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

MLA: A Multisensory Language-Action Model for Multimodal Understanding and Forecasting in Robotic Manipulation

Zhuoyang Liu[1*], Jiaming Liu[1*†], Jiadong Xu[1], Nuowei Han[1], Chenyang Gu[1], Hao Chen[3], Kaichen Zhou[1], Renrui Zhang[3], Kai Chin Hsieh[1], Kun Wu[2], Zhengping Che[2†], Jian Tang[2], Shanghang Zhang[1✉]

***Abstract*— Vision-language-action models (VLAs) have shown generalization capabilities in robotic manipulation tasks by inheriting from vision-language models (VLMs) and learning action generation. Most VLA models focus on interpreting vision and language to generate actions, whereas robots must perceive and interact within the spatial-physical world. This gap highlights the need for a comprehensive understanding of robotic-specific multisensory information, which is crucial for achieving complex and contact-rich control. To this end, we introduce a multisensory language–action (MLA) model that collaboratively perceives heterogeneous sensory modalities and predicts future multisensory objectives to facilitate physical world modeling. Specifically, to enhance perceptual representations, we propose an encoder-free multimodal alignment scheme that innovatively repurposes the large language model itself as a perception module, directly interpreting multimodal cues by aligning 2D images, 3D point clouds, and tactile tokens through positional correspondence. To further enhance MLA's understanding of physical dynamics, we design a future multisensory generation post-training strategy that enables MLA to reason about semantic, geometric, and interaction information, providing more robust conditions for action generation. For evaluation, the MLA model outperforms the previous state-of-the-art 2D and 3D VLA methods by 12% and 24% in complex, contact-rich real-world tasks, respectively, while also demonstrating improved generalization to unseen configurations. Project website: `https://robotic-mla.github.io/`**

I. INTRODUCTION

Recent robot imitation learning has achieved remarkable advances in training policies from expert demonstrations to perform diverse vision–language manipulation tasks. Meanwhile, vision–language models (VLMs) [1], [2] pre-trained on internet-scale data have been proven to possess strong capabilities in common-sense reasoning in general scenarios. Building on these progresses, vision–language–action (VLA) models have been proposed [3], [4], [5], which not only inherit the properties of VLMs but also extend them by training with robot demonstrations for action prediction. As a result, VLA models demonstrate impressive generalization and precise manipulation, effectively mapping human instructions and visual observations to the robot control signal.

In the real world, robots must perceive spatial environments, reason about semantic relationships, and interact with dynamic environment configurations. However, most existing VLA models rely primarily on 2D image integration [6], [3], which is fundamentally inadequate for capturing spatial dependencies and modeling physical dynamics. **On the one hand**, to address these limitations, several studies enhance VLAs with richer multimodal observations. Specifically, some approaches incorporate 3D inputs to improve geometric scene understanding [7], [8], while others introduce tactile signals to capture interaction feedback from manipulated objects [9], [10]. Existing VLA models often require modality-specific encoders to enrich perceptual capacity, which undermines efficiency. Furthermore, without pre-training on multisensory inputs, the large language model (LLM) backbone of VLAs exhibits limited representation to align with the newly introduced multimodal features. **On the other hand**, several VLA studies attempt to reason about the physical dynamics by predicting future states, such as subgoal images

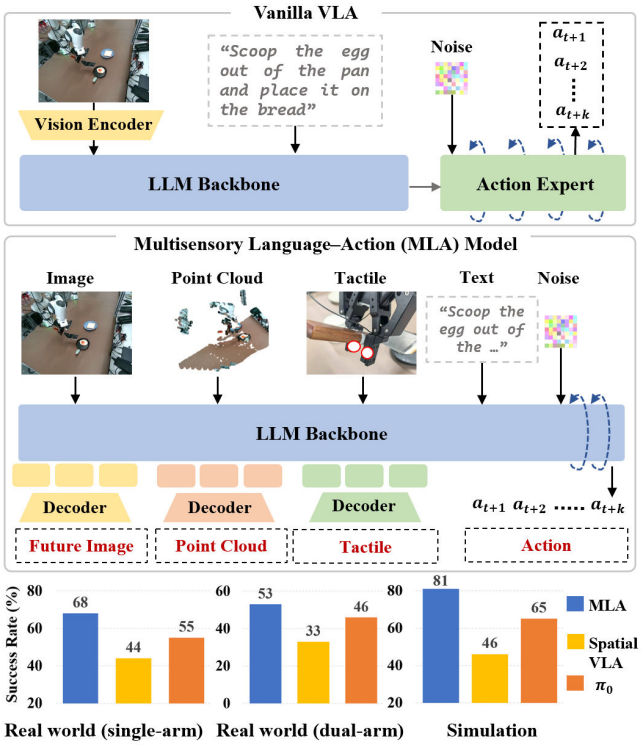

Fig. 1: (a) Unlike vanilla VLA methods that rely on 2D images and natural language instructions to generate actions, (b) we propose MLA, a multisensory language–action model that collaboratively processes diverse robotic-specific modalities and predicts their future states to enhance physical dynamics modeling in robotic control. (c) MLA achieves state-of-the-art performance across a variety of real-world and simulation tasks.

*Equal contribution. †Project lead. [1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University. [2]Beijing Innovation Center of Humanoid Robotics (X-Humanoid). [3]The Chinese University of Hong Kong (CUHK).

and camera-view depth maps [11], [12], [13]. However, these approaches remain limited in predicting complete point cloud structures and tactile interaction information, which are essential not only for understanding complex, contact-rich scenes but also for effective motion planning in robotic manipulation. Consequently, a critical question arises: how can multisensory modalities be integrated into a unified representation and predicted in their future states to collaboratively enhance VLA models' physical-world understanding and action generation?

