A Universal Embodied Artificial Intelligent Scheme Combined Large Language Model with Diffusion Policy for Robot Manipulation Tasks

Abstract—The development of large language model (LLM) and visual language model(VLM) has driven rapid progress in embodied artificial intelligence (Embodied AI) technology, opening up new avenues for robots to interact with the physical world and providing vast space for robots to perform more complex manipulation tasks. Moreover, the diffusion policy can elegantly handle multimodal action distributions, it's suitable for high-dimensional action spaces, and exhibits impressive training stability, making it highly suitable for robot control. Traditional robot trajectory planning and control methods are sensitive to disturbances and have strict hardware requirements, making them difficult to apply in precise robot grasping tasks. In addition, it is difficult to ensure the generalization of these methods for different task scenarios. To address these issues, this article proposes an Embodied AI scheme based on the LLM and the diffusion policy. This solution firstly converts human natural language instructions into robot manipulation subtasks by LLM and then generates safe robot trajectories and actions based on the diffusion policy. The combination of LLM and diffusion policy addresses the existing problems in robot manipulation tasks, and most importantly, this method has strong generality and generalization. To validate our idea, we conducted physics experiments on a 7 dof Franka Emika Panda. The experimental results show that compared to previous embodied intelligence schemes, the scheme designed in this paper greatly improves the success rate and safety of task execution. In addition, this scheme has strong generalization ability for different task scenarios.

Index Terms—Large language model, visual language model, embodied artificial intelligence, diffusion policy, Franka Emika Panda.

I. Introduction

IN recent years, large language model(LLM) and visual language model(VLM) have been greatly applied in the field of robot manipulation. Embodied Artificial Intelligence(Embodied AI) [1], [2] refers to an intelligent agent that has a body and supports physical interaction. Under Embodied AI, artificial intelligence has a body and robots have a brain.

In previous robot operations, we needed to predefine trajectories, which made robots more limited. More importantly, obtaining large-scale robot data was difficult, which limited the development of the robotics field. The excellent response of ChatGPT 4 makes us feel that it is possible for robots to become universal robots. We can use LLM to reason, provide useful steps for robots, and then use VLM to plan paths. In theory, robots can interact with the real world through natural language.

The current technical roadmap for multimodal large models can be divided into two coarse-grained approaches. One major approach is to use multimodal VLM and traditional control algorithms to achieve multimodal perception and decision-making for robots [3], [4], [5], [6], [7], [8], enabling robots to handle complex and unfamiliar scenes. Then, traditional control algorithms can be called through API interfaces to achieve motion control for robots. Another major approach is to attempt to build an integrated visual language action(VLA) model [9], [10], [11], [12] that directly maps perception signals to robot operation and control instructions, skipping the complex signal conversion process and achieving a closed-loop of perception decision action. Although the latter approach is more direct in controlling robots, it is currently not mature. The former utilizes both the generation capability of large models and the advantages of stable and efficient traditional control, making it the mainstream technology solution for embodied intelligence.

Diffusion Policy [13], [14] is a novel robot trajectories [15], [16], [15], [17] or/and actions [18] generation method that represents the visual trajectories or/and actions policy of a robot as a conditional denoising diffusion process. The advantage of using a diffusion model to plan robot trajectories or/and actions is that the diffusion model can sample from complex arbitrary distributions. The distribution of robot trajectories or/and actions sequences is usually complex and multimodal, meaning that robots can typically complete a task in multiple ways. However, diffusion strategies can easily handle these complexities. Using a universal diffusion policy to train robot strategies can enable robots to learn multimodal behavior, making robot training more stable.

Although there are currently some Embodied AI achievements [19], [20], security issues are rarely mentioned. There are many aspects to the safety of robots, such as physical limitations of actuators, constraints on system states, and speed constraints during operation. Safety is a prerequisite for robots to operate stably, so the actual deployment of embodied intelligence solutions must consider safety issues.

