training data, thereby mitigating the challenges posed by limited 3D robotic datasets [11]. Furthermore, to enhance perception of scene details and resolve the 'contextual mismatch' common in robotics pre-training, we design a novel local feature alignment strategy. Instead of forcing a flawed global alignment between a cropped point cloud and a full image, our strategy uses deformable attention [12] to adaptively match local 3D patches with corresponding 2D regions. This achieves a robust cross-modal local alignment that focuses learning on meaningful, shared information, respecting the irregular structure of the point cloud.

To summarize, our main contributions are as follows:

- We propose *CLAR*, a novel representation learning framework tailored for robotic manipulation. *CLAR* integrates 3D reconstruction to enhance spatial perception and employs global contrastive learning to strengthen semantic understanding.
- We propose a novel adaptive local alignment mechanism designed to preserve the fine-grained details essential for manipulation, overcoming the limitations of global alignment. By leveraging deformable attention for adaptive local feature matching, our mechanism also effectively resolves the contextual mismatch introduced by standard cropping procedures.
- We conduct extensive experiments to validate the effectiveness of *CLAR*. It outperforms state-of-the-art (SOTA) approaches in visuomotor robotic manipulation, with results on MetaWorld (82.6% vs. 76.8%), RLBench (82.0% vs. 77.0%), and real-world tasks (83.0% vs. 61.0%), demonstrating enhanced spatial awareness and semantic understanding.

II. RELATED WORK

A. 3D Representation Pre-training

Recent studies attempt to enhance the 3D understanding capabilities by pre-training on 3D point clouds [10], [11], [13], [14], [15]. PointMAE [10] and PointBert [16] leverage MAE to reconstruct masked point clouds for strengthening the models' geometric comprehension. Following the success of contrastive learning in 2D foundation models, [11], [13] apply contrastive learning for aligning textual, image, and point cloud features to facilitate the learning of 3D representations. Recon [14] and ShapeLLM [15] attempt to combine both methods simultaneously, achieving promising results. However, these approaches have been trained and tested on datasets comprising virtually limitless object data, making them challenging to apply to robotic tasks characterized by data scarcity, complex environments, and significant variations in point cloud scales. Crucially, these methods typically perform feature alignment at a global level, overlooking the fine-grained, local correspondences that are essential for precise manipulation tasks. As a result, these pre-trained models encounter difficulties when applied to robotic tasks. Our aim is to propose a 3D representation learning framework tailored for robotic manipulation tasks, capable of acquiring spatial comprehension skills while also attaining a degree of advanced semantic understanding.

B. Robot Representation Learning

With the maturation of the pre-trained visual representation learning paradigm, many studies [4], [17], [18], [19] focus on enhancing visual representations in robotics using pre-training models. Currently, most efforts are concentrated on 2D-based representations, with frameworks such as OpenVLA [20], and RDT [21] utilizing pre-trained 2D foundation models for extracting visual features. R3M [17] attempts to learn visual representations from human video

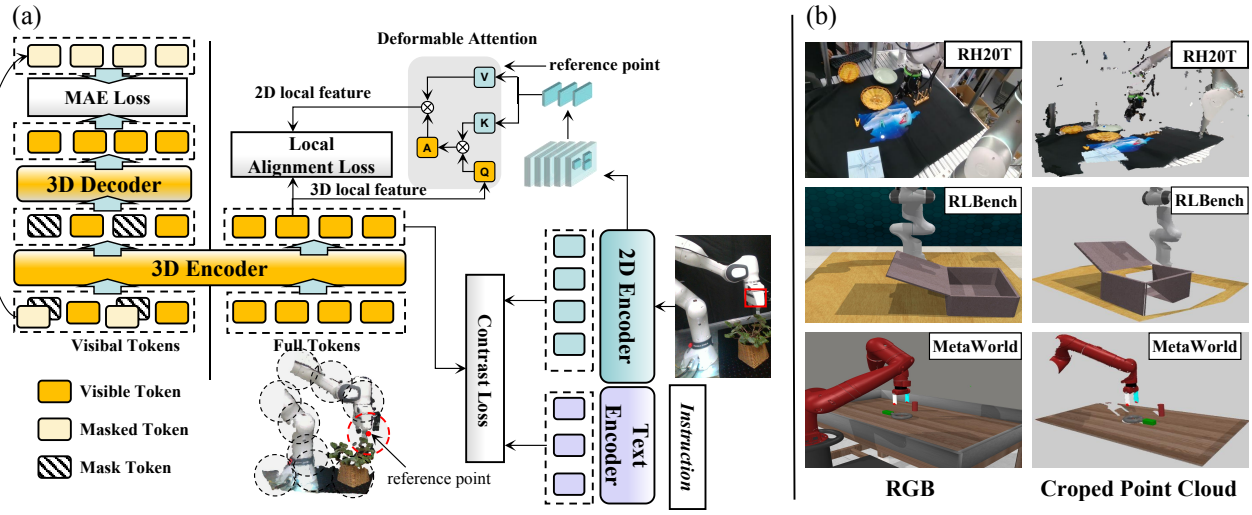

Fig. 2. **(a) The *CLAR* Pre-training Framework**. ***CLAR*** enhances spatial understanding via MAE and semantic comprehension through contrastive learning. To capture the fine-grained local details essential for robotic tasks, we supplement the global contrastive loss with an adaptive local feature alignment mechanism. **(b) Contextual Mismatch Induced by Point Cloud Cropping**. The common practice of cropping a point cloud in robotics removes background context, leading to a feature discrepancy between the partial point cloud and the full RGB image.

