

CLAMP: Contrastive Learning for 3D Multi-View Action-Conditioned Robotic Manipulation Pretraining

I-Chun Arthur Liu^{*†}, Krzysztof Choromanski^{*}, Sandy Huang^{*}, Connor Schenck^{*}

^{*}Google DeepMind [†]University of Southern California

Abstract—Leveraging pre-trained 2D image representations in behavior cloning policies has achieved great success and has become a standard approach for robotic manipulation. However, such representations fail to capture the 3D spatial information about objects and scenes that is essential for precise manipulation. In this work, we introduce Contrastive Learning for 3D Multi-View Action-Conditioned Robotic Manipulation Pretraining (CLAMP), a novel 3D pre-training framework that utilizes point clouds and robot actions. From the merged point cloud computed from RGB-D images and camera extrinsics, we re-render multi-view four-channel image observations with depth and 3D coordinates, including dynamic wrist views, to provide clearer views of target objects for high-precision manipulation tasks. The pre-trained encoders learn to associate the 3D geometric and positional information of objects with robot action patterns via contrastive learning on large-scale simulated robot trajectories. During encoder pre-training, we pre-train a Diffusion Policy to initialize the policy weights for fine-tuning, which is essential for improving fine-tuning sample efficiency and performance. After pre-training, we fine-tune the policy on a limited amount of task demonstrations using the learned image and action representations. We demonstrate that this pre-training and fine-tuning design substantially improves learning efficiency and policy performance on unseen tasks. Furthermore, we show that CLAMP outperforms state-of-the-art baselines across six simulated tasks and five real-world tasks. The project website and videos can be found at <https://clamp3d.github.io/CLAMP/>.

I. INTRODUCTION

Contrastive pre-training on internet-scale image–text pairs has become a standard approach for learning strong computer vision backbones [38, 61, 33]. However, an analogous standard approach has not emerged in robotics due to the limited availability of large-scale robot data. Moreover, it remains unclear which modalities and image representations, such as RGB, point clouds, or voxel grids, are best suited for pre-training. Pre-training in robotics is further complicated by high-dimensional state and action spaces.

Recent works on robotic manipulation [24, 34] have explored leveraging web-scale images, videos, and text with simulated and real-world robot data for pre-training. However, it is not well understood which data sources contribute most to policy fine-tuning performance. Moreover, these approaches do not pre-train on 3D data or incorporate 3D perception, which provides the spatial information robots need to plan

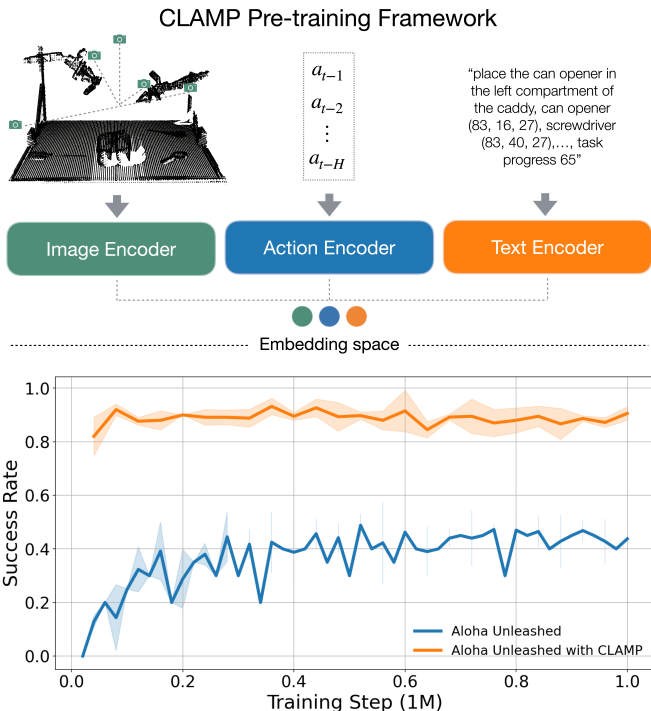


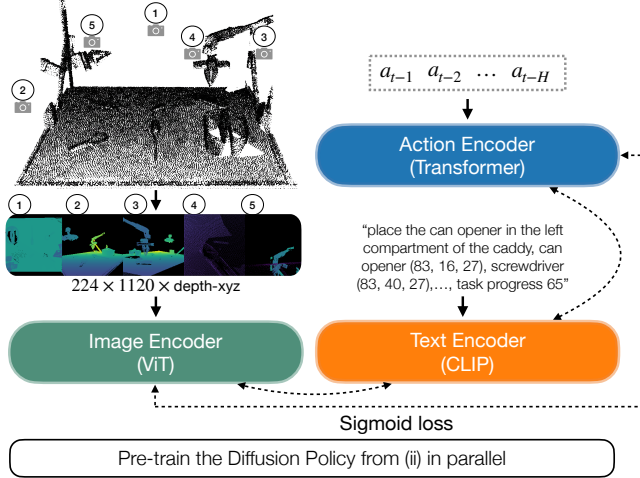
Fig. 1: CLAMP is a 3D pre-training framework for robotic manipulation. We contrastively pre-train an image encoder with fixed and dynamic multi-view observations, an action encoder with histories of action chunks, and a text encoder with spatial and temporal object- and task-level information. CLAMP learns image and action representations that improve downstream policy fine-tuning efficiency and performance. The bottom figure compares the ALOHA Unleashed evaluation curves during fine-tuning without and with CLAMP on the Mug on Plate task.

and act effectively in the physical world. A recent work, 3D-MVP [37], proposes a 3D multi-view pre-training framework using masked autoencoders for manipulation. This method learns image representations through image reconstruction, but it does not leverage robot action data during pre-training. We hypothesize that 3D perception and robot actions are valuable pre-training modalities because they are fundamental to how robots perceive and act in the world.

In this work, we propose Contrastive Learning for

[†]Work done as a student researcher at Google DeepMind.

(i) Contrastive pre-training on simulated robot data and Diffusion Policy pre-training



(ii) Diffusion Policy fine-tuning and inference

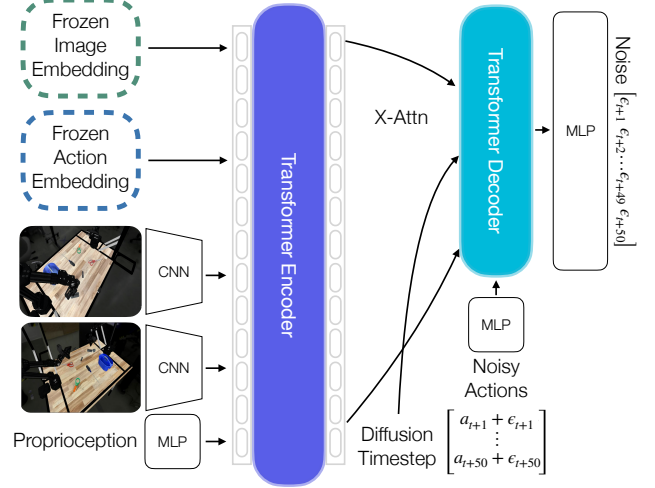


Fig. 2: **Overview of CLAMP.** (i): CLAMP consists of three encoders: image, action and text. The image encoder is a Vision Transformer (ViT) [8] that takes five multi-view observations (overhead, back-right, front-left, wrist-left, and wrist-right) as input. These views are rendered from a merged point cloud and include depth and 3D coordinates. The action encoder is a Transformer encoder that takes a history of previous actions ($a_{t-1}, \dots, a_{t-H}$) as input. The text input includes a general task description, each object’s name and normalized position in the global coordinate frame, and an integer indicating task progress, and is processed by a CLIP [38] text encoder. The intuition is to learn image and action representations that capture correlations between action patterns and robot states as observed in images, grounded by text that describes the spatial and temporal information about the objects and task. We also pre-train a Diffusion Policy [66] in parallel. (ii): The Diffusion Policy is first initialized with the pre-trained weights from stage (i) and use the frozen CLAMP encoders to extract image and action embeddings. These embeddings are fed into a Transformer encoder together with RGB feature maps from ResNet-50 backbones and proprioceptive features from the robot to predict noise of shape 50×14 for the next 50 actions ($\epsilon_{t+1}, \dots, \epsilon_{t+50}$).