This article integrates LLM into robots and combines the advantages of stable, efficient, and universal diffusion policy to convert complex human natural language instructions into specific robot trajectories and action outputs without the need for additional data and training, solving the problem of scarce robot training data. In this article, the operable objects are also open and do not require a predetermined range. Opening bottles, pressing switches, and unplugging charging wires can all be completed. In theory, as long as the above basic processes are

mastered, any given task can be completed, achieving zero sample daily operation task trajectory synthesis. Robots in the real world can directly perform this task without training. Overall, this paper makes the following main contributions:

1) In the field of embodied intelligence, we have innovatively proposed combining the strong generalization ability of LLM with the stable and efficient diffusion strategy for robot manipulation tasks. The proposed solution does not require specific dynamic modeling and can achieve planning and control of long-term tasks.
2) Based on the iterative diffusion process, both safety trajectories and safety actions are generated simultaneously, simplifying the overall design scheme.
3) Compared to existing Embodied AI schemes, our proposed scheme considers the safety of the robotic arm during operation, significantly reducing the number of constraint violations and greatly improving the success rate and safety of task completion.
4) The integration of large language models into robots converts complex human instructions into specific robot trajectories and actions without the need for additional training data, solving the problem of scarce training data for robots. This scheme has strong generality and theoretically can complete given tasks with zero samples without training.

The remainder of this paper is organized as follows: Section II describes the problem statement and details some preliminaries. Section III gives unified description of large language model and visual language model. This is followed by the design of our diffusion policy in Section IV. Section V documents experimental studies. Section VI draws conclusions.

II. Problem Description and Preliminaries

A. System Description

Consider the system is governed by a unknown discrete-time system model:

$$s_{t+1} = f(s_t, a_t) \tag{1}$$

where $s_t \in \mathbb{R}^n$ is the state of system (1), $a_t \in \mathbb{R}^m$ is the action of system (1).

The constraint set of state S_t and action A_t is

$$\mathrm{s}_t \subseteq S, \qquad \mathrm{a}_t \subseteq A \tag{2}$$

B. Problem Statement

The goal of this article is to generate safe trajectories and actions for robots based on human natural language instructions. Therefore, the objectives of this article can be divided into two major aspects:

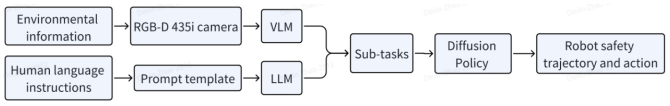

Fig. 1. The overall framework of embodied artificial intelligent scheme.

Problem 1: Firstly, because robots cannot understand human natural language instructions. Therefore, how to automatically convert human instructions into robot executable steps for different task scenarios?

Problem 2: Secondly, based on the instructions of robot operability, how to apply strategies to plan robot safety paths and robot safety actions?

C. Overall Framework

To solve these problems in section II-B, we propose a novel Embodied AI scheme, the overall framework is shown in Figure 1.

As shown in Figure 1, we first utilize the powerful reasoning ability of large language models and the powerful environmental recognition ability of visual language models to split human natural language instructions into several sub-tasks that robots can execute. Then, we train a universal diffusion model to generate stable and safe robot trajectories and actions. These technical details will be detailed in Section III and Section IV.

III. Large Language Model and Visual Language Model

Firstly, given the environmental information (capturing RGB-D images with a camera) and the natural language instructions we need to execute. LLM can extract the information of interest from human instructions, such as "open the drawer, be careful of the vase" [3]. At this point, LLM knows that the drawer is the target and the vase needs to be avoided. Then, VLM is used to obtain the specific location of the vase and drawer. LLM parses the instructions and automatically decomposes the task based on the prompt template, dividing the total task L assigned by humans into several sub tasks $\{L_1, L_2, \cdots, L_n\}$ that can be executed by robots.