data through contrastive learning, while MVP [22], VC-1 [19], and Voltron [23] apply MAE to strengthen the visual representation capabilities in robotics. However, these 2D-based methods lack spatial awareness and suffer from challenges in cross-dataset camera view ambiguities, making it difficult to obtain a robust robotic representation module through pre-training on effective robotic datasets. Other studies also attempt to leverage 2D modality data to learn 3D representation capabilities. Lift3D [4], MV-MWM [24], 3D-MVP [25], and SPA [18] aim to acquire 3D visual representations through multi-view data and depth image. While these methods strengthen the spatial awareness of 2D models, they remain imprecise in capturing fine geometric structures due to limitations of data modality. Unlike these methods, our method *CLAR* conducts 3D modality pre-training, leveraging multi-view embodied datasets to enhance spatial awareness and semantic understanding capabilities in robotic manipulation tasks.

III. METHOD

A. Globla Contrastive Learning and 3D Reconstruction for Robotic Representation

Existing 3D pre-training approaches often struggle to generalize from conventional object-centric datasets to complex robotic scenes, and inadequately integrate the geometric reconstruction advantages of masked autoencoding (MAE) with the semantic distillation capabilities of contrastive learning. To address these limitations, we employ 3D point clouds as input to enhance spatial representation and enforce coordinate consistency in 3D pre-training models. Our key insight is that both spatial awareness and semantic understanding are essential for robotic tasks. To enhance spatial awareness, we utilize point cloud MAE in *CLAR*, while cross-modal contrastive learning is employed to effectively transfer semantic knowledge from pre-trained 2D foundation models into *CLAR*, as illustrated in Fig. 2-(a). First, during pre-training, we project all 3D point cloud data into a unified robot coordinate system using the extrinsic parameters of each camera. This step mitigates ambiguity caused by varying camera viewpoints across datasets, enabling our method to establish generalizable 3D perception with fewer 3D data during pre-training. Next, given an input point cloud $X \in \mathbb{R}^{N\times 3}$ with N points, we use the farthest point sampling (FPS) to sample n center points $P_c \in \mathbb{R}^{n\times 3}$. We then apply the K-Nearest Neighbors (KNN) algorithm to search the neighborhoods of these n points, generating corresponding n point patches $P \in \mathbb{R}^{n\times k\times 3}$. A high masking ratio m (60%-80%) is randomly applied to P, with the masked patches denoted as $P_{gt} \in \mathbb{R}^{mn\times k\times 3}$, which is used as the ground truth (GT) for computing the reconstruction loss. The visible point patches $P_v \in \mathbb{R}^{(1-m)n\times k\times 3}$ and full point patches P are embedded into visible tokens T_v and full tokens T, respectively. Subsequently, a standard transformer $f_\theta(\cdot)$ embeds T_v and T, along with their positional embeddings, generating encoded visible patches $\mathcal{F}_v^P$ and encoded complete patches $\mathcal{F}^P$, respectively.

Enhancing Geometric Perception with MAE. *CLAR* inputs $\mathcal{F}_v^P$ and masked tokens $\mathcal{F}_m^P$ into the MAE decoder, where a complete set of positional embeddings is added to each transformer block. The final reconstructed point patches P_{pre} are obtained after passing through a prediction head. The MAE loss is computed as the $\ell 2$ Chamfer Distance between P_{pre} and P_{gt}.

$$\mathcal{L}^R = \frac{1}{|P_{pre}|}\sum_{r\in P_{pre}} \min_{g\in P_{gt}} \|r-g\|_2^2 + \frac{1}{|P_{gt}|}\sum_{g\in P_{gt}} \min_{r\in P_{pre}} \|r-g\|_2^2 \tag{1}$$

Enhancing semantic understanding with contrastive learning. *CLAR* utilizes pre-trained image encoder E_I and text encoder E_T to extract image features $\mathcal{F}^I$ and text features $\mathcal{F}^T$, respectively. We align $\mathcal{F}^P$ with $\mathcal{F}^I$ and $\mathcal{F}^T$ through a distillation-like paradigm, which efficiently transfers the semantic understanding capabilities of the pre-

trained models to *CLAR*. The 3D-to-image and 3D-to-text alignment is formulated using a contrastive loss function, which corresponds to the "Contrast Loss" module in Fig. 2-(a):

$$\mathcal{L}_{PI}^{C} = -\frac{1}{2}\sum_{i}\left[\mathcal{S}\left(\mathcal{F}^{P},\mathcal{F}^{I}\right)_{i} + \mathcal{S}\left(\mathcal{F}^{I},\mathcal{F}^{P}\right)_{i}\right], \quad (2)$$

$$\mathcal{L}_{PT}^{C} = -\frac{1}{2}\sum_{i}\left[\mathcal{S}\left(\mathcal{F}^{P},\mathcal{F}^{T}\right)_{i} + \mathcal{S}\left(\mathcal{F}^{T},\mathcal{F}^{P}\right)_{i}\right], \quad (3)$$

$$\mathcal{S}\left(\mathcal{U},\mathcal{V}\right)_{i} = \log\frac{\exp\left(\mathcal{U}_{i}\mathcal{V}_{i}/\tau\right)}{\sum_{j,i\neq j}\exp\left(\mathcal{U}_{i}\mathcal{V}_{j}/\tau\right)}, \quad (4)$$

where τ is a learnable temperature parameter, and i, j are the sampling indices.