To address this question, we propose **MLA**, a multisensory language–action model that collaboratively processes diverse sensory inputs and predicts their corresponding future states to enhance physical-world modeling for robotic control. As shown in Figure 1, to avoid introducing additional modality-specific encoders that lack pretraining alignment with LLM's embeddings, MLA adopts an encoder-free multimodal alignment mechanism, repurposing the initial transformer blocks of the LLM as a perception module to directly interpret visual, geometric, and tactile cues. In particular, we project 3D points and the spatial positions of the tactile gripper onto 2D image planes using camera parameters, thereby constructing cross-modal positional mappings. These positional correspondences serve as positive pairs for token-level contrastive learning, aligning multimodal features within the LLM's embedding space. This position-guided consistency constraint enhances the multimodal representations of our MLA model and supports more comprehensive physical-world perception. To further enhance the LLM's understanding of physical robotic scenes, we propose a future multisensory generation post-training strategy. Specifically, the lightweight transformer-based decoders and tailored generation scheme are designed to process the LLM's final-layer features and generate the future states of multiple modalities, including 2D images, 3D point clouds, and tactile signals. Through this predictive process, MLA is able to reason about physical dynamics from multiple dimensions, encompassing semantic information, geometric structures, and object-centric interactions. Notably, the proposed methods are applied only during training and do not affect MLA's inference efficiency, while enriching feature conditions for action generation.

Since existing open-source real-world datasets [14], [15], [16] lack multisensory information, we pretrain the LLM solely on large-scale image–action paired datasets following common practice [5], [4], including more than 570K trajectories. Subsequently, we perform supervised fine-tuning (SFT) on downstream task datasets using the proposed encoder-free multimodal alignment mechanism, and finally conduct future multisensory generation post-training, progressively equipping our model with the ability to integrate perception, understanding, and action generation from multisensory inputs in the real physical world. To systematically evaluate our model, we design six complex, contact-rich real-world robotic experiments covering both single- and dual-arm manipulation tasks, where MLA achieves state-of-the-art success rates and demonstrates strong generalization to unseen objects and backgrounds. For reproducibility, we further evaluate MLA on the RLBench [17] simulator and also obtain competitive performance. As tactile sensing in simulation is not realistic, we incorporate tactile signals only in real-world experiments. Our contributions are summarized as follows:

- We propose MLA, a multisensory language-action model with an encoder-free multimodal alignment mechanism, repurposing the LLM itself to directly align with and interpret image, point cloud, and tactile cues.
- To further strengthen MLA's understanding of physical dynamics, we introduce a future multisensory generation post-training strategy that enables it to reason about semantic, geometric, and interaction information, providing more robust conditions for action generation.
- Through a progressive pipeline of pretraining, SFT, and post-training, MLA achieves state-of-the-art success rates and strong generalization on complex real-world tasks, including both single- and dual-arm manipulation.

II. RELATED WORK

Vision-language-action (VLA) models [5], [3], [6], [8], [4] have advanced rapidly with the development of vision-language models (VLMs) and large-scale robotic datasets. PaLM-E [18] pioneered the adaptation of VLMs to robotic data, and subsequent works followed this paradigm, further extending its capabilities [5], [3]. Meanwhile, diffusion and flow modeling have emerged as effective tools for modeling the multimodal distributions of robotic actions [19], [20]. This has motivated approaches that condition continuous action experts on VLM representations [6], [21], [3], as well as recent scaling efforts using transformer-based diffusion architectures [22], [23], [4]. Moreover, some studies further enhance action generation by incorporating richer sensory inputs, such as 3D point clouds [8], [24] and tactile signals [9], [10]. These methods often require modality-specific encoders which undermines efficiency. Also, without multisensory pretraining, the LLM backbone struggles to align and interpret them efficiently.

Robotic world knowledge forecasting policy, which predicts and reasons about future world knowledge, has gained considerable attention in robotics for its ability to capture the dynamics of the physical environment. Early attempts [25], [26] employed generative models to directly predict future images, and then leveraged the learned representations to train an action generator. Subsequent work [11], [27], [28] explored the use of latent action tokens as forward-dynamics representations for action planning and generation. Another line of VLA research [29], [13], [30], [12] focuses on leveraging future state prediction to facilitate action generation. While these methods are confined to 2D image prediction and struggle with complex, contact-rich scenes, MLA introduces comprehensive multisensory forecasting for robotics, yielding more robust representations for action generation.