3D Multi-View Action-Conditioned Robotic Manipulation Pretraining (CLAMP). We contrastively pre-train an image encoder, a text encoder, and an action encoder on large-scale simulated robot data by pulling positive image–text, image–action, and text–action pairs closer together while pushing negative pairs apart. See Figure 1 for an overview.

For image representation, we re-render multi-view four-channel observations from the merged point cloud that include depth and 3D coordinates, using fixed viewpoints and dynamic wrist views, which provide clearer observations for high-precision manipulation tasks. We apply STRING [40] relative positional encoding to the image encoder, enabling tokens from different views that correspond to nearby regions in 3D space to be correlated. For action data, we use a history of previous actions. Our intuition is that many manipulation tasks share common action patterns or trajectory segments that can be learned during pre-training, thereby reducing the amount of task-specific learning required during fine-tuning. For text, we extract object spatial relationships and task progress from the simulator and use this information to ground image and action representations. Note that the text encoder is used exclusively during pre-training; therefore, CLAMP does not rely on privileged state information during policy deployment. In addition, we also pre-train a Diffusion Policy, the visuomotor controller for the robot, in parallel to initialize the policy for fine-tuning. We find that appropriate initialization substantially improves fine-tuning efficiency and performance. We then fine-tune the

policy using task-specific demonstrations for each task.

In summary, our contributions are as follows: (i) We introduce a novel 3D contrastive pre-training framework for robotic manipulation that leverages robot actions and dynamic wrist views. (ii) We are the first to apply STRING relative positional encoding using 3D coordinates derived from point clouds. (iii) We show that dynamic wrist views are critical for strong results on high-precision tasks, and that pre-training both the encoders and the policy yields substantially better performance than pre-training the encoders alone. (iv) Extensive empirical results show that CLAMP pre-training greatly improves sample efficiency and policy performance during fine-tuning, outperforming state-of-the-art baselines.

II. RELATED WORK

Contrastive learning has seen significant interest in recent years [20] with success on tasks such as view-invariant dense object features [41, 19] and robotic manipulation [14, 30]. One major area of interest in contrastive learning has been training large-scale vision encoders. Modern approaches such as CLIP [38], ALIGN [26], COCA [55], OpenCLIP [5], SigLIP [61], Florence [57], InternVL [4], and EVA [9, 10, 43] use contrastive learning to create semantically meaningful image embeddings by aligning vision encoders with text embeddings derived from image captions. TIPS [33] takes this a step further by introducing additional loss terms to maintain spatial features as well as semantic features in the image embeddings.

ULIP [52] and ULIP-2 [53] lift contrastive learning into 3D by aligning a 3D encoder with pre-aligned text and image encoders. In this paper, we utilize contrastive learning to pre-train 3D, text, and action encoders.

Interfacing unordered point clouds with learned models has been a challenge in computer vision and robotics. PointNet [35] and PointNet++ [36] showed early success by utilizing an MLP and max pooling layers for object classification and segmentation. PointMLP [31] added to this a lightweight geometric affine module and residual connections. DGCNN [47] takes a different approach, applying a convolutional kernel on a graph neural network constructed from the point cloud. The advent of transformers [46], operating on sets of tokens, created natural opportunities for new point-cloud model architectures. PCT [15] and Point Transformer [64] used point-based features directly as the tokens for the self-attention mechanism. However, these methods required significant downsampling of the point cloud to achieve tractability, even when utilizing local attention. Point-BERT [56] solved this by first pooling local regions of points together using a mini PointNet, and generating one token per region. This methodology has been used and iterated on successfully by many follow-on works [67, 23, 51, 21, 62, 49, 50] with ULIP [52] and ULIP-2 [53] evaluating both PointNet and Point-BERT as point-cloud encoder backbones. An alternative approach to Point-BERT style architectures is the multi-view projection approach, i.e., projecting the point cloud to multiple virtual image planes and applying standard 2D image architectures to the resulting images. This has shown success for point-cloud text alignment [63, 69], point-cloud captioning [32], shape classification and recognition [12, 16, 17], and 3D visual scene grounding [22].

Prior work in robotics has shown that these point-cloud encoders increase performance. DP3 [60] showed improved success rate across 72 simulation tasks using a PointNet-style backbone. ArticBot [48] and HACMan [68] utilized PointNet++ backbones to solve articulated object and non-prehensile manipulation tasks respectively. PerAct [42] and GNFator [59] voxelize their point clouds and feed them into a Perceiver Transformer [25] showing high success rates across multiple tasks. Act3D [11] generates point clouds from multiple RGB-D images, samples potential keypoints in a coarse-to-fine manner, and cross-attends them to select grasp points. RVT [13] and 3D-MVP [37] took the multi-view projection approach and used five orthogonal image planes (1 for each cardinal direction plus a top-down view) to maximize view coverage of the point cloud, yielding significant gains in success rate. Lift3D [27] combined both the multi-view projection approach with a Point-BERT backbone by using projection to get features from 2D models, then averaging them over views in the point cloud and generating embeddings using Point-BERT. In our work, we found the multi-view approach worked best (see Sec. V-C), with three fixed views (overhead, front-left, and back-right) and two dynamic views (a view attached to the wrist of each arm).

III. METHODOLOGY

A. Problem Formulation

In contrastive pre-training, we aim to learn a model ψ that generates image and action embeddings that can be utilized to improve downstream fine-tuning of a behavior cloning policy. Given a mini-batch of image-text-action triplets $\mathcal{B} = \{(I_1, T_1, A_1), \dots, (I_n, T_n, A_n)\}$, the contrastive objective pulls the embeddings of matched triplets (I_i, T_i, A_i) closer together while pushing apart embeddings of mismatched triplets (I_i, T_j, A_k) for $i \neq j$ and $i \neq k$. We denote the per-pixel 4-channel image observation by $I \in \mathbb{R}^{H_I \times W_I \times 4}$, where H_I and W_I are the image height and width, and the four channels (D, X, Y, Z) correspond to depth and 3D coordinates. Let T be a text string containing (i) a high-level task description, (ii) each object's name and its normalized position in the global coordinate frame, and (iii) an integer indicating task progress. Note that privileged information, such as an object's name and position, is obtained from the simulator and is used only during contrastive pre-training. A denotes the history of previous actions from a dual-arm robot: $A_t = \{A_{t-1}, \dots, A_{t-H}\}$, where H is the action chunk size and missing terms are zero-padded if $t - H < 0$.