For example, when the natural language instructions L is: put the sour fruit into the top drawer. Then, LLM will split L into seven sequential subtasks $L_1, \cdots, L_7$. (L_1:"grasp the top drawer handle", L_2:"move away from the top drawer handle by 25cm", L_3:"open gripper", L_4:"back to default pose", L_5:"grasp the lemon"), L_6:"move to 10cm on top of the top drawer", L_7:"open gripper"). For another natural language instructions: Could you please set up the fork for the steak for me? The LLM will decompose into the following five subtasks: "grasp the fork"→("back to default pose"→"move to 10cm to the right of the plate"→"open gripper"→"back to default pose". This process utilizes the powerful reasoning ability of LLM. The LLM used in this article is the GPT-4 API, and the template format is referenced in reference [21].

Visual language model is a technology that combines image and natural language processing. Its main purpose is to understand and interpret the relationship between images and text, and generate accurate and vivid natural language descriptions based on images. This model generates relevant textual descriptions by analyzing the

content and context of images, giving computers a visual understanding ability closer to that of humans.

In this paper, we Supervised Fine-Tuning the local detector OWL ViT [22] to to serve as VLM. We utilize the fined-tuning VLM to obtain the bounding box, and uses Segment Anything Model to obtain the mask, and then uses video tracker XMEM to track the mask. The tracked mask uses RGB-D to reconstruct the 3D point cloud.

IV. Design of Diffusion Policy

One of the main advantages of diffusion policy is that they use simple and efficient loss functions during training and can generate highly realistic data. The working mechanism of the diffusion model is divided into two stages. Firstly, they introduce noise into the dataset, which is a fundamental step in the forward diffusion process, and then systematically reverse this process. Below is a detailed breakdown of the diffusion model lifecycle. The diffusion model is trained by a temporal U-Net [23], [24].

The Denoising Diffusion Probabilistic Model (DDPM) is a generative model designed to fit a given target distribution so that samples can be taken from it. It uses Markov chain modeling, with noise as input, gradually denoised through neural networks, and finally produces an output that conforms to the target distribution. The generated data can be arbitrary, including images, speech, robot trajectories, and actions.The execution of robot manipulation is currently a shortcoming, which limits the generality of robot operation. Diffusion Policy has shown excellent performance and great potential in solving the problem of agile robot operation execution.

A. Denoising diffusion probabilistic models [25]

The predefined forward diffusion process involves adding noise to a variable step by step until the variable becomes pure noise. The form and parameters of this process are manually defined. The forward process is a predefined Markov chain, which is defied as:

$$q(x^k|x^{k-1}) = N(\sqrt{1-\beta_k}x^k, \beta_k I) \tag{3}$$

where $q_\theta(x^{k-1}|x^k)$ represents the distribution of real data. $x^0 = x$, $k = 1, \cdots, K$ is the diffusion time step, and $\beta_k \in (0, 1)$.

From Eq.(3), we can get

$$q(x^k|x^0) = N(\sqrt{\alpha_k}x^0, (1-\alpha_k)I) \tag{4}$$

where $\alpha_k = \prod_{i=1}^{k}(1-\beta_i)$, and $q(x^K|x^0) \approx N(0, I)$.

The reverse denoising diffusion process that needs to be learned is to gradually denoise from pure noise until the variables are obtained. This process is represented by a learnable neural network. The reverse denoising diffusion process is defied as:

$$p_\theta(x^{k-1}|x^k) = N(\mu_\theta x^k, \sum x_k) \tag{5}$$

And, μ_θ in Eq.(5) is expressed as:

$$\mu_\theta = \frac{1}{\sqrt{1-\beta_k}}(x^k - \frac{\beta_k}{\sqrt{1-\alpha_k}}\varepsilon_\theta x^k)\sum x_k) \tag{6}$$

where $\varepsilon_\theta \sim N(0, I)$.

B. Data Collection and Sampling Trajectories

The collection of dataset D was obtained by randomly sampling a pre-defined action space A, and we collected a total of 10^6 trajectories. The termination condition for the data collection process is when the state leaves the state space S or the current position reaches the target point g.