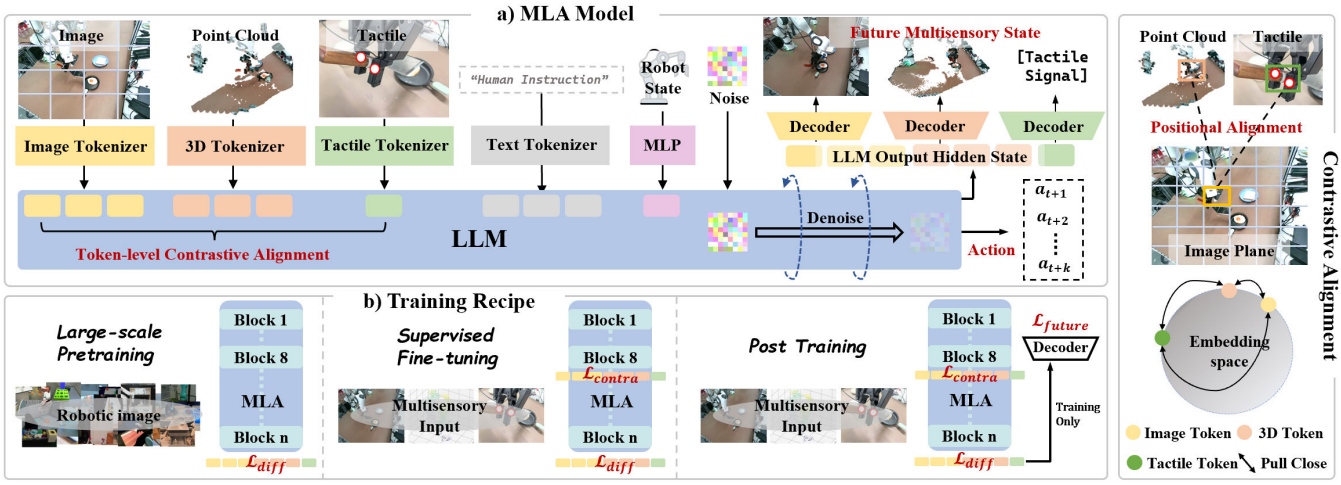

Fig. 2: **Overall Framework of MLA.** a) Beyond language instructions and robot states, MLA introduces an innovative encoder-free multimodal alignment mechanism that directly enables the LLM to integrate RGB images, point clouds, and tactile signals, aligning them through token-level contrastive learning. Furthermore, MLA incorporates a future multisensory generation post-training strategy, allowing the model to generate future multisensory states and providing more robust conditions for action generation. b) MLA adopts a three-stage training paradigm: large-scale pretraining, supervised fine-tuning with cross-modal alignment, and post-training with future state prediction.

III. METHOD

A. Preliminary

Similar to the VLA problem [3], [6], our MLA imitation learning is formulated as a probabilistic sequence decision-making task. At each timestep t, the policy π_θ takes multimodal inputs, including the image observation I_t, point cloud P_t, tactile signal T_t, robot state S_t, and language instruction L. It then predicts both the immediate action sequence $a_{t:t+H}$ and the future keyframe observations across modalities $I_{t+N}, P_{t+N}, T_{t+N}$. Formally, the generative process is expressed as:

$$a_{t:t+H}, I_{t+N}, P_{t+N}, T_{t+N} \sim \pi_\theta\big(\cdot \mid I_t, P_t, T_t, S_t, L\big).$$

We follow the experimental setup with Franka single- and dual-arm configurations and represent actions as end-effector pose [5], [4]. In the single-arm setting, each action is a 7-DoF vector $a_t = (\Delta x, \Delta y, \Delta z, R_r, R_p, R_y, g)$, where $\Delta x, \Delta y, \Delta z$ denote the Cartesian position delta, R_r, R_p, R_y correspond to the Euler angles for rotation, and g is the gripper width. In the dual-arm configuration, the action is represented by concatenating two 7-DoF vectors into a 14-DoF representation.

B. MLA Architecture

As shown in Figure 2, our proposed MLA model is built upon a LLM, where the parameters are initialized from the LLM backbone of Prismatic VLM [1], similar to prior work [5]. Distinct from conventional VLA frameworks that employ vision encoders, our model introduces lightweight tokenizers that directly convert raw multisensory inputs into a shared token sequence and repurpose the LLM itself as a unified MLA model. Furthermore, we incorporate transformer-based decoders that predict future multimodal states.

Image Tokenizer. For each input image $I \in \mathbb{R}^{H\times W\times 3}$, our Vision Tokenizer converts it into a compact token sequence. Following previous works [31], [32], the image is divided into non-overlapping patches of size 14×14, yielding a token sequence of length $N_{\text{img}} = 256$ with a batchsize of B and an embedding dimension of $d_h = 4096$, i.e., $f^{\text{img}} \in \mathbb{R}^{B\times N_{\text{img}}\times d_h}$.

3D Point Cloud Tokenizer. Given raw point clouds $P \in \mathbb{R}^{B\times 1024\times 3}$, our 3D Tokenizer partitions the points into local groups centered at sampled anchor points. Following [33], the tokenizer consists of three blocks, each incorporating farthest point sampling (FPS) [34] for downsampling, k-nearest neighbors (KNN) for local aggregation, and a learnable linear layer for feature encoding. After 3D tokenization, we obtain a compact representation consisting of $N_{\text{pc}} = 256$ tokens, $f^{\text{pc}} \in \mathbb{R}^{B\times N_{\text{pc}}\times d_h}$.

Tactile Tokenizer. For tactile sensing, we design a simple MLP-based tokenizer to embed low-dimensional tactile signals into the shared token space. Specifically, we attach two tactile sensors to the gripper fingers. From each sensor, we extract six values: normal force, tangential force, and tangential force direction (two components each). The raw signal is processed by a lightweight MLP, producing a tactile token $f^{\text{tac}} \in \mathbb{R}^{B\times 1\times d_h}$.

LLM Backbone. We adopt LLaMA-2 7B as our base model and repurpose it into a unified perception-and-reasoning policy. Specifically, tokens from images, point clouds, tactile signals, and language are projected into a shared embedding space $f \in \mathbb{R}^{B\times N_t\times 4096}$ and jointly processed by the LLM. The noise tokens required by the diffusion-based action head are appended to the end of the token sequence, enabling the model to perform diffusion modeling. Diffusion noise and timesteps are embedded through MLP projectors. This design eliminates the need

for separate modality-specific encoders and fully leverages the large-scale pretrained LLM to directly interpret robotic-specific multisensory cues and generate robust actions.

Future Prediction Decoder. For future multisensory generation, we adopt transformer decoders to predict future sensory observations from the LLM's final hidden states h. Each decoder maps the unified multimodal embeddings into its target space, such as image, point cloud, and tactile vectors, and is supervised by the corresponding future state. The transformer-based decoder follows a standard query–key–value attention design, consisting of four stacked self-attention and feed-forward layers, enabling effective modeling of the multimodal embeddings.