In policy fine-tuning, the objective is to train a policy ϵ_θ , parameterized by θ , that learns from a dataset of expert demonstrations $\mathcal{D} = \{E_1, \dots, E_M\}$ where $E_i = (I_1, I_1^R, A_1, \mathbf{a}_1, \dots, I_t, I_t^R, A_t, \mathbf{a}_t)$ for a demonstration with t time steps, leveraging pretrained model ψ . Here, I^R denotes an RGB image observation $I^R \in \mathbb{R}^{H_R \times W_R \times 3}$ where (H_R, W_R) may differ from (H_I, W_I) , and $\mathbf{a}$ denotes the target joint positions for the robot. Note that both I and I^R can be obtained from a standard RGB-D camera.

B. Contrastive Pre-training

Rendering. Given RGB-D images captured from a set of cameras, we construct a scene point cloud using the known camera intrinsics and extrinsics. Following [60], we omit RGB colors from the point cloud for better appearance generalization. We apply 3D bounding-box cropping and voxel-grid downsampling to remove points outside the workspace and reduce point density. We then re-render the scene from five virtual viewpoints, each with depth and 3D coordinates: overhead, front-left, back-right, wrist-left, and wrist-right. The virtual wrist cameras are defined from the left and right gripper poses, with a small offset for improved visibility. These wrist views reduce occlusion during object manipulation. We empirically find that these dynamic views outperform standard orthogonal views and are particularly effective on high-precision tasks.

Model Architecture. The encoders in CLAMP consist of an image encoder $f(\cdot)$, a text encoder $g(\cdot)$, and an action encoder $h(\cdot)$. For the image encoder, we use a Vision Transformer (ViT) [8] with STRING [40] relative positional encoding. Specifically, we use a 94M-parameter, 12-layer, 768-wide, B/16 model with 12 attention heads and multihead attention pooling (MAP) [29]. For STRING, we use the Cayley

STRING formulation, where the 3D coordinates for relative positional encodings are computed as the average global XYZ of the point-cloud points within each image patch. Because the coordinates are in the global coordinate frame, attention between image patches from different images can be seamlessly attenuated by the STRING mechanism without any additional modification. More analysis is provided in Sec. III-D.

The text encoder takes a tokenized text [39, 54] with 256 tokens as input, and we train a CLIP text encoder [38] on these tokens. We use a 277M-parameter, 12-layer, 768-wide model with 12 attention heads. The action encoder is a standard Transformer encoder with MAP pooling. We use an 18M-parameter, 2-layer, 768-wide model with 12 attention heads. Each encoder outputs a 768-dimensional embedding.

Contrastive Loss. The encoders are trained to minimize a SigLIP-style loss [61] over the three modality pairs: image-text, image-action, and text-action.

$$\mathcal{L}_{\text{ImageText}} = -\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} \log \left(\frac{1}{1 + e^{\ell_{ij}(-t \mathbf{x}_i^\top \mathbf{y}_j + b)}} \right) \quad (1)$$

$$\mathcal{L}_{\text{ImageAction}} = -\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \sum_{k=1}^{|\mathcal{B}|} \log \left(\frac{1}{1 + e^{\ell_{ik}(-t \mathbf{x}_i^\top \mathbf{z}_k + b)}} \right) \quad (2)$$

$$\mathcal{L}_{\text{TextAction}} = -\frac{1}{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} \sum_{k=1}^{|\mathcal{B}|} \log \left(\frac{1}{1 + e^{\ell_{jk}(-t \mathbf{y}_j^\top \mathbf{z}_k + b)}} \right) \quad (3)$$

where $\mathbf{x}_i = \frac{f(I_i)}{\|f(I_i)\|_2}$, $\mathbf{y}_i = \frac{g(T_i)}{\|g(T_i)\|_2}$, and $\mathbf{z}_i = \frac{h(A_i)}{\|h(A_i)\|_2}$. ℓ_{ij} , ℓ_{ik} , ℓ_{jk} are the labels for image-text, image-action, and text-action inputs, respectively; they are set to 1 for matched pairs and to -1 otherwise. Bias b and temperature t are learnable scaling parameters that help stabilize training and prevent overly large updates early on. Following [61], we initialize $b = -10$ and $t = \exp(\log 10) = 10$. The total loss of CLAMP encoders is $\frac{1}{3}(\mathcal{L}_{\text{ImageText}} + \mathcal{L}_{\text{ImageAction}} + \mathcal{L}_{\text{TextAction}})$. This formulation encourages paired modalities to be close while pushing unpaired modalities apart in a shared embedding space. The intuition is to learn action patterns or representations that can be used to infer future states, while also learning image representations that can be used to infer the history of preceding robot actions, conditioned on text describing the task type, the spatial arrangement of objects in the scene, and the temporal progress of the task.

C. Diffusion Policy Pre-training and Fine-Tuning

Diffusion Policy Architecture. The Diffusion Policy is based on the architecture proposed by Zhao et al. [66]. It has been shown to excel at handling multimodal action distributions [6, 66], which is essential for challenging tasks with multiple action modes (e.g., grasp-to-place vs. grasp-to-hand-over-to-place) or for learning from demonstrations collected by different human operators. We use the Denoising Diffusion Probabilistic Model (DDPM) formulation [18, 6]. During training, a denoising step k is randomly selected.

Conditioned on k , we randomly sample random noise ϵ^k with corresponding variance. The training objective of the noise prediction network ϵ_θ is to predict the injected noise ϵ^k : $\mathcal{L} = \text{MSE}(\epsilon^k, \epsilon_\theta(\mathbf{O}_t, \mathbf{a}_t^0 + \epsilon^k, k))$. At test time, the policy samples an action by iteratively denoising an action sequence:

$$\mathbf{a}_t^{k-1} = \alpha \left(\mathbf{a}_t^k - \gamma \epsilon_\theta(\mathbf{O}_t, \mathbf{a}_t^k, k) + \mathcal{N}(0, \sigma^2 I) \right), \quad (4)$$

where $\mathbf{a}_t^k$ is sampled from Gaussian noise; α , γ , and σ are determined by the DDPM noise schedule; $\mathcal{N}(0, \sigma^2 I)$ denotes additive Gaussian noise; and $\mathbf{O}_t$ is the observation at time t .

We use a 192M-parameter Diffusion Policy with a 4-layer, 768-wide Transformer encoder with 8 attention heads and a 7-layer, 768-wide Transformer decoder with 8 attention heads. Following [66], we apply action chunking with a chunk size of 50, and the policy outputs 14 absolute target joint positions. For each camera’s RGB image, we use a separate ResNet-50 vision backbone initialized with weights pretrained on ImageNet classification [7]. We extract feature maps from the stage-four output of ResNet. These feature maps are concatenated with unpooled frozen image and action embeddings from pretrained CLAMP encoders, taken before MAP pooling, along with features from the robot’s proprioception. We add positional embeddings before feeding the resulting tokens into the encoder. The encoder latents are then passed to the decoder together with a noised action chunk, a learnable positional embedding, and an one-hot encoding of the diffusion timestep. Finally, the decoder output is passed through a projection head to predict noise with shape 50×14 for the next 50 actions.

Pre-training and Fine-Tuning. For pre-training, we train the policy on the same dataset used to pre-train the CLAMP encoders for simulated tasks. For real-world tasks, we pre-train the policy on real-world training data, since weights learned from real data provide better initialization due to reduced sim-to-real gap compared with policies pre-trained on simulated data. Note that frozen image and action embeddings are not used during policy pre-training because the policy is trained in parallel with the encoders. For fine-tuning, we initialize all policy weights, including each ResNet backbone, from the pre-trained checkpoint. This pre-training initialization is essential for achieving high fine-tuning sample efficiency and performance, as shown in Table 1. Our key insight is to pre-train both the encoders and the policy before fine-tuning the policy. See Figure 2 for an overview of the model architecture for pre-training and fine-tuning, and refer to Appendix VII-B for implementation details.