B. Adaptive Fine-Grained Cross-Modal Alignment of Local Features.

To overcome the failure of global feature alignment in capturing fine-grained details for manipulation, we introduce an adaptive local alignment mechanism using deformable attention. This forces precise local 3D-to-2D correspondences, capturing the essential details for manipulation while simultaneously resolving the contextual mismatch caused by cropping techniques, as visually illustrated in Fig. 2-b. Specifically, our mechanism performs this alignment via cross-attention. The process begins by constructing a dense mapping $M^{P->I}$ from the point cloud coordinate system P to the image coordinate system I, which projects a 3D spatial point (x^P, y^P, z^P) to a 2D pixel coordinate (u^I, w^I). For each 3D point cloud patch, its encoded geometric feature serves as the **query** $q = \mathcal{F}^{P}_{\text{local}}$. The entire 2D image feature map serves as the **value** $\mathbf{v}$. However, since 3D point cloud patches are defined by KNN, they have a variable spatial scale. This presents a mismatch with the typically fixed receptive fields used on 2D images.

To resolve this mismatch, we employ deformable attention [12], [26] to extract local features. Departing from conventional methods that attend to features within a static, grid-like receptive field, the core of our approach is using the local patch centers of the point cloud as queries. Since the queries originate from the 3D point cloud, the attention mechanism is guided to sample features only from corresponding valid regions within the 2D image, thereby overcoming the contextual mismatch caused by point cloud cropping. For each query q with a corresponding 2D reference point p_q, the model predicts sampling offsets Δp_{qk} and attention weights A_{qk}. The aligned 2D visual feature is then computed as a weighted sum of the sampled features:

$$\mathcal{F}^{I}_{\text{local}} = \sum_{k=1}^{K} A_{qk}\cdot\mathbf{v}(p_q + \Delta p_{qk}) \quad (5)$$

where p_q is the normalized 2D coordinate projected from the 3D patch center, while Δp_{qk} and A_{qk} are learned adaptively from the query q. This mechanism dynamically adjusts its receptive field on the image according to the input 3D query, allowing it to flexibly capture visual information corresponding to 3D geometric structures of varying scales. This strategy ensures that the model learns precise and robust local correspondences between the 3D shape and its 2D appearance. The Local Alignment Loss shown in Fig. 2-(a), denoted as $\mathcal{L}_{PI}^{local}$:

$$\mathcal{L}_{PI}^{local} = -\frac{1}{2}\sum_{i}\left[\mathcal{S}\left(\mathcal{F}^{P}_{local},\mathcal{F}^{I}_{local}\right)_{i} + \mathcal{S}\left(\mathcal{F}^{I}_{local},\mathcal{F}^{P}_{local}\right)_{i}\right] \quad (6)$$

Finally, the total loss L is a weighted sum of all components:

$$\mathcal{L} = \alpha\mathcal{L}^{R} + \beta\mathcal{L}_{PI}^{C} + \gamma\mathcal{L}_{PT}^{C} + \theta\mathcal{L}_{PI}^{local} \quad (7)$$

where $\alpha = 1.5$, $\beta = 0.5$, $\gamma = 0.5$ and $\theta = 0.5$ represent the weighting coefficients for each component of the loss function.

C. Pre-training Dataset

Due to the scarcity of embodiment data containing camera parameters and depth images, we select RH20T [27] and RLBench [28] datasets for pre-training. RH20T contains 110K contact-rich robot trajectories, over 140 tasks, and a large multi-view dataset with RGB, depth, and other modalities, providing a strong foundation for unified spatial coordinate systems. To compensate for the lack of real-world data, we supplement it with simulation data from RLBench (*close-jar*, *insert-onto-square-peg*, *move-hanger*, *open-drawer*, *push-buttons*, *put-groceries-in-cupboard*, *stack-blocks*, *stack-cups*, *turn-oven-on*). By co-training RH20T real-world data with RLBench simulation data, we partially alleviate the data scarcity issue.

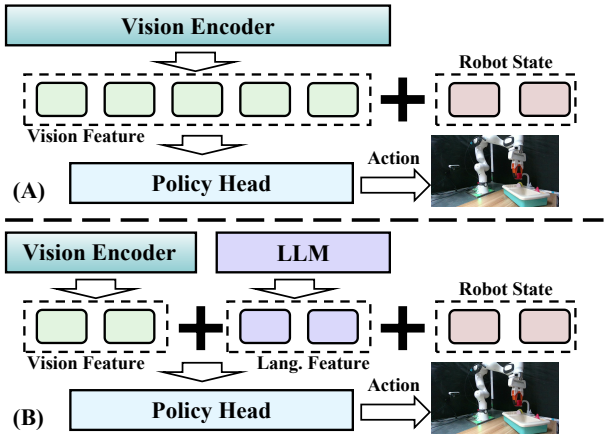

Fig. 3. **Policy Architecture for Imitation Learning.** (A) Single-task experimental setup. (B) Multi-task experimental setup.