C. Encoder-Free Multimodal Alignment

Previous VLA models [5], [3] rely on vision encoders that are large-scale pretrained on general-domain data, such as SigLIP [35], to process robotic observations. However, these encoders are rarely trained on robot-domain datasets and have not been exposed to robotic-specific sensors. As a result, their representations are limited in aligning with and interpreting robotic data. In addition, pretraining newly introduced encoders often incurs substantial computational cost, and their incorporation constrains inference efficiency. Inspired by prior works on contrastive learning [35], which adopt a self-supervised approach to align semantic information from heterogeneous modalities into a unified embedding space, we propose an *Encoder-Free Multimodal Alignment* method. This method repurposes the initial transformer blocks of the LLM as a unified perception module via token-level contrastive loss, enhancing multisensory representations without the need for additional modality-specific encoders. In practice, we employ the embedding features from the 8th transformer block for self-supervised alignment and further examine the effect of using different block outputs in our ablation study.

Formulation of Positive and Negative Pairs. For the Transformer-based model, the positional indicators of tokens can provide both positional alignment and semantic contextual alignment [36]. Therefore, we construct cross-multisensory positional mappings to formulate the positive and negative pairs in our token-level contrastive loss. In contrast, directly treating multimodal tokens with misaligned positional information as positive pairs may lead to semantic misalignment. As shown in the right part of Figure 2, we project 3D points and the 3D positions of tactile grippers onto 2D image planes using the camera parameters. Since each 3D point cloud token ($\{f_i^{pc}\}_{i=1}^{N_{pc}}$) is aggregated from a set of 3D points, we unproject its center point into 2D image coordinates. For the tactile token (f^{tac}), we directly read the robot state and project the tactile gripper's position in the 3D world coordinate onto the 2D image plane. Subsequently, we identify the corresponding 2D image patch onto which these features project, and align the 3D token and tactile token with the corresponding 2D token ($\{f_j^{img}\}_{j=1}^{N_{img}}$) as positive pairs ($f_j^{img}$–$f_i^{pc}$–$f^{tac}$), while the remaining unmatched tokens are treated as negative pairs.

Image–Point Cloud Alignment. Since image and 3D embeddings have the same token sequence length (256), we apply a token-level InfoNCE loss to pull positive pairs together in the embedding space and push negative pairs apart, where τ denotes the temperature.

$$\mathcal{L}_{\text{img\_pc}} = -\frac{1}{256}\sum_{i=1}^{256}\log\frac{\exp\left(\langle f_j^{\text{img}}, f_i^{\text{pc}}\rangle/\tau\right)}{\sum_{j=1}^{256}\exp\left(\langle f_j^{\text{img}}, f_i^{\text{pc}}\rangle/\tau\right)}$$

Tactile–Image and Point Cloud Alignment. In the single-arm setting, since tactile embeddings consist of a single token, this yields one positive sample per tactile token $(f^{\text{tac}}, f_j^{\text{img}})$ and $(f^{\text{tac}}, f_i^{\text{pc}})$, while other tokens serve as negatives. A unidirectional contrastive loss is applied to pull the tactile embedding toward its corresponding tokens:

$$\mathcal{L}_{\text{tac\_img/pc}} = -\log\frac{\exp(\langle f^{\text{tac}}, f_{j/i}^{\text{img/pc}}\rangle/\tau)}{\sum_{j/i=1}^{256}\exp(\langle f^{\text{tac}}, f_{j/i}^{\text{img/pc}}\rangle/\tau)}$$

The overall contrastive objective is the sum of the three losses: $\mathcal{L}_{\text{contrastive}} = \mathcal{L}_{\text{img\_pc}} + \mathcal{L}_{\text{tac\_img}} + \mathcal{L}_{\text{tac\_pc}}$. Through this contrastive learning objective, the model effectively captures consistent semantic and spatial information, enabling multimodal features to be seamlessly integrated within the LLM's unified embedding space.

D. Future Multisensory Generation

While some VLA studies [11], [30], [13] adopt future observation prediction to enable models to reason about physical dynamics, they still fall short in forecasting diverse robotic-specific modalities that are essential for fully capturing the semantic, geometric, and interaction information of the physical world. To further enhance MLA's understanding of robotic physical scenes, we propose a future multisensory generation post-training strategy, making the first attempt to jointly forecast the future states of images, point cloud, and tactile modalities that are most relevant for manipulation.

Image Prediction. For the visual stream, we adopt a transformer-based decoder, where the last-layer hidden states of the LLM are injected as input features and the future image generation is supervised with an MSE loss. Unlike previous VLA methods that predict dense future frames (close to the current timestep), MLA predicts future keyframes. Following prior work [33], keyframes are identified based on changes in the robotic joint velocity and action transitions. To ease optimization of the image generation loss, background pixels are removed using the corresponding depth map, restricting prediction to foreground regions.

Point Cloud Prediction. For 3D geometry, we adopt a transformer-based decoder to reconstruct the next-keyframe point cloud. Inspired by the masked autoencoder method for point clouds [37], we partition the ground-truth point cloud into G local patches by sampling G center points with FPS and grouping M neighboring points for each center using KNN. The decoder then outputs the predicted coordinates $\hat{P} \in \mathbb{R}^{G\times M\times 3}$, supervised with Chamfer Distance against the ground truth P. This operation enhances the stability of