D. Theoretical Analysis

CLAMP is the first model that applies STRING [40] relative positional encoding (RPE) mechanism with coordinate vectors coming directly from the (x, y, z) coordinates of points from the point cloud (rather than by leveraging depth, as it was the case in the 3D-STRING variants from [40]).

a) *Multi-view setting with STRING:* CLAMP solves in an elegant way the problem of modeling the correlations between tokens coming from different 2D-views. Rather than

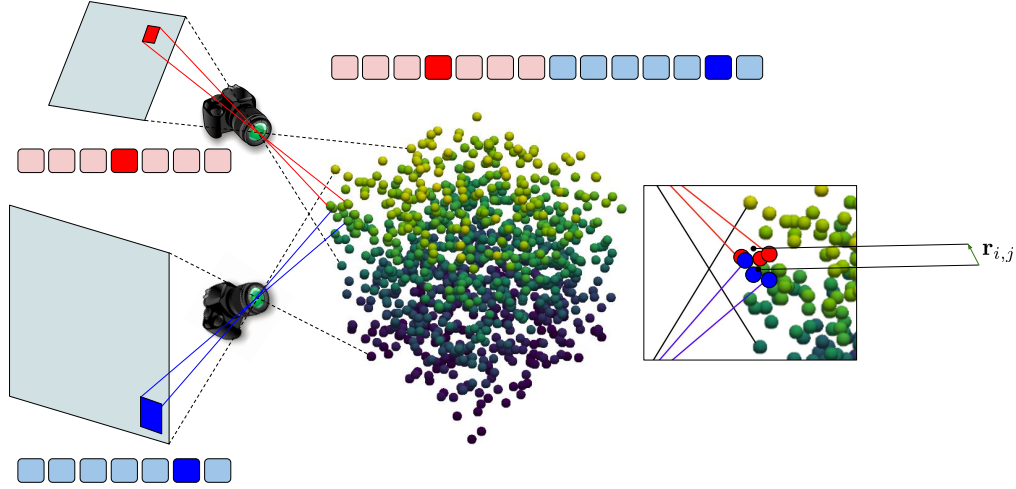


Fig. 3: Pictorial description of the STRING mechanism applied in CLAMP’s image encoder to correlate tokens from different views corresponding to similar regions of the 3D space. Tokens from different 2D views are concatenated, but also equipped with STRING positional encoding mechanism [40]. Consequently, the attention score between two highlighted tokens from two different views will be modulated by a function that depends only on the vector $\mathbf{r}_{i,j} \in \mathbb{R}^3$ between the centers of mass of two sub-point clouds (highlighted in red and blue correspondingly) of the visible sets of points related to those tokens.

leveraging non-trivial methods for finding matchings between patches from different views, but corresponding to the same or similar regions of 3D space, CLAMP does it implicitly via STRING positional encoding, associating with every token its coordinate vector, pointing to the center of mass of the set of visible points corresponding to the token. Those coordinate vectors are then used in STRING for positional encoding (see: Fig. 3). Below, we explain the mechanism in greater depth.

Lemma 3.1: Consider two tokens: $t_{\mathcal{P}_1}$ and $t_{\mathcal{P}_2}$ of the image encoder’s self-attention module used in CLAMP, corresponding to visible sub-point clouds $\mathcal{P}_1$ and $\mathcal{P}_2$ respectively. Denote by $(\mathbf{q}_{\mathcal{P}_1}, \mathbf{k}_{\mathcal{P}_1}) \in \mathbb{R}^{d_{QK}} \times \mathbb{R}^{d_{QK}}$ and $(\mathbf{q}_{\mathcal{P}_2}, \mathbf{k}_{\mathcal{P}_2}) \in \mathbb{R}^{d_{QK}} \times \mathbb{R}^{d_{QK}}$ their corresponding query-key vector-pairs and by $\mathbf{r}_{i,j} \in \mathbb{R}^3$ a vector joining the center of mass of $\mathcal{P}_i$ with the center of mass of $\mathcal{P}_j$ for $i, j \in \{1, 2\}$. For a given vector $\mathbf{u} \in \mathbb{R}^{d_{QK}}$, denote by $\mathbf{u}(\mathcal{C})$ its representation in the coordinate system $\mathcal{C}$ (we assume that the canonical system $\mathcal{C}(\mathbf{e}_1, \dots, \mathbf{e}_{d_{QK}})$ is the default one, thus: $\mathbf{u}(\mathcal{C}(\mathbf{e}_1, \dots, \mathbf{e}_{d_{QK}})) = \mathbf{u}$). Then the not-normalized logit score $s_{i,j}$ of the self-attention module in CLAMP’s image encoder is given as follows, for $\alpha(\mathbf{v}, \mathbf{w}) \in [0, \pi]$ denoting an angle between vectors $\mathbf{v}, \mathbf{w}$, learnable coordinate system $\mathcal{L}$ and linear functions δ_k ($k = 1, \dots, l = \frac{d_{QK}}{2}$; without loss of generality, we assume that query/key dim is even):

$$s_{i,j} = \sum_{k=0}^{l-1} \|\mathbf{q}_{\mathcal{P}_i}(\mathcal{L})[2k, 2k+1]\|_2 \|\mathbf{k}_{\mathcal{P}_j}(\mathcal{L})[2k, 2k+1]\|_2 \cdot \cos(\alpha(\mathbf{q}_{\mathcal{P}_i}(\mathcal{L})[2k, 2k+1], \mathbf{k}_{\mathcal{P}_j}(\mathcal{L})[2k, 2k+1]) + \delta_k(\mathbf{r}_{i,j})) \quad (5)$$

Lemma 3.1 sheds a light on the mechanism used in CLAMP to correlate tokens corresponding to sub-point clouds within different distances from each other (proof in Appendix VII-A).

Tokens encoding the same sub-point cloud: In that scenario, $\mathbf{r}_{i,j} = 0$ and thus $\delta_k(\mathbf{r}_{i,j}) = 0$ (since δ_k is a linear function). Since coordinate system change is an isometry, by

Lemma 3.1, it is easy to see that the formula for $s_{i,j}$ reduces to: $s_{i,j} = \mathbf{q}_{\mathcal{P}_i}^\top \mathbf{k}_{\mathcal{P}_j}$. Thus we obtain a regular attention score.