D. Policy Architecture

We evaluate all models using a unified imitation learning policy architecture, as depicted in Fig. 3. The policy head is consistently an MLP-based structure. In standard, single-task experiments (non-multi-task experiments, Fig. 3-A), the input to the policy head is formed by concatenating two components: the visual features from the vision encoder being evaluated and the robot's proprioceptive state. For multi-task experiments (Fig. 3-B), an additional modality is incorporated into the input: language features representing

TABLE I

QUANTITATIVE RESULTS (SUCCESS RATE, %) ON REPRESENTATION MODELS (METAWORLD & RLBENCH)

Methods	Type	Input Data	Pretrained	Easy(5)	Medium(5)	Hard(5)	Mean	RLBench(4)
CLIP [29]	2D Rep.	RGB	✓	58.0%	62.4%	48.8%	56.4%	48.0%
R3M [17]	2D Rep.	RGB	✓	80.0%	68.8%	56.4%	68.4%	53.0%
VC-1 [19]	2D Rep.	RGB	✓	50.8%	72.4%	47.2%	56.8%	43.0%
VGGT [30]	2D Rep.	RGB	✓	16.8%	17.6%	2.4%	12.2%	-
PointNet [31]	3D Rep.	PC	✘	70.4%	52.4%	44.0%	55.6%	55.5%
PointNet++ [32]	3D Rep.	PC	✘	82.0%	60.8%	57.6%	66.8%	60.0%
PointNext [33]	3D Rep.	PC	✘	76.8%	63.6%	39.2%	59.8%	56.0%
DP3 [34]	3D Rep.	PC	✘	79.2%	51.2%	50.5%	60.3%	-
SPA [18]	3D Rep.	RGB	✓	65.2%	76.8%	57.6%	66.5%	63.0%
Lift3D [4]	3D Rep.	PC	✓	89.2%	80.4%	60.8%	76.8%	77.0%
CLAR (Ours)	3D Rep.	PC	✓	**96.0%**	**80.8%**	**71.2%**	**82.6%**	**82.0%**

* 2D Rep. and 3D Rep. denote 2D and 3D representation methods, respectively.

the task instruction, which are encoded by a pre-trained LLM. In all settings, features from different modalities (such as vision and language) are projected to a shared dimensionality before being fused and processed by the policy head to predict the action.

IV. EXPERIMENT

A. Experimental Setup

Benchmarks. We evaluate various methods on two widely used manipulation benchmarks: MetaWorld [35], based on MuJoCo, and RLBench [28], based on CoppeliaSim. We select 15 tasks of varying difficulty levels from MetaWorld: **easy** (*reach, lever-pull, handle-pull, peg-unplug-side, dial-turn*), **medium** (*hammer, sweep-into, bin-picking, push-wall, box-close*), and **hard** (*assembly, hand-insert, shelf-place, pick-place-wall, disassemble*). In RLBench, we select four tasks where the front camera view is employed: *close-box, close-laptop-lid, water-plants, and toilet-seat-down*.

Data preparation. Expert policies from MetaWorld and RLBench are used to collect trajectories. For MetaWorld, 30 demonstrations are gathered per task, with 25 used for training and 5 for validation, each containing 200 steps. For RLBench, the demonstrations comprise keyframe waypoint data rather than the dense trajectory data used in MetaWorld.

Baselines. We compare eleven methods across four categories: (1) **2D pre-training representation methods**, including CLIP [29] (pre-trained on general internet datasets), VGGT[1] [30] (pre-trained on large-scale datasets with 3D supervision) and two robotic-specific methods, R3M [17] and VC-1 [19] (both pre-trained on robotic datasets). (2) **3D representation methods without pre-training**, including widely used approaches such as PointNet [31], PointNet++ [32], and PointNext [33]. (3) **3D robotic pre-training representation methods**, including SPA [18] and Lift3D [4]. SPA employs differentiable neural rendering on multi-view images to improve 3D representation. Lift3D, a previous SOTA 3D robotic pre-training representation method, employs task-aware MAE during the pre-training phase to reconstruct the masked depth image. (4) **3D policies**, including DP3 [34] and RVT-2 [6]. *CLAR* is compared with DP3 on MetaWorld and with RVT-2 on RLBench.

[1]For VGGT, we extract the intermediate feature map from its DPT depth head and use a convolutional projection head to aggregate it into a single feature vector for the subsequent action head.

B. Experiments on Simulation Benchmarks

Evaluation Details. Experiments in this section aim to evaluate various representation methods on simulation benchmarks for robotic manipulation (MetaWorld and RLBench). Each method is trained for 150 epochs, with 25 rollouts performed every 10 epochs, and the average success rate (SR) of the best performing model during training is reported. Note that our *CLAR* focuses solely on the representation modules without modifying downstream robot action prediction models.