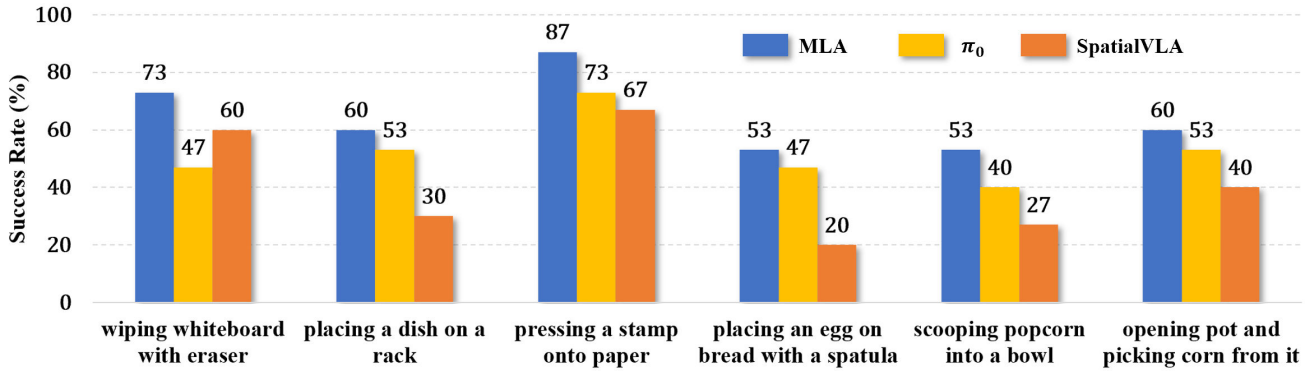

Fig. 3: **Real-world results.** All models are evaluated over 15 rollouts from different manipulated object positions on the tabletop, with task completion determined by human judgment.

future point cloud prediction by first aligning coarse center points to establish the basic 3D structure, and subsequently refining local neighbor points.

Tactile Prediction. For tactile feedback, the decoder outputs a low-dimensional tactile embedding supervised by an MSE loss against the ground truth.

By jointly predicting the future states of images, point clouds, and tactile signals, MLA achieves more comprehensive feature representations across semantic, geometric, and interaction dimensions. It is worth noting that these future-state prediction losses are applied only during the post-training stage and do not affect inference efficiency.

E. Overall Training Recipe

Large-Scale Pretraining. Similar to previous VLA methods [5], we construct a large-scale dataset of over 570K trajectories by combining diverse open-source datasets, such as Open-X-Embodiment [14] and RoboMIND [16]. Since the observations in these datasets primarily consist of image and language modalities, we pretrain MLA using only these inputs for 10 epochs. For the other modalities, we reserve their token positions in the sequence, ensuring a smooth transition to subsequent training stages. For the action generation loss ($\mathcal{L}_{\text{diff}}$), we adopt a standard DDPM objective, minimizing the MSE between the predicted and ground-truth noise.

Supervised Fine-Tuning. The pretrained model is subsequently adapted to high-quality, task-specific datasets using the proposed encoder-free multimodal alignment mechanism. In this stage, all multisensory modalities are introduced, including image, point cloud, tactile signals, and language instruction. We incorporate the proposed contrastive loss (as detailed in Section III-C) to enhance MLA's cross-modal alignment and multimodal representations. The overall training objective loss is $\mathcal{L}_{\text{sft}} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{contrastive}}$.

Post-Training. Finally, the model undergoes future multisensory generation post-training. In this stage, the training data and input modalities are the same as in the SFT phase. Additionally, the training incorporates future multimodal prediction supervision, as described in Section III-D, enabling the model to capture physical dynamics and thereby achieve more robust action generation. The overall supervision is $\mathcal{L}_{\text{post}} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{contrastive}} + \mathcal{L}_{\text{future}}$. Note that we perform SFT followed by post-training to progressively equip our model with the ability to integrate perception, understanding, and action generation from multisensory inputs in the real physical world. During inference, we employ DDIM [38] with n sampling steps (e.g., $n = 4$).

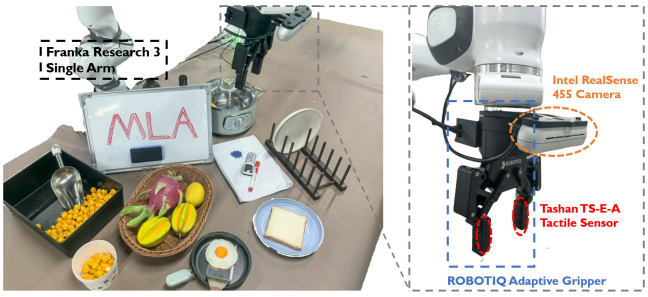

Fig. 4: **Single-arm Experiment Setup.** We show the details about single-arm setup and assets of real-world experiments.

IV. EXPERIMENTS

In Section IV-A, we compare MLA model with recent VLA models on single- and dual-arm real-world tasks. Section IV-B presents an ablation study of each component, while Section IV-C demonstrates MLA's generalization in real-world settings. Section IV-D benchmarks MLA against VLA baselines in the RLBench simulator for reproducibility.

A. Real-World Experiment

Real-World Experiment Setup. We evaluated four complex contact-rich tasks on a single-arm Franka robot and two tasks on a dual-arm setup combining two Franka robots. As shown in Figure 4, for the single arm, two RealSense D455 cameras were used to provide image and point cloud data from a third-person view and a wrist view, with only the third-person view contributing to cross-modal alignment. Each gripper was equipped with two tactile sensors (Tashan TS-E-A). For the dual arm, three D455 cameras were employed, including one third-person view and two wrist views.

Self-collected Data. For the single-arm setting, we designed four contact-rich tasks: (1) pressing a stamp onto paper, (2) wiping a whiteboard with an eraser, (3) placing a dish on a rack, and (4) placing an egg on bread with a spatula. For the dual-arm setting, we evaluated two collaborative tasks: (1) scooping popcorn into a bowl and (2) opening