It turns out that STRING method applied in CLAMP has the ability to modulate large attention values of the regular attention mechanism in such a way that larger distances $\|\mathbf{r}_{i,j}\|_2$ between centers of the sub-point clouds imply smaller modulated attention scores. Intuitively, this is a very useful property, since the mechanism is capable of implicitly learning the 3D structure of the input, correlating tokens from different views, but corresponding to the same or similar regions of space (by explicitly reducing attention scores for sub-point clouds within larger distances from each other). The following is true:

Lemma 3.2: Take some $T > 0$ and $\phi \in [0, \frac{\pi}{3}]$. Consider two sub-point clouds $\mathcal{P}_i$ and $\mathcal{P}_j$ with the corresponding vector $\mathbf{r}_{i,j}$ joining their centers of masses. Take a *translated* version $\hat{\mathcal{P}}_j(\mathbf{r})$ of $\mathcal{P}_j$ such that the vector joining the centers of masses becomes $\mathbf{r} = \eta \mathbf{r}_{i,j}$ for some $\eta \in \mathbb{R}$. Assume that the query $\mathbf{q}_{\mathcal{P}_i} \in \mathbb{R}^{d_{QK}}$ and key $\mathbf{k}_{\mathcal{P}_j} \in \mathbb{R}^{d_{QK}}$ are taken from the ball $\mathcal{B}(0, R)$ centered at zero and of radius R and that the absolute values of all their dimensions are lower-bounded by some $r > 0$. Then there exists $\epsilon(R, r, \phi) > 0$ and a linear function $\delta = \delta(\mathbf{r})$ such that the logit score $s_{i,j}(\mathbf{r})$ corresponding to $\mathcal{P}_i$ and $\hat{\mathcal{P}}_j(\mathbf{r})$ is a decreasing function of $\|\mathbf{r}\|_2$ for $|\eta| \in [2T\phi, T(\pi - \phi)]$ if the cosine similarity between $\mathbf{q}_{\mathcal{P}_i}$ and $\mathbf{k}_{\mathcal{P}_j}$ is $\geq 1 - \epsilon$.

Proof: Consider a function: $f : \mathbb{R}^{d_{QK}} \times \mathbb{R}^{d_{QK}} \rightarrow \mathbb{R}$ defined as follows:

$$f(\mathbf{x}, \mathbf{y}) = \frac{\frac{d_{QK}}{2} - 1}{\max_{k=0} \arccos(\frac{\mathbf{x}[2k, 2k+1]^\top \mathbf{y}[2k, 2k+1]}{\|\mathbf{x}[2k, 2k+1]\|_2 \|\mathbf{y}[2k, 2k+1]\|_2})} \quad (6)$$

Denote by $L(r) > 0$ the Lipschitz constant of the function f in the region $\{(\mathbf{x}, \mathbf{y}) : \mathbf{x}, \mathbf{y} \in \mathcal{B}(0, R), |\mathbf{x}[i]|, |\mathbf{y}[i]| \geq r; i = 0, 1, \dots, d_{QK} - 1\}$. Note that since $r > 0$, $L(r)$ is well defined (i.e. is finite). Straightforward trigonometric analysis

then leads to the conclusion that if $\epsilon > 0$ is chosen to be small enough such that: $L(r) \cdot 2R \sin(\frac{\arccos(1-\epsilon)}{2}) \leq \phi$ then the following holds: each angle $\alpha_k \in [0, \pi]$ between $\mathbf{q}_{\mathcal{P}_i}[2k, 2k+1]$ and $\mathbf{k}_{\mathcal{P}_j}[2k, 2k+1]$ for $k = 0, 1, \dots, \frac{d_{QK}}{2}$ is upper-bounded by ϕ . Define $\delta(\mathbf{r})$ as follows: $\delta(\mathbf{r}) = \frac{1}{T} \frac{\mathbf{r}^\top \mathbf{r}_{i,j}}{\|\mathbf{r}_{i,j}\|_2^2}$ (in other words: $\delta(\mathbf{r}) = \frac{1}{T} \eta$). But then for $|\eta| \in [2T\phi, T(\pi - \phi)]$, the cos-expression from Eq. 5 is a decreasing function of $\|\mathbf{r}\|_2$ (note that the vanishing cos-expression corresponds to the regular logit score $s_{i,j} = \mathbf{q}_{\mathcal{P}_i}^\top \mathbf{k}_{\mathcal{P}_j}$, as discussed above). That completes the proof. ■

Lemma 3.2 quantifies the intuition that high attention scores between tokens for which the corresponding sub-point clouds are close do not change much when the STRING-based modulation is applied, whereas those for pairs of tokens with larger distances can significantly decrease. Lemma 3.2 does it in the curated setting with one query-key pair and varying distances via the translation operation, whereas in the most general setting, we have many pairs and the analysis is more convoluted. Note however that the fact that angles α_k can correspond to the learnable (rather than canonical, as in the proof of Lemma 3.2) coordinate system $\mathcal{L}$ (see: Lemma 3.1) makes δ learning easier. Even random $\mathcal{L}$ are useful, since they tend to distribute the norm of the queries/keys uniformly across different dimensions. This leads to the less constrained statements (for instance, the assumption about the lower bound on the absolute values of query/key dimensions is not needed).

IV. EVALUATION

We utilized the ALOHA 2 model from Mujoco Menagerie [45, 58] for our simulation environment, the same model used in ALOHA Unleashed [66]. This is a bimanual tabletop environment with two arms attached on either side of a table. In simulation, we use a two-camera setup with front-left and back-right views. These cameras are placed on opposite corners of the table, 70 cm above the surface, and pointed at the center of the table. In the real world, we use a four-camera setup consisting of the two views used in simulation and a wrist-mounted camera on each robot arm. All cameras record RGB-D images. The simulation and real robots were identical. We follow the same simulation evaluation protocol as in [66]: we perform 50 evaluations per checkpoint every 40k training steps and report the maximum evaluation score across all checkpoints. We also report the maximum success rate across all checkpoints for real-world experiments, and we evaluated at most two checkpoints per method per task.

A. Evaluation Tasks

Simulation. Can Opener in Caddy: Six objects are placed on the table: a caddy, screwdriver, magnifier, can opener, thumb drive, and scissors. They are evenly spaced in the 2×3 grid. The caddy, screwdriver, and magnifier may appear in any of the three top positions, while the can opener, thumb drive, and scissors may appear in any of the three bottom positions. Objects are randomly rotated by ± 90 degrees. The robot must grasp the can opener and place it in the left caddy.

Screwdriver in Caddy: Same setup as Can Opener in Caddy, but objects are randomly rotated by ± 180 degrees. The robot must grasp the screwdriver and place it in the right caddy.

Pen in Container: Six objects are placed on the table: a plate, bowl, container, mug, pen, and banana. They are also evenly spaced in a 2×3 grid. The plate, bowl, and container may appear in any of the three top positions, while the mug, pen, and banana may appear in any of the three bottom positions. Objects are randomly rotated by ± 180 degrees. The robot must grasp the pen and place it in the container.

Mug on Plate: Same setup as Pen in Container. The robot must grasp the mug and place it on the plate.

Plate on Rack: Two objects are placed on the table: a plate and drying rack. The drying rack is placed in the top half of the workspace, and the plate in the bottom half of the workspace. The drying rack’s z-axis orientation is randomly varied by ± 86.4 degrees. The robot must grasp the plate and place it into one of the dividers in the drying rack.

Drawer Open: A drawer is randomly placed in the workspace and rotated by ± 45 degrees. The robot must open the bottom drawer.

Real-World. Open Drawers: a drawer is randomly placed in the workspace and rotated by ± 90 degrees. The robot must open both the top and bottom drawers, with partial success awarded if the bottom drawer is opened.

Close Drawers: Same setup as Open Drawers, but with the drawers already open. The robot must close both the top and bottom drawers, with partial success awarded if the top drawer is closed.

Mug on Plate: Same setup as the Mug on Plate task in simulation. The robot must grasp the mug and place it on the plate, with partial success awarded for a successful grasp of the mug.

Plate on Rack: Same setup as the Plate on Rack task in simulation, but the drying rack’s z-axis orientation is randomly varied by ± 45 degrees. The robot must grasp the plate and place it into one of the dividers in the drying rack, with partial success awarded for a successful grasp of the plate.