Results and Analysis. As shown in Table I, our *CLAR* achieves average success rates of 82.6% on MetaWorld and 82.0% on RLBench, significantly outperforming all baselines. The experimental results clearly reveal the superiority of 3D pre-training methods compared to their 2D counterparts. This indicates that even when 2D methods are pre-trained on larger-scale data, their performance is ultimately capped by an inherent lack of spatial awareness and multi-view ambiguity. For example, although VGGT demonstrates excellent performance in 3D reconstruction, its highly specialized training strategy impairs its generalization capabilities for robotics tasks that require broad scene understanding. In contrast to these approaches, *CLAR* acquires superior spatial awareness and geometric comprehension by pre-training directly on 3D data in a self-supervised manner within a unified spatial coordinate system, thereby achieving SOTA performance. Furthermore, *CLAR* shows strong competitive performance against other advanced 3D policy methods like DP3[34] and RVT-2[6], proving that its extracted 3D spatial features are better aligned with the demands of robotic manipulation.

C. Real World Experiments

Evaluation Details. Our real-world evaluation is conducted on a 7-DOF Franka Emika robot arm, with a static side view captured by an Intel RealSense D455 RGB-D camera. In this setup, we compare our *CLAR* against three baselines (VC-1, PointNet, Lift3D) on five tasks from

TABLE II

QUANTITATIVE RESULTS (SUCCESS RATE, %) ON POLICY METHODS (RLBENCH)

Method	close b.	close l.	water p.	toilet s. d.	Mean
PointNet [31]	52%	88%	20%	96%	64%
RVT-2(M) [6]	88%	**100%**	12%	**100%**	75%
RVT-2 [6]	**96%**	76%	16%	96%	71%
Lift3D [4]	92%	92%	24%	**100%**	77%
CLAR (Ours)	**96%**	**100%**	**32%**	**100%**	**82%**

* close b., close l., water p., and toilet s. d. denote close box, close laptop lid, water plants, and toilet seat down, respectively. RVT-2(M) denotes RVT-2 in the multi-view setting.

RLBench [28]: *pick-banana*, *put-bread*, *water-flower*, *place-cube-in-basin*, and *open-drawer*. For each task, we collect 30 teleoperated demonstrations at 10Hz for training, with randomized initial states. The final trained policy is then evaluated 20 times per task. The policy outputs a 7D action space, comprising a 6D end-effector delta pose and a 1D binary gripper action. All experiments are run on an NVIDIA RTX A6000 GPU.

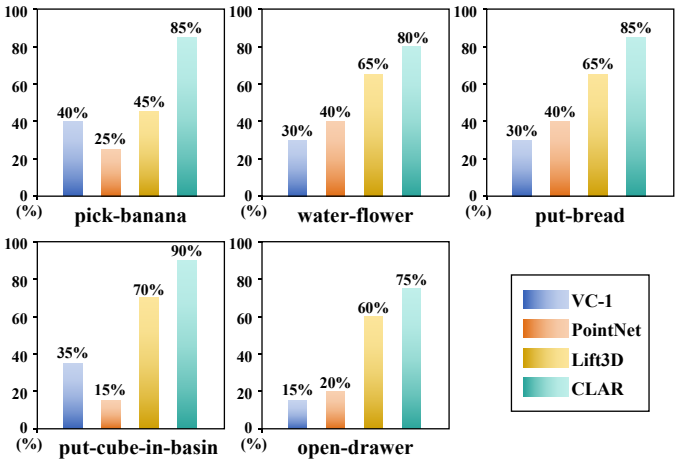

Fig. 4. **Quantitative Results (Success Rate, %) for Real World Experiments**.

Results and Analysis. As shown in Fig. 4, *CLAR* achieves SOTA performance across multiple real-world tasks, demonstrating the robustness of its representation module for robotic manipulation. The failures of the baseline methods highlight the advantages of our approach. The 2D-based method, VC-1, struggles to accurately perceive the geometry of fine structures, such as the kettle handle in the *water-flower* task. The non-pre-trained PointNet performs poorly across all tasks due to its inability to adapt from limited few-shot demonstration data (30 trajectories). Even the 3D-aware Lift3D falters in tasks requiring precise manipulation (e.g., grasping in *water-flower* or insertion in *open-drawer*). This limitation stems from a modality mismatch, as it is pre-trained on 2D depth reconstruction but uses point clouds for downstream tasks, diminishing its pre-training effectiveness. In contrast, *CLAR* overcomes these issues by pre-training directly in 3D space with a unified coordinate system, leveraging both MAE and contrastive learning to enhance spatial and semantic understanding. This allows it to better comprehend object geometries and spatial relations, leading to superior real-world success rates.

D. Multi-Task Experiments

Evaluation Details. Experiments in this section demonstrate the semantic understanding capabilities of different methods by introducing language instructions to enable robots to learn multi-task policies. We select three tasks from MetaWorld [35]: *button-press*, *reach*, and *lever-pull*. In addition, we generate several semantically similar robot instructions using ChatGPT for each task, such as "Move to the goal position," "Go to the target position," etc., resulting in a total of 18 instructions across the three tasks. Each method uses CLIP to extract text features, which are concatenated with visual features and robot state information before being fed into the policy head. We employ multi-task training but evaluate each task separately to validate semantic understanding.