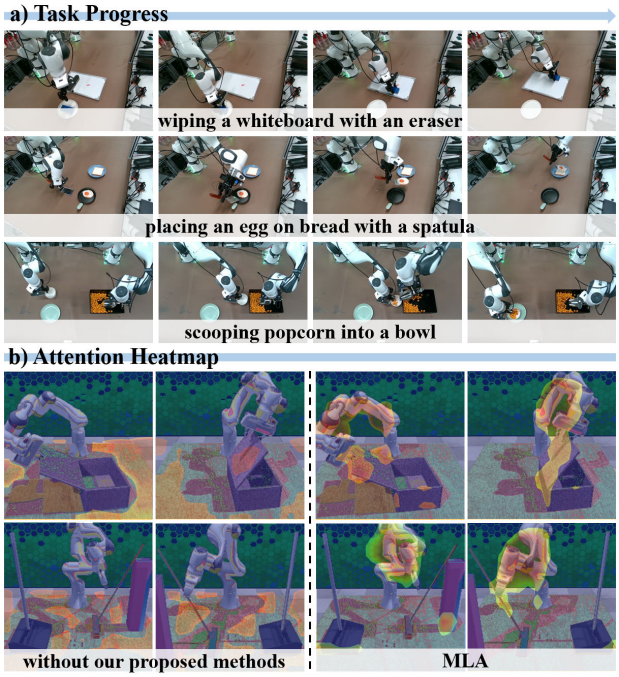

Fig. 5: **Visualization** of real-world task progress and attention heatmaps from the final-layer output features of MLA.

a pot lid and picking corn from the pot. All demonstrations were collected using the Gello [39] platform, with 200 high-quality demonstrations per task.

Training and Evaluation Details. We train MLA for 300 epochs during SFT and 100 epochs during post-training using the AdamW optimizer. Baselines are initialized with their pretrained parameters and fine-tuned under their respective protocols. We compare against two closely related baselines: π_0 [3], a state-of-the-art 2D VLA model, and SpatialVLA [8], a state-of-the-art 3D VLA model. All models use the same number of camera viewpoints, and each task is evaluated with 15 rollouts under consistent test conditions.

Results. As shown in Figure 3, MLA achieves superior performance across six tasks, outperforming π_0 and SpatialVLA by an average of 12% and 24%, respectively. In the Wiping a Whiteboard task, MLA effectively leverages tactile sensing to regulate the downward and lateral movements of the end effector during wiping. The superior performance is attributed to MLA's ability to better align with and interpret robotic multisensory inputs, thereby enhancing its perceptual representation of the physical environment compared to the baselines. Furthermore, relative to SpatialVLA, MLA's advantage arises from its capability to generate future multisensory states, which enables improved modeling of physical dynamics and provides more robust conditions for action generation. As shown in Figure 5 a), we visualize the robot execution progress for several tasks.

B. Ablation Study

To validate each of our proposed contributions, we conducted an ablation study on two real-world tasks, including pressing a stamp onto paper and placing an egg on bread.

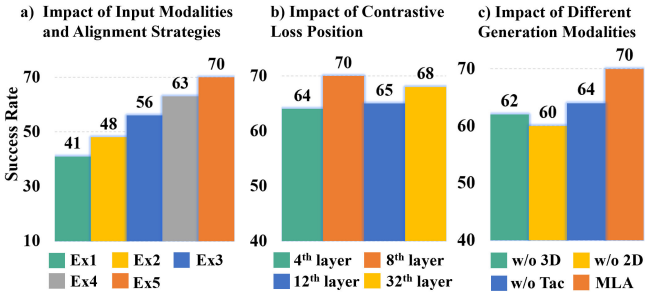

Fig. 6: **Ablation study.** We systematically analyze the contributions of each component in the MLA model.

Impact of Input Modalities and Alignment Strategies in the Encoder-Free Multimodal Alignment Scheme. As shown in Figure 6 a), we first examine the role of different input modalities and alignment strategies under the following configurations: (Ex1) 2D image input only, (Ex2) 2D image + 3D point cloud with simple token-level concatenation, (Ex3) 2D image + 3D point cloud + tactile signals with simple token-level concatenation, (Ex4) all modalities with image-level contrastive alignment, and (Ex5) our proposed all modalities with token-level contrastive alignment. Compared with Ex1 and Ex2, Ex3 demonstrates that semantic, spatial, and interactive perception are all critical for contact-rich manipulation. Compared with Ex1–Ex3, Ex5 achieves significant improvements, showing that the proposed position-guided consistency constraint strengthens multimodal representations. Furthermore, compared to Ex4, where multimodal inputs from the same timestep are treated as positives and those from different timesteps as negatives, Ex5 still achieves a 7% accuracy gain, highlighting the advantage of token-level cross-modal alignment in physical-world perception.

Impact of Contrastive Loss Position. As shown in Figure 6 b), we investigate the effect of applying contrastive loss at different layers of the LLaMA-2 backbone. Specifically, we select the 4th, 8th, 12th, and 32nd layers for cross-modal alignment during the SFT and post-training stages. The results reveal that applying token-level contrastive learning at the 8th layer yields the best performance, as it aligns features at relatively shallow layers while leaving sufficient subsequent transformer blocks to focus on future state prediction and action generation. Interestingly, applying self-supervision at the 32nd layer yields limited gains, as the final hidden states are already optimized for multiple objectives.

Impact of Multimodal Data Encoding Methods. We also compare injecting additional 2D [35] and 3D [33] with our proposed approach that repurposes the LLM itself as a unified perception module. The results show that introducing extra encoders not only decreases performance (–7%) but also reduces inference efficiency.

Impact of Different Generation Modalities in Future State Generation. As shown in Figure 6 c), building upon the MLA model following SFT, we further evaluated three ablation variants during post-training: (1) without image generation, (2) without point cloud generation, and (3) without tactile signal generation. The results indicate that

TABLE I: **Results on the RLBench benchmark.** Each model is evaluated over 20 rollouts, with success determined by the built-in RLBench evaluation module. Results report average manipulation success rates (S.R., %) with variance.