Define the trajectory of the system (1) is

$$\tau(z) = ((s_0, a_0), \cdots, (s_n, a_n)) \tag{7}$$

Trajectories (7) follow the following distribution

$$q(\tau) = q(A)\prod_{t=1}^{m}\delta(s_t - \hat{s}_t) \tag{8}$$

where $A = (a_1, \cdots, a_n)$ is the action, $q(A)$ is the distribution of action. $q = (q_1, \cdots, a_n)$ is the trajectory, $q(A)$ is the distribution of trajectory, and $\hat{s}_t = f(s_{t-1}, a_{t-1})$.

Based on the collected dataset D, we obtain the trajectory distribution by distributing the action space according to Eq. (8), and then sample the trajectory using Eq. (9).

$$p_\theta(\tau^{k-1}|\tau^k) = N(\mu_\theta\tau^k, \sum \tau_k) \tag{9}$$

where $k = K, \cdots, 1$ and $\tau^K \sim N(0, I)$.

The sampled trajectories are neither optimal to goal g nor satisfy the state constraint S. Then, we will consider the optimality and safety issues of the trajectories.

C. Optimality of the Trajectories [15]

Define the cumulative reward function as:

$$R_g(\tau) = \sum r(s_t, a_t, g) = -d^2(s_t, g) \tag{10}$$

where $r(s_t, a_t, g)$ is the reward function. $d(s_t, g)$ represents the distance between the current positon and the goal g.

Since the calculation of $R_g(\tau)$ is difficult, we train a value diffusion model V_ψ to predict the $R_g(\tau)$ by minimizing the loss.

$$L_\psi = E_{\tau\sim D}[|V_\psi(\tau, g) - R_g(\tau)|_2] \tag{11}$$

Then, we can approximate the Eq.(9) as:

$$p_\theta(\tau^{k-1}|\tau^{k-1}) \approx N(\mu_\theta(\tau^k) + m\Sigma_k\upsilon, \Sigma_k) \tag{12}$$

where $\upsilon = \nabla_\tau V_\psi(\tau, g)|_{\tau=\mu\theta,k}$, and $m > 0$.

D. Safely of the Trajectories [26]

We define the unsafe regions of state space as $\bar{S}_1, \cdots, \bar{S}_M$ and the unsafe regions of action space as $\bar{A}_1, \cdots, \bar{A}_N$, then the constraint of state and action (2) becomes:

$$S = S \backslash \bigcup_{M}^{i=2} \bar{S}_i, \qquad A = A \backslash \bigcup_{j=0}^{N} \bar{A}_j \tag{13}$$

In this way, unsafe states and actions can be easily removed.

The overall implementation of the diffusion policy Part consists of multiple steps, as shown in Algorithm 1.

Algorithm 1 The implementation process of the entire task execution.

Input: Environmental information (capturing RGB-D images with a realsence depth camera D435i) and the Human natural language instructions, goal g, robot action space A

Output: Optimal safe trajectory τ^* and safe action a_0^*

According to the given prompt template, use ChatGPT 4 to split the total task L and obtain several sub tasks $L_1, \cdots, L_n$ that robots can execute
Initialization $t = 0$
repeat
  Collect data set D by applying randomly uniformly selected actions from action space A
until Get 10^6 trajectories
Training a Trajectory diffusion model ε_τ, value model V_ψ utilize the collected data set D
repeat
  for each $k = \overline{K}, \cdots, 1$ do
    Compute gradient $\upsilon = \nabla_\tau V_\psi(\tau, g)|_{\tau=\tau_t^{k,1:B}}$
    Implement the denoising process: Sample $\tau_t^{k-1,1:B}$ $N(\mu_\theta(\tau_t^{k,1:B}, k) + m\Sigma_{k-1}\upsilon, \Sigma_{k-1})$
    Set the first state in $\tau_t^{k-1,1:B}$ $= (s_0, a_0, \cdots, s_T, a_T)$ as the current state
    for each $i = 0, \cdots, M$ do
      if $\hat{s}_t \in \bar{S}_i$ then $\hat{s}_t \leftarrow proj_{S \setminus \bar{S}_i(\hat{s}_t)}$
    end for
    for each $j = 0, \cdots, N$ do
      if $\hat{a}_t \in \bar{A}_j$ then $\hat{a}_t \leftarrow proj_{A \setminus \bar{A}_j(\hat{a}_t)}$
    end for
  end for
  Get the optomal trajectory $\tau^* = arg\max_{\tau_t^{0,i}} V_\psi(\tau_t^{0,i})$
  Apply the first action a_0^* in τ^*
until Reaches the goal g
return a_0^*