TABLE III

QUANTITATIVE RESULTS (SUCCESS RATE, %) FOR MULTI-TASK EXPERIMENTS ON METAWORLD

Method	button-press	reach	lever-pull	Mean
CLIP [29]	60%	20%	20%	33.3%
R3M [17]	40%	**60%**	60%	53.3%
VC-1 [19]	60%	**60%**	80%	66.6%
PointNet [31]	60%	40%	60%	53.3%
PointNet++ [32]	**80%**	20%	20%	40.0%
PointNext [33]	60%	40%	20%	40.0%
SPA [18]	**80%**	0%	**100%**	60.0%
Lift3D [4]	60%	40%	80%	60.0%
CLAR (Ours)	60%	**60%**	**100%**	**73.3%**

Results and Analysis. The multi-task experiment results in Table III show that *CLAR* achieves an average success rate of 73.3%. Models pre-trained on robotic datasets, such as Lift3D, SPA, VC-1, and R3M, generally surpass non-pre-trained 3D methods like PointNet, PointNet++, and PointNext. This suggests that these models benefit from the robust semantic understanding of pre-trained 2D foundation models. In contrast, CLIP, despite its strong general semantic understanding, performs poorly in robotic tasks due to the lack of pre-training in robotic scenarios. Our *CLAR*, however, effectively transfers the semantic understanding of 2D foundation models to 3D through contrastive learning loss, ensuring alignment between the 3D feature space and the 2D foundation model's feature space. Finally, leveraging the semantic understanding distilled from 2D pre-trained model, *CLAR* outperforms others and achieves a high success rate in this instruction-guided, multi-task imitation learning setting.

E. Perspective Variation Experiments

Evaluation Details. This experiment examines whether perspective variations cause ambiguity in 2D representation methods and assesses their impact on 3D representation methods with a unified spatial coordinate system. We evaluate our method against 2D representation methods (including R3M, VC-1, CLIP and Lift3D) on MetaWorld [35], using training data from the "corner" perspective and testing under

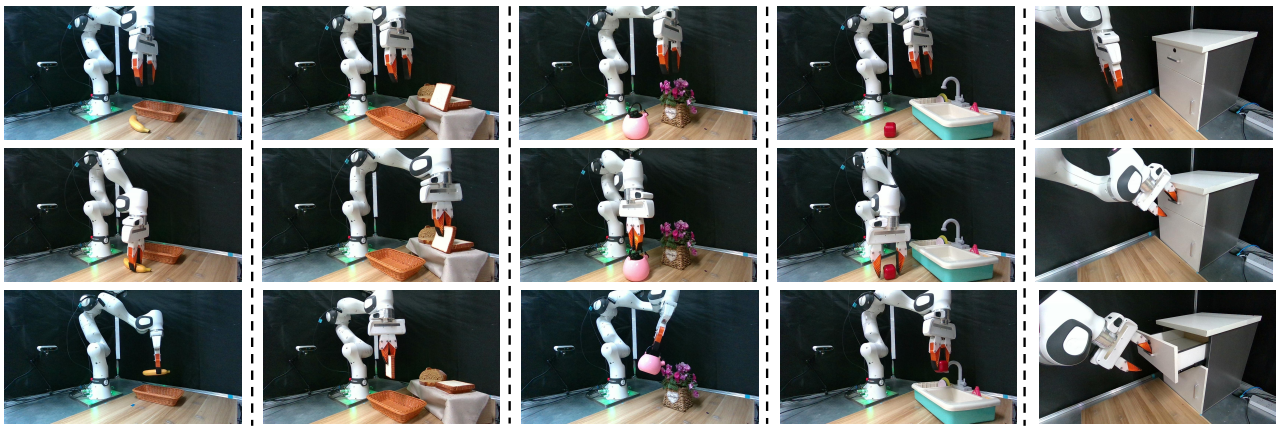

Fig. 5. **Visualization of Real world Experiments**. The camera perspectives used in the experiments differ from the one used for visualization.

TABLE IV

QUANTITATIVE RESULTS (SUCCESS RATE, %) FOR PERSPECTIVE VARIATION EXPERIMENTS ON METAWORLD

Methods	button-press	reach	box-close	Mean
CLIP [29]	50% (100%)	28% (60%)	4% (64%)	27.3% (74.6%)
R3M [17]	0% (92%)	8% (66%)	4% (96%)	18.6% (84.6%)
VC-1 [19]	8% (92%)	28% (36%)	4% (66%)	13.3% (64.6%)
Lift3D [4]	100% (100%)	80% (84%)	56% (92%)	78.6% (92%)
CLAR (Ours)	**100%** (100%)	**80%** (88%)	**72%** (96%)	**84%** (94.6%)

* $(\cdot)$ represents S.R. from the normal camera perspective.

the "corner2" perspective. Other experimental conditions and details remain consistent with the benchmark experiments.

Results and Analysis. As shown in Table IV, 2D representation methods suffer a sharp performance drop under perspective variations, whereas 3D-based methods demonstrate significantly greater robustness. The core reason is that 2D representations lack a unified spatial coordinate system, making their understanding of object relations viewpoint-dependent. Even advanced methods like VC-1 and R3M, pre-trained on multi-view data, are confined to a camera-centric system and thus fail when the test viewpoint differs from the training one. In contrast, 3D methods learn viewpoint-invariant spatial relationships by projecting point clouds into a unified coordinate system. This explains why both *CLAR* and Lift3D exhibited slight performance degradation. Notably, the distinction between them is *CLAR* unifies the spatial coordinate system during pre-training, whereas Lift3D's pre-training is camera-perspective-based. This fundamental difference allows our method to better handle perspective shifts, evidenced by a mere 10.6% drop in its success rate.