Models	Close box	Close laptop lid	Toilet seat down	Sweep to dustpan	Close fridge	Phone on base	Take umbrella out	Take frame off hanger	Place wine at rack	Water plants	Mean S.R. & Var
OpenVLA [5]	0.60	0.35	0.75	0.55	0.85	0.20	0.30	0.15	0.20	0.05	0.40±0.02
π_0 [3]	0.85	**0.95**	0.90	0.85	**1.00**	0.05	0.10	**0.90**	0.65	0.25	0.65±0.04
HybridVLA [4]	0.85	0.75	**1.00**	0.80	0.95	0.50	0.50	0.30	0.70	0.25	0.66±0.05
SpatialVLA [8]	0.80	0.70	0.85	0.20	0.80	0.15	0.25	0.40	0.15	0.30	0.46±0.03
UP-VLA [40]	0.80	0.40	0.65	0.10	0.80	0.15	0.35	0.55	0.20	0.20	0.42±0.04
*DreamVLA** [11]	**0.95**	0.75	0.95	0.25	**1.00**	0.35	**0.55**	0.50	**0.85**	0.35	0.65±0.05
MLA	**0.95**	0.90	**1.00**	**1.00**	0.95	**0.60**	0.50	**0.90**	0.75	**0.55**	**0.81±0.03**

removing future state generation from any modality leads to a drop in accuracy, reaffirming that generating comprehensive semantic, spatial, and interactive information provide more robust feature conditions for action generation. Finally, we investigate the impact of predicting **future adjacent frames (64%)** versus **future keyframes (70%)** on manipulation performance. The results show that predicting adjacent frames introduces high redundancy, leading to limited improvements in motion planning and dynamic representation of MLA.

TABLE II: **Generalization experiments.** Visualization of the two generalization scenarios along with the corresponding quantitative results. The red boxes highlight the differences from the training setup.

Unseen object | Unseen object | Unseen background | Unseen background

Model	Original	Unseen Object	Unseen Background
π_0	47	35 (-26%)	25 (-47%)
MLA	53	45 (-15%)	40 (-25%)

C. Generalization Experiment

As shown in Table II, we designed two common generalization scenarios to compare our MLA with π_0, including unseen manipulated objects and unseen complex backgrounds. The most challenging task, placing an egg on bread with a spatula, is selected as the evaluation task. For unseen objects, we replace the egg with lettuce and change the color of the target plate. Across these foreground modifications, MLA shows almost no decrease in success rate for the initial subtask. For unseen backgrounds, cluttered scenes are introduced during testing by adding unseen objects around the manipulated object. Even under such challenging background conditions, MLA maintains a 40% success rate in completing the entire task. These results demonstrate that MLA can better perceive and reason about robotic manipulation scenes, whether facing semantic variations in manipulated objects or background interference. This robustness is attributed to its strong multimodal perception capability and its ability to anticipate the future states of manipulated objects.

D. Simulation Experiment

Simulation Benchmark. To systematically evaluate the performance of MLA, we conducted experiments on 10 tasks in the RLBench [17] benchmark, which is based on the CoppeliaSim simulator. For each task, 100 demonstration trajectories were collected using the official Motion Planning Library [41]. Observations consist of a front-view camera image and the corresponding point cloud data. We extract the keyframes following the approach in [33].

Training and Evaluation Details. As tactile sensing in simulation is not realistic, we provide only image and point cloud modalities to MLA and all baseline methods. We selected several state-of-the-art baselines from relevant domains, including OpenVLA [5], π_0 [3], HybridVLA [4], SpatialVLA [8], UP-VLA [40], and $DreamVLA^*$ [11]. For each baseline, we loaded the officially released pretrained checkpoints. Since $DreamVLA^*$ does not provide a general large-scale pretrained checkpoint, we re-implemented its input and generation strategy on our backbone for a fair comparison. All tasks were trained jointly, and evaluation was performed using 20 rollouts per task.

Results. As shown in Table I, MLA achieves an average score of 81% across 10 tasks, significantly outperforming π_0 (65%), SpatialVLA (46%), and other baselines. The improvements are particularly notable on more challenging tasks, such as *Place Wine at Rack Location* and *Water plants*. These results validate the effectiveness of our proposed multimodal alignment and future multisensory generation post-training, which enable MLA to progressively enhance its representations and achieve higher action accuracy. They also demonstrate that our paradigm remains effective even without access to expensive sensors such as tactile devices. Furthermore, as shown in Figure 5 b), we visualize the attention heatmaps from the output features of MLA and a variant without our proposed approach. The results clearly show that MLA learns better feature representations and focuses more effectively on both the robot and the manipulated objects.

V. CONCLUSIONS

In this work, we introduced MLA, a multisensory language–action model that integrates 2D visual, 3D geometric, and tactile cues through encoder-free multimodal alignment and enhances physical-world understanding via future multisensory generation. We progressively equip a LLM with the ability to integrate perception, understanding, and action generation from multisensory inputs in the real world through large-scale pretraining, supervised fine-tuning, and post-training. MLA not only achieves state-of-the-art performance and demonstrates strong generalization across both real-world and simulation tasks, but also provides a new multimodal foundation model paradigm for the community.

VI. Acknowledgement

This work was supported by the National Natural Science Foundation of China (625B2007). This work was also supported by the National Natural Science Foundation of China (62476011). This work was also supported by Beijing Innovation Center of Humanoid Robotics.

References

[1] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh, "Prismatic vlms: Investigating the design space of visually-conditioned language models," in *Forty-first International Conference on Machine Learning*, 2024.

[2] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[3] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn *et al.*, "pi0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[4] J. Liu, H. Chen, P. An, Z. Liu, R. Zhang, C. Gu, X. Li, Z. Guo, S. Chen, M. Liu *et al.*, "Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model," *arXiv preprint arXiv:2503.10631*, 2025.

[5] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[6] Q. Li, Y. Liang, Z. Wang, L. Luo, X. Chen, M. Liao, F. Wei, Y. Deng, S. Xu, Y. Zhang *et al.*, "Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation," *arXiv preprint arXiv:2411.19650*, 2024.

[7] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, "3d-vla: a 3d vision-language-action generative world model," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 61 229–61 245.