Recycle Cans: Two trash bins are placed at the top-right of the workspace, and three soda cans are placed in the center. The cans are randomly rotated by ± 45 degrees. The robot must grasp at least two cans and place them in the bins, with partial success awarded for grasping and placing one can.

See Figure 8 for examples of the initial configurations for each real-world task.

B. Datasets

The datasets are collected using a combination of a VLA policy in simulation and teleoperation in real, as described in the following sections.

Sim Data. The sim data was collected using a pre-trained VLA (vision-language-action) model. Specifically, we used Gemini Robotics 1.5 (GR1.5) [44], which takes camera images, a text command, and robot proprioception as inputs and outputs target joint positions. It generated data across 48 tasks

TABLE I: Results from simulation experiments comparing CLAMP against five baselines, averaged over three training seeds. The ‘CLAMP Encoders Only’ variants pre-train only the encoders, while the policy is fine-tuned from scratch.

Method	Can Opener in Caddy	Screwdriver in Caddy	Pen in Container	Mug on Plate	Plate on Rack	Drawer Open
VLA Backbone	7.3%	38.0%	2.0%	14.7%	56.7%	33.3%
ALOHA Unleashed	68.0%	78.0%	16.0%	54.0%	76.0%	80.7%
Pretrained ALOHA Unleashed	70.0%	86.7%	20.7%	62.0%	76.7%	88.7%
ACT	34.0%	82.0%	9.3%	38.7%	70.0%	44.7%
Pretrained ACT	7.3%	20.0%	0.7%	6.0%	29.3%	16.7%
ACT with CLAMP Encoders Only	72.7%	92.0%	10.7%	62.7%	88.7%	28.0%
ALOHA Unleashed with CLAMP Encoders Only	71.3%	83.3%	34.0%	86.0%	88.7%	82.7%
ALOHA Unleashed with CLAMP	94.0%	98.0%	51.3%	95.3%	92.7%	95.3%

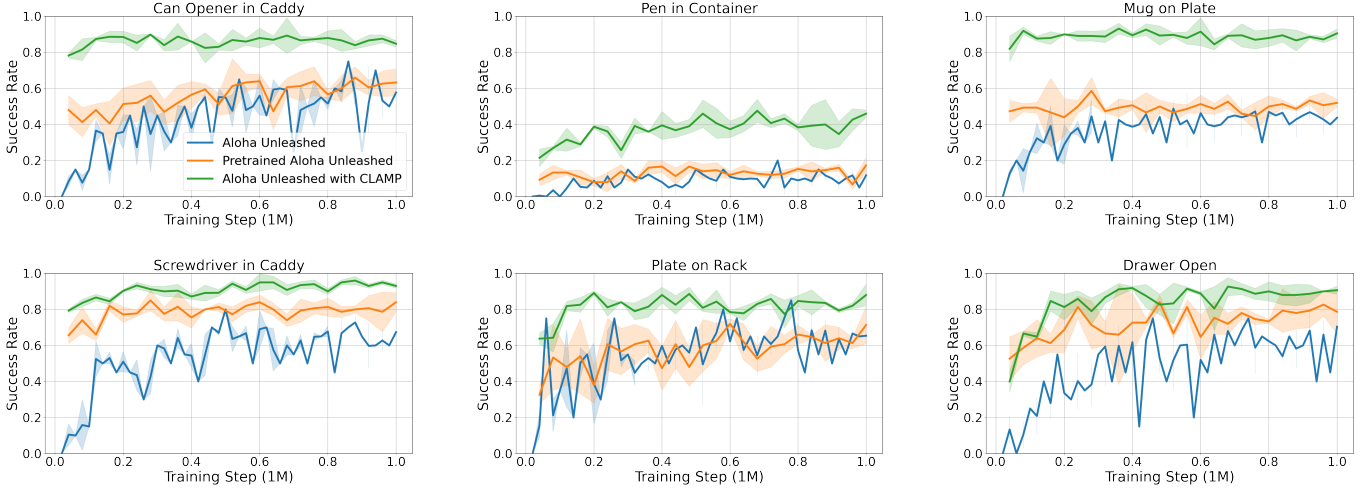


Fig. 4: Evaluation curves comparing our method (green) to baselines across three seeds in simulation. All methods are trained on 4500 demonstrations for 1M environment steps and evaluated for 50 trials per checkpoint every 40K steps.

from which we only kept episodes that successfully completed the task. It averaged 11,533 successful episodes per task for a total of 553,592 episodes to be used for pre-training our encoders (though not uniformly distributed across tasks, see Appendix VII-E). Additionally we generated 4,500 episodes for 6 of the tasks to be used for fine-tuning the policies.

In total, 358,510,758 image-text-action triplets were used for pre-training, while each method was fine-tuned on a single task using 4,500 demonstrations. Note that the pre-training tasks do not overlap with the fine-tuning tasks. See Appendix VII-E for a list of pre-training tasks and count of successful episodes per task.

Real Data. Experienced operators used the teleoperation system [66] to collect demonstrations: 500 each for Open Drawers and Close Drawers, 100 each for Mug on Plate and Plate on Rack, and 63 for Recycle Cans.

C. Baselines

We compare CLAMP against strong baseline methods: **VLA Backbone**, **ALOHA Unleashed** [66], **Pretrained ALOHA Unleashed**, **Action Chunking with Transformers (ACT)** [65], and **Pretrained ACT**. VLA Backbone is a VLA model with the π_0 architecture [2], pre-trained on the same dataset used for CLAMP pre-training. Note that this model is trained from scratch and does not use pretrained weights from Physical Intelligence, the company that produced π_0 . ALOHA

Unleashed and ACT are state-of-the-art imitation learning methods for bimanual manipulation, and ‘Pretrained’ denotes pre-training on the same dataset as CLAMP.

V. RESULTS

A. Sim Results

Table 1 reports the maximum success rate across all checkpoints in simulation, averaged over three training seeds, with evaluations conducted over 50 trials every 40k steps. The top five rows correspond to baseline methods, while the bottom three rows show the same baselines with CLAMP pre-training. ACT with CLAMP pre-training outperforms its non-pretrained counterpart on five of the six tasks. All variants of ALOHA Unleashed with CLAMP pre-training outperform the non-pretrained ALOHA Unleashed across all tasks. Notably, ‘ALOHA Unleashed with CLAMP’ significantly surpasses all baselines on every task. On the whole, methods with CLAMP pre-training exhibit substantial performance gains, demonstrating the benefits of leveraging learned image and action representations during fine-tuning. Moreover, we show that pre-training the encoders and the policy in parallel further improves fine-tuning performance.

Figure 4 illustrates the evaluation curves of our method (green line) and the baselines, averaged over three training seeds. ALOHA Unleashed with CLAMP pre-training shows

TABLE II: Real-world experiment results comparing CLAMP with baseline.