F. Ablation Study

Evaluation Details. To validate the effectiveness of each module, we perform ablation experiments on *CLAR* across four tasks from MetaWorld [35]: *hammer*, *bin-picking*, *assembly*, and *shelf-place*. By removing the MAE loss, contrastive learning loss and local alignment loss separately, we examine the impact of each component on the robotic manipulation task. The average success rate measures the effect of each module.

Results and Analysis. From Table V, we observe that both

TABLE V

QUANTITATIVE RESULTS (SUCCESS RATE, %) FOR ABLATION STUDY ON METAWORLD

Method	Pretrained	MAE	Con.	Local	Mean
CLAR	✓	✓	✓	✓	**83%**
CLAR w/o MAE	✓	✘	✓	✓	73%
CLAR w/o Local	✓	✓	✓	✘	77%
CLAR w/o Local and Con.	✓	✓	✘	✘	70%
CLAR w/o Pre.	✘	✘	✘	✘	65%

* Pretrained denotes whether the model has undergone pretraining, MAE refers to the use of a point cloud masked autoencoder, Con. indicates the application of contrastive loss, and Local signifies the exclusion of the local alignment loss.

MAE and contrastive losses enhance the visual representation capabilities of the 3D pre-training model for robotic manipulation tasks during pre-training. Notably, MAE has the most significant impact in these tasks, as robust spatial awareness is more critical than semantic understanding when policy is required to tackle only one specific task. However, semantic understanding becomes equally important when the policy is required to tackle multiple tasks simultaneously. The results also show that incorporating contrastive learning alongside MAE does not hinder spatial understanding; instead, MAE loss and contrastive loss complement each other, further enhancing the 3D pre-training model's representation capabilities. Furthermore, the local feature alignment module yields an 6% gain for *CLAR*. Visualizations of the attention maps (Fig. 6) confirm that this component directs the model's focus toward task-relevant local geometric features, which in turn boosts its downstream imitation learning performance.

V. CONCLUSION AND LIMITATIONS

In this work, we presented *CLAR*, a novel 3D pre-training framework for robotic manipulation. Our approach bridges the gap between spatial-geometric and semantic representation learning by synergistically integrating a Masked Autoencoder with cross-modal contrastive learning. To make this fusion effective for manipulation, we introduced an adaptive local alignment mechanism to focus on fine-grained features, overcoming the limitations of conventional global alignment. Extensive experiments validate that *CLAR* establishes a new state-of-the-art, providing a robust and generalizable representation for visuomotor policies.

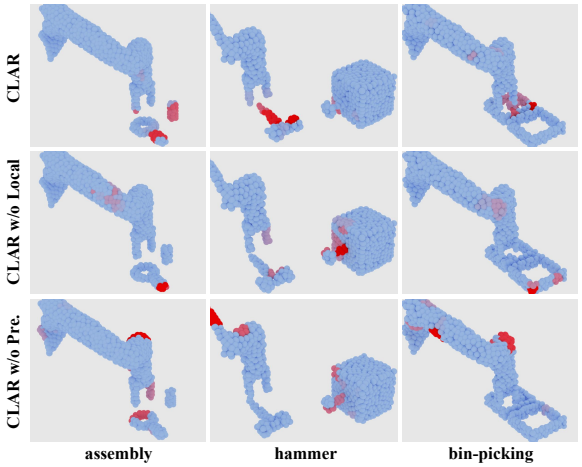

Fig. 6. **Visualization of attention maps.** Red highlights the regions with the highest attention scores.

Limitations. A primary limitation is the scarcity of suitable 3D robotics datasets with precise camera and depth information, which restricts our pre-training scale compared to 2D foundation models. Furthermore, our experiments are confined to tabletop manipulation; while the model excels in this domain, its performance in more complex environments involving mobile navigation remains to be explored.

REFERENCES

[1] Y. Chen, S. Tian, S. Liu *et al.*, "ConRFT: A Reinforced Fine-tuning Method for VLA Models via Consistency Policy," *Robotics: Science and Systems*, 2025.

[2] Y. Chen, W. Cui, Y. Chen *et al.*, "RoboGPT: an LLM-based Long-term Decision-making Embodied Agent for Instruction Following Tasks," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 17, no. 5, pp. 1163–1174, 2025.

[3] H. Li, Z. Jiang, Y. Chen, and D. Zhao, "Generalizing consistency policy to visual rl with prioritized proximal experience regularization," *Advances in Neural Information Processing Systems*, vol. 37, pp. 109 672–109 700, 2024.

[4] Y. Jia, J. Liu, S. Chen *et al.*, "Lift3D Policy: Lifting 2D Foundation Models for Robust 3D Robotic Manipulation," *Computer Vision and Pattern Recognition Conference*, pp. 17 347–17 358, June 2025.

[5] A. Goyal, J. Xu, Y. Guo *et al.*, "RVT: Robotic View Transformer for 3D Object Manipulation," *Conference on Robot Learning*, pp. 694–710, 2023.