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

[9] Z. Cheng, Y. Zhang, W. Zhang, H. Li *et al.*, "Omnivtla: Vision-tactile-language-action model with semantic-aligned tactile sensing," *arXiv preprint arXiv:2508.08706*, 2025.

[10] J. Huang, S. Wang, F. Lin, Y. Hu, C. Wen, and Y. Gao, "Tactile-vla: Unlocking vision-language-action model's physical knowledge for tactile generalization," *arXiv preprint arXiv:2507.09160*, 2025.

[11] W. Zhang, H. Liu, Z. Qi, Y. Wang, X. Yu, J. Zhang, R. Dong, J. He, H. Wang, Z. Zhang *et al.*, "Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge," *arXiv preprint arXiv:2507.04447*, 2025.

[12] Y. Wang, X. Li, W. Wang, J. Zhang, Y. Li, Y. Chen, X. Wang, and Z. Zhang, "Unified vision-language-action model," *arXiv preprint arXiv:2506.19850*, 2025.

[13] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn *et al.*, "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1702–1713.

[14] Open X-Embodiment Collaboration, A. Padalkar, A. Pooley *et al.*, "Open X-Embodiment: Robotic learning datasets and RT-X models," https://arxiv.org/abs/2310.08864, 2023.

[15] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," 2024.

[16] K. Wu, C. Hou, J. Liu, Z. Che, X. Ju *et al.*, "Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation," in *Robotics: Science and Systems (RSS) 2025*. Robotics: Science and Systems Foundation, 2025.

[17] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.

[18] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: an embodied multimodal language model," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 8469–8488.

[19] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[20] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy," *arXiv preprint arXiv:2403.03954*, 2024.

[21] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn *et al.*, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," 2025. [Online]. Available: https://arxiv.org/abs/2504.16054

[22] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu *et al.*, "Rdt-1b: a diffusion foundation model for bimanual manipulation," in *The Thirteenth International Conference on Learning Representations*.

[23] H. Chen, J. Liu, C. Gu, Z. Liu, R. Zhang, X. Li, X. He, Y. Guo, C.-W. Fu, S. Zhang *et al.*, "Fast-in-slow: A dual-system foundation model unifying fast manipulation within slow reasoning," *arXiv preprint arXiv:2506.01953*, 2025.

[24] C. Li, J. Wen, Y. Peng, Y. Peng, F. Feng, and Y. Zhu, "Pointvla: Injecting the 3d world into vision-language-action models," 2025. [Online]. Available: https://arxiv.org/abs/2503.07511

[25] Y. Hu, Y. Guo, P. Wang, X. Chen, Y.-J. Wang, J. Zhang, K. Sreenath, C. Lu, and J. Chen, "Video prediction policy: A generalist robot policy with predictive visual representations," *arXiv preprint arXiv:2412.14803*, 2024.

[26] S. Zhou, Y. Du, J. Chen, Y. Li, D.-Y. Yeung, and C. Gan, "Robodreamer: Learning compositional world models for robot imagination," *arXiv preprint arXiv:2404.12377*, 2024.

[27] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Hu, X. Huang *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," *arXiv preprint arXiv:2503.06669*, 2025.

[28] Z. Liu, J. Liu, H. Chen, Z. Guo, C. Hou, C. Gu, J. Yu, X. Mi, R. Zhang, Z. Che, J. Tang, P.-A. Heng, and S. Zhang, "Last$_0$: Latent spatio-temporal chain-of-thought for robotic vision-language-action model," 2026. [Online]. Available: https://arxiv.org/abs/2601.05248

[29] H. Wu, Y. Jing, C. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong, "Unleashing large-scale video generative pre-training for visual robot manipulation," *arXiv preprint arXiv:2312.13139*, 2023.

[30] J. Zhang, Y. Guo, Y. Hu, X. Chen, X. Zhu, and J. Chen, "Up-vla: A unified understanding and prediction model for embodied agent," 2025. [Online]. Available: https://arxiv.org/abs/2501.18867

[31] J. Chen, Z. Cai, P. Chen, S. Chen, K. Ji, X. Wang, Y. Yang, and B. Wang, "Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation," *arXiv preprint arXiv:2506.18095*, 2025.

[32] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu *et al.*, "Emu3: Next-token prediction is all you need," *arXiv preprint arXiv:2409.18869*, 2024.

[33] Y. Jia, J. Liu, S. Chen, C. Gu, Z. Wang, L. Luo, X. Li, P. Wang, Z. Wang, R. Zhang *et al.*, "Lift3d policy: Lifting 2d foundation models for robust 3d robotic manipulation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17 347–17 358.

[34] C. R. Qi, H. Su, K. Mo, and L. 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*, 2017, pp. 652–660.

[35] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *International Conference on Computer Vision (ICCV)*, 2023.

[36] Y. Tang, R. Zhang, J. Liu, Z. Guo, B. Zhao, Z. Wang, P. Gao, H. Li, D. Wang, and X. Li, "Any2point: Empowering any-modality large models for efficient 3d understanding," in *European Conference on Computer Vision*. Springer, 2025, pp. 456–473.

[37] Y. Pang, W. Wang, F. E. H. Tay, W. Liu, Y. Tian, and L. Yuan, "Masked autoencoders for point cloud self-supervised learning," 2022. [Online]. Available: https://arxiv.org/abs/2203.06604

[38] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[39] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," 2024. [Online]. Available: https://arxiv.org/abs/2309.13037

[40] J. Zhang, Y. Guo, Y. Hu, X. Chen, X. Zhu, and J. Chen, "Up-vla: A unified understanding and prediction model for embodied agent," *arXiv preprint arXiv:2501.18867*, 2025.

[41] I. A. Şucan, M. Moll, and L. E. Kavraki, "The Open Motion Planning Library," *IEEE Robotics & Automation Magazine*, vol. 19, no. 4, pp. 72–82, December 2012, https://ompl.kavrakilab.org.