Method	Open Drawers		Close Drawers		Mug on Plate		Plate on Rack		Recycle Cans	
	Partial	Success	Partial	Success	Partial	Success	Partial	Success	Partial	Success
ALOHA Unleashed	2 / 10	0 / 10	8 / 10	0 / 10	4 / 10	3 / 10	8 / 10	8 / 10	9 / 10	5 / 10
ALOHA Unleashed with CLAMP	8 / 10	7 / 10	8 / 10	5 / 10	8 / 10	8 / 10	9 / 10	8 / 10	10 / 10	6 / 10

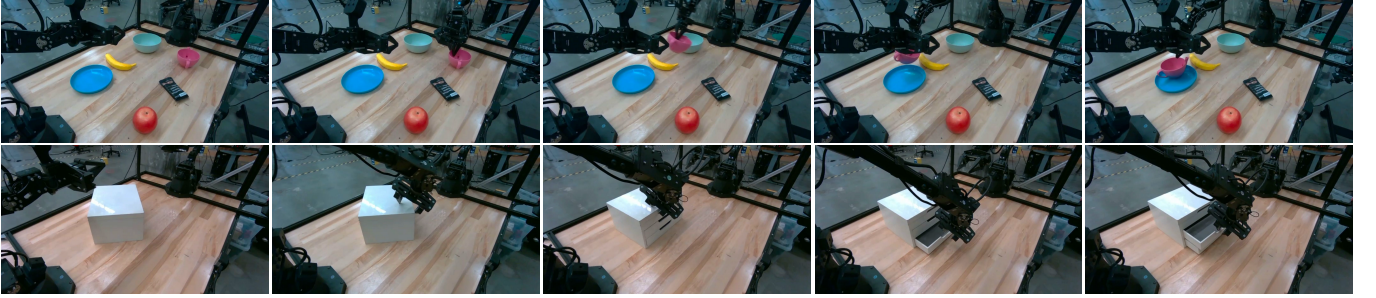


Fig. 5: Successful rollouts of our method on the ALOHA robot. Top: Mug on Plate. Bottom: Open Drawers.

TABLE III: Ablation experiment results in simulation, averaged over three seeds.

Variation	Can Opener in Caddy	Pen in Container	Mug on Plate
Without virtual wrist views	-2.6%	-22.0%	-28.6%
Standard virtual five views [37]	-9.3%	-9.3%	-6.6%
Pre-training w/o text encoder	+0.7%	-2.0%	0.0%
Replace ViT [8] with DP3 [60]	-15.3%	-15.3%	-7.3%
Without STRING [40]	-2.0%	-3.3%	-4.6%
Without pre-training encoders	-9.3%	-7.3%	-4.6%
Unfrozen encoders (fine-tuning)	-23.3%	-16.0%	-3.3%
ACT with CLAMP	-67.4%	-10.7%	-59.4%

strong performance at very early training stages, indicating significant gains in sample efficiency and overall performance due to pre-training. On challenging tasks such as Pen in Container, its performance continues to improve with more training.

B. Real-World Results

Real-world results are reported in Table II, with example rollouts of our method shown in Figure 5. Our method outperforms the baseline on four out of five tasks, while maintaining comparable performance on Plate on Rack. For Open Drawers and Close Drawers, our method successfully opens and closes both drawers, whereas the baseline struggles with the final drawer. In Mug on Plate, our method shows noticeable improvements in grasping and placing the mug compared to the baseline. For Plate on Rack and Recycle Cans, we observe slightly better grasping performance, while placement performance remains comparable to the baseline. We hypothesize that these tasks are relatively simple and the training data less diverse, limiting gains over the baseline.

C. Ablations

We ablate key design choices of CLAMP in Table III. Experiments are conducted in simulation on Can Opener in Caddy, Pen in Container, and Mug on Plate, and

results are averaged over three training seeds. As shown in the first two rows, incorporating virtual wrist views substantially improves performance, especially on high-precision tasks such as Pen in Container and Mug on Plate. Moreover, the standard virtual five-view setup used in related work (e.g., [37]) underperforms our virtual five-view configuration.

In row three, we show that including a text encoder during pre-training yields slightly better performance in simulation, but we observe that the text encoder noticeably improves real-world performance. We hypothesize that this discrepancy stems from differences between the environments. In simulation, object shapes are consistent and noise-free, so the model relies less on textual information describing the objects. In contrast, objects captured by real-world cameras exhibit texture variations, lighting changes, and other forms of visual noise, making textual information more valuable.

In the following, we ablate the image encoder. Replacing the ViT encoder with a point-cloud encoder (DP3) results in worse performance. Removing STRING positional encoding from ViT, which eliminates learned correlations between tokens from different image patches, also degrades performance. As shown in rows six and seven, it is preferable to pre-train the encoders and keep them frozen during fine-tuning. Finally, the last row shows that ACT with CLAMP performs poorly; we hypothesize this is due to ACT's limited ability to effectively learn from multimodal demonstrations.

D. Simulation Pre-training Results

Figure 6 shows the retrieval results of the CLAMP encoders on six multimodal retrieval tasks on the validation examples. To ensure diversity and accurate assessment, we select examples from the validation set using the following criteria: (1) the time step is a multiple of 10, and (2) the workspace is evenly divided into a 3×3 grid, and the first object's position has not previously appeared in any grid cell. This results in a total of 3216 examples. As shown in the figure, the CLAMP encoders perform particularly well on tasks that involve retrieving the previous action, highlighting the potential usefulness of

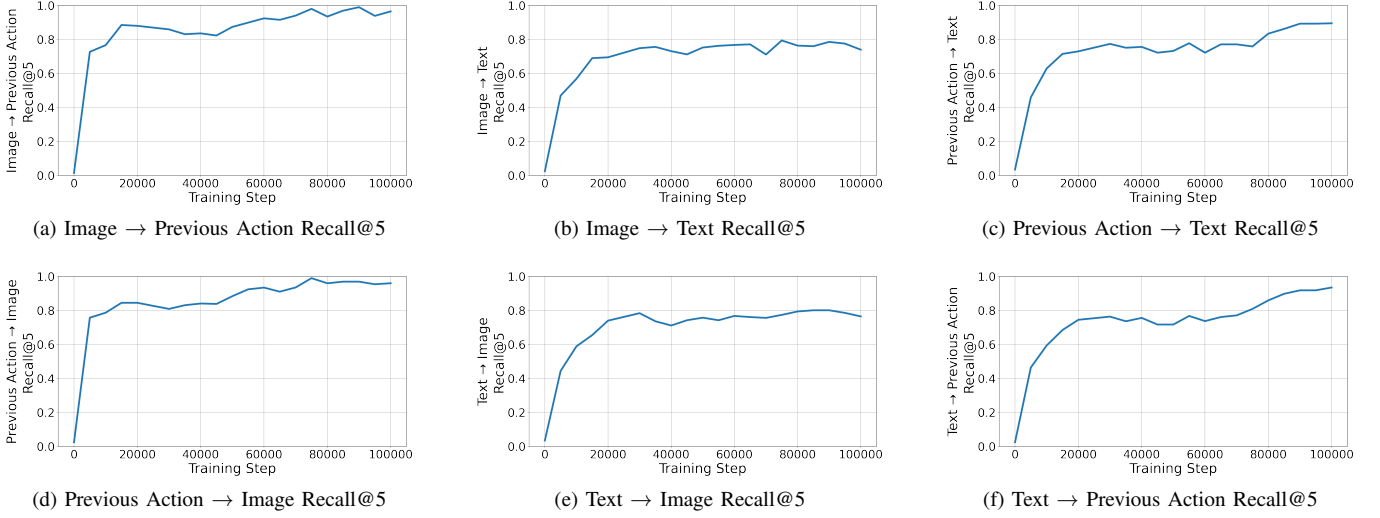


Fig. 6: Recall@5 results for cross-modal retrieval tasks on the validation data during pre-training.