[6] A. Goyal, V. Blukis, J. Xu *et al.*, "RVT-2: Learning Precise Manipulation from Few Demonstrations," in *RSS 2024 Workshop: Data Generation for Robotics*.

[7] Y. Geng, B. An, H. Geng *et al.*, "RLAfford: End-to-End Affordance Learning for Robotic Manipulation," *2023 IEEE International Conference on Robotics and Automation*, pp. 5880–5886, 2023.

[8] W. Cui, C. Zhao, S. Wei *et al.*, "GAPartManip: a large-scale dataset for generalizable and actionable part manipulation with material-agnostic articulated objects," *IEEE International Conference on Robotics and Automation*, 2025.

[9] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding *et al.*, "Spatialvla: Exploring spatial representations for visual-language-action model," *arXiv preprint arXiv:2501.15830*, 2025.

[10] Y. Pang, W. Wang, F. E. Tay *et al.*, "Masked autoencoders for point cloud self-supervised learning," *European Conference on Computer Vision*, pp. 604–621, 2022.

[11] L. Xue, N. Yu, S. Zhang *et al.*, "ULIP-2: Towards Scalable Multimodal Pre-Training for 3D Understanding," *Computer Vision and Pattern Recognition*, pp. 27 081–27 091, 2024.

[12] Z. Yang, N. Song, W. Li, X. Zhu, L. Zhang, and P. H. Torr, "Deepinteraction++: Multi-modality interaction for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 8, pp. 6749–6763, 2025.

[13] J. Zhou, J. Wang, B. Ma *et al.*, "Uni3d: Exploring unified 3d representation at scale," *International Conference on Learning Representations*, 2024.

[14] Z. Qi, R. Dong, G. Fan *et al.*, "Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining," *International Conference on Machine Learning*, pp. 28 223–28 243, 2023.

[15] Z. Qi, R. Dong, S. Zhang *et al.*, "ShapeLLM: Universal 3D Object Understanding for Embodied Interaction," *European Conference on Computer Vision*, pp. 214–238, 2024.

[16] X. Yu, L. Tang, Y. Rao *et al.*, "Point-BERT: Pre-training 3D Point Cloud Transformers with Masked Point Modeling," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[17] S. Nair, A. Rajeswaran, Kumar *et al.*, "R3M: A Universal Visual Representation for Robot Manipulation," *Conference on Robot Learning*, vol. 205, pp. 892–909, 2023.

[18] H. Zhu, H. Yang, Y. Wang *et al.*, "SPA: 3D Spatial-Awareness Enables Effective Embodied Representation," *International Conference on Learning Representations*, 2024.

[19] A. Majumdar, K. Yadav, S. Arnaud *et al.*, "Where are we in the search for an artificial visual cortex for embodied intelligence?" *Advances in Neural Information Processing Systems*, vol. 36, pp. 655–677, 2023.

[20] M. J. Kim, K. Pertsch, S. Karamcheti *et al.*, "OpenVLA: An Open-Source Vision-Language-Action Model," *Conference on Robot Learning*, 2024.

[21] S. Liu, L. Wu, B. Li *et al.*, "RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation," *International Conference on Learning Representations*, 2024.

[22] I. Radosavovic, T. Xiao, S. James *et al.*, "Real-world robot learning with masked visual pre-training," *Conference on Robot Learning*, pp. 416–426, 2023.

[23] S. Karamcheti, S. Nair, A. S. Chen *et al.*, "Language-Driven Representation Learning for Robotics," *Robotics: Science and Systems*, 2023.

[24] Y. Seo, J. Kim, S. James *et al.*, "Multi-View Masked World Models for Visual Robotic Manipulation," *International Conference on Machine Learning*, pp. 30 613–30 632, 2023.

[25] S. Qian, K. Mo, V. Blukis *et al.*, "3D-MVP: 3D Multiview Pretraining for Robotic Manipulation," *Computer Vision and Pattern Recognition*, 2025.

[26] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *Computer Vision and Pattern Recognition*, 2020.

[27] H.-S. Fang, H. Fang, Z. Tang *et al.*, "RH20T: A Comprehensive Robotic Dataset for Learning Diverse Skills in One-Shot," *2024 IEEE International Conference on Robotics and Automation*, 2024.

[28] S. James, Z. Ma, D. R. Arrojo *et al.*, "RLBench: The Robot Learning Benchmark & Learning Environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.

[29] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *International Conference on Machine Learning*, vol. 139, pp. 8748–8763, 18–24 Jul 2021.

[30] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "VGGT: Visual Geometry Grounded Transformer," *Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025.

[31] R. Q. Charles, H. Su, M. Kaichun *et al.*, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 77–85, 2017.

[32] C. R. Qi, L. Yi, H. Su *et al.*, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Neural Information Processing Systems*, pp. 5105–5114, 2017.

[33] G. Qian, Y. Li, H. Peng, J. Mai, H. Hammoud *et al.*, "PointNeXt: Revisiting PointNet++ with Improved Training and Scaling Strategies," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 192–23 204, 2022.

[34] Y. Ze, G. Zhang, K. Zhang *et al.*, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," *2nd Workshop on Dexterous Manipulation: Design, Perception and Control*, 2024.

[35] T. Yu, D. Quillen, Z. He *et al.*, "Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning," *Conference on Robot Learning*, pp. 1094–1100, 2020.