Fig. 7: SigLIP pairwise matching probabilities on an unseen task: image-to-previous-action (top) and previous-action-to-image (bottom). CLAMP correctly retrieves all image-previous-action pairs on both tasks.

the learned action representation. To further examine this representation, we visualize the SigLIP pairwise matching probabilities for image-to-previous-action and previous-action-to-image retrieval on an unseen task in [Figure 7](#). The model correctly retrieves all examples for both tasks, suggesting that the learned image and action representations generalize and can transfer to unseen tasks.

VI. CONCLUSION

In this paper, we introduce CLAMP, a 3D action-conditioned pre-training framework for robotic manipulation.

CLAMP contrastively pre-trains image, text, and action encoders using 3D multi-view observations, simulator-derived state descriptions, and action histories. The learned image and action representations are then used to fine-tune a Diffusion Policy, which is pre-trained in parallel with the encoders, from limited task demonstrations. Across simulation and real-world experiments, CLAMP substantially improves fine-tuning sample efficiency and final performance over state-of-the-art pretrained and non-pretrained baselines.

REFERENCES

- [1] Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Big vision. https://github.com/google-research/big_vision, 2022.
- [2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmaail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0 : A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*, 2024.
- [3] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL <http://github.com/google/jax>.
- [4] 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*, pages 24185–24198, 2024.
- [5] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.
- [6] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.
- [7] 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*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- [8] 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. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>.
- [9] 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*, pages 19358–19369, 2023.
- [10] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.
- [11] Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. In *Conference on Robot Learning*, pages 3949–3965. PMLR, 2023.
- [12] Ankit Goyal, Hei Law, Bowei Liu, Alejandro Newell, and Jia Deng. Revisiting point cloud shape classification with a simple and effective baseline. In *International conference on machine learning*, pages 3809–3820. PMLR, 2021.
- [13] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*, pages 694–710. PMLR, 2023.
- [14] Christian Graf, David B Adrian, Joshua Weil, Miroslav Gabriel, Philipp Schillinger, Markus Spies, Heiko Neumann, and Andras Gabor Kupcsik. Learning dense visual descriptors using image augmentations for robot manipulation tasks. In *conference on Robot Learning*, pages 871–880. PMLR, 2023.
- [15] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational visual media*, 7(2): 187–199, 2021.
- [16] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2021.
- [17] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Voint cloud: Multi-view point cloud representation for 3d understanding. *arXiv preprint arXiv:2111.15363*, 2021.
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.
- [19] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5693–5702, 2021.
- [20] Haigen Hu, Xiaoyuan Wang, Yan Zhang, Qi Chen, and Qiu Guan. A comprehensive survey on contrastive learning. *Neurocomputing*, 610:128645, 2024.
- [21] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023.
- [22] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer*

- Vision and Pattern Recognition*, pages 15524–15533, 2022.
- [23] Xiaoshui Huang, Zhou Huang, Sheng Li, Wentao Qu, Tong He, Yuenan Hou, Yifan Zuo, and Wanli Ouyang. Frozen clip transformer is an efficient point cloud encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2382–2390, 2024.
 - [24] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. URL <https://arxiv.org/abs/2504.16054>.
 - [25] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.
 - [26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.
 - [27] Yueru Jia, Jiaming Liu, Sixiang Chen, Chenyang Gu, Zhilve Wang, Longzan Luo, Xiaoqi Li, Pengwei Wang, Zhongyuan Wang, Renrui Zhang, et al. Lift3d policy: Lifting 2d foundation models for robust 3d robotic manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17347–17358, 2025.
 - [28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1412.6980>.
 - [29] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosior, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019.
 - [30] Sung-Wook Lee, Xuhui Kang, Brandon Yang, and Yen-Ling Kuo. Class: Contrastive learning via action sequence supervision for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2025.
 - [31] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022.
 - [32] Zijing Ma, Zhi Yang, Aihua Mao, Shuyi Wen, Ran Yi, and Yongjin Liu. A multi-view projection-based object-aware graph network for dense captioning of point clouds. *Computers & Graphics*, 126:104156, 2025.
 - [33] Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, et al. Tips: Text-image pretraining with spatial awareness. *arXiv preprint arXiv:2410.16512*, 2024.
 - [34] NVIDIA, :, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llonet, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots, 2025. URL <https://arxiv.org/abs/2503.14734>.
 - [35] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.
 - [36] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.
 - [37] Shengyi Qian, Kaichun Mo, Valts Blukis, David F Fouhey, Dieter Fox, and Ankit Goyal. 3d-mvp: 3d multi-view pretraining for manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22530–22539, 2025.
 - [38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
 - [39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
 - [40] Connor Schenck, Isaac Reid, Mithun George Jacob, Alex Bewley, Joshua Ainslie, David Rendleman, Deepali Jain, Mohit Sharma, Avinava Dubey, Ayzaan Wahid, et al. Learning the ropes: Better 2d and 3d position encodings with string. *arXiv preprint arXiv:2502.02562*, 2025.
 - [41] Tanner Schmidt, Richard Newcombe, and Dieter Fox. Self-supervised visual descriptor learning for dense cor-

- responsiveness. *IEEE Robotics and Automation Letters*, 2 (2):420–427, 2016.
- [42] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.
- [43] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.
- [44] Gemini Robotics Team. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer, 2025. URL <https://arxiv.org/abs/2510.03342>.
- [45] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.
- [46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [47] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019.
- [48] Yufei Wang, Ziyu Wang, Mino Nakura, Pratik Bhowal, Chia-Liang Kuo, Yi-Ting Chen, Zackory Erickson, and David Held. Articubot: Learning universal articulated object manipulation policy via large scale simulation. *arXiv preprint arXiv:2503.03045*, 2025.
- [49] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022.
- [50] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4840–4851, 2024.
- [51] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer, 2024.
- [52] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1179–1189, 2023.
- [53] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27091–27101, 2024.
- [54] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 483–498, 2021.
- [55] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.
- [56] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19313–19322, 2022.
- [57] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.
- [58] K. Zakka, Y. Tassa, and MuJoCo Menagerie Contributors. MuJoCo Menagerie: A collection of high-quality simulation models for MuJoCo. http://github.com/google-deepmind/mujoco_menagerie, 2022. URL http://github.com/google-deepmind/mujoco_menagerie. Accessed: 2025-12-15.
- [59] Yanjie Ze, Ge Yan, Yueh-Hua Wu, Annabella Macaluso, Yuying Ge, Jianglong Ye, Nicklas Hansen, Li Erran Li, and Xiaolong Wang. Gnfactor: Multi-task real robot learning with generalizable neural feature fields. In *Conference on robot learning*, pages 284–301. PMLR, 2023.
- [60] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024.
- [61] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.
- [62] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Pointm2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in neural information processing systems*, 35:27061–27074, 2022.
- [63] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer*

- vision and pattern recognition*, pages 8552–8562, 2022.
- [64] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.
 - [65] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. 2023.
 - [66] Tony Z Zhao, Jonathan Tompson, Danny Driess, Pete Florence, Kamyar Ghasemipour, Chelsea Finn, and Ayzaan Wahid. Aloha unleashed: A simple recipe for robot dexterity. *arXiv preprint arXiv:2410.13126*, 2024.
 - [67] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023.
 - [68] Wenxuan Zhou, Bowen Jiang, Fan Yang, Chris Paxton, and David Held. Hacman: Learning hybrid actor-critic maps for 6d non-prehensile manipulation. *arXiv preprint arXiv:2305.03942*, 2023.
 - [69] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2639–2650, 2023.