In Algorithm 1, based on the environmental information collected by the RGB-D depth camera, LLM is first used to decompose human natural language instructions into several sub-tasks that can be executed by robots according to the designed prompt template. Then, a dataset of actions is obtained by randomly sampling the pre-defined robot action space. Based on this dataset, the distribution of actions is obtained, and the trajectory distribution is further obtained through equation 1. Then, based on the dataset, a trajectory diffusion model is trained to approximate the trajectory distribution. Then, based on the diffusion model, a reverse denoising process is carried out, which considers both optimality and safety issues, ultimately obtaining the optimal and safe robot trajectory and action.

V. Numerical Experiments

In this section, we validated the correctness of the proposed algorithm on a 7-Dof Franka Emika Panda Robot.

A. Dynamics of Robot

The dybamic model of 7-Dof Franka Emika Panda Robot is expressed as:

The state space and action space and their bounds are defined in table 1. And, the hyperparameters of the training process are shown in table 2.

TABLE I
The state and action and their bounds of system (??).

Symbol	Definition	Bounds
R_s	stator resistance	$\pm 5N$
L_d, L_q	inductance along	$\pm 5N$
n_p	the number of poles	$\pm 5N$
B_m	viscous friction torque	$\pm 5N$

TABLE II
The hyperparameters in training process.

Hyperparameter	Value
Optimizer	Adam
Batch size	32
Learning rate	2×10^{-5}
Traning steps	10^6
Epochs	100
Diffusion steps k	20

B. Experiment Results

The experimental setup is shown in Figure 2 below:

As shown in Figure 2 above, a 7-degree-of-freedom Franka Emika Panda is used to complete the grasping task. We install a Realsense RGB-D camera at the left diagonal position of the robotic arm to observe environmental information. LLM calls the GPT-4 API, and we trained a VLM based on OWL-VIT to recognize objects in the scene.

VI. Conclusion

This paper has presented a embodied artificial intelligence scheme for robot manipulation tasks, which combines the generation ability of large language model and the environmental recognition ability of visual language model, and utilizes the stable and efficient advantages of

Fig. 2. An experimental scenario for a Franka Emika Panda.

TABLE III
Task decomposition of human natural language instructions by the ChatGPT 4.

Task 1→	Put the tennis into the black bin
Subtask 1→	grasp the tennis
Subtask 2→	back to default pose
Subtask 3→	move to 10cm above the black bin
Subtask 4→	open gripper
Task 2→	Take out a tissue paper and put it on the cabinet
Subtask 1→	grasp the tissue paper
Subtask 2→	back to default pose
Subtask 3→	move to 10cm on top of the cabinet
Subtask 4→	open gripper
Subtask 5→	back to default pose
Task 3→	Unplug the charger plug from the socket and put it aside
Subtask 1→	grasp the charger plug
Subtask 2→	back to default pose
Subtask 3→	move to 30cm to the left
Subtask 4→	open gripper
Subtask 5→	back to default pose
Task 4→	Open the top drawer by 20cm, watch out for that vase
Subtask 1→	grasp the top drawer handle while keeping at least 15cm away from the vase
Subtask 2→	move 20cm away from the drawer handle while keeping at least 15cm away from the vase
Subtask 3→	open gripper
Subtask 4→	back to default pose

TABLE IV
Index in different task scenarios.

Task scenario in voxposer→	task 1	task 2	task 3	task 4
Index 1: Safety→	safe	safe	safe	safe
Index 2: Success rate→	80%	80%	80%	80%
Index 3: Time consuming→	45s	45s	45s	45s
Task scenario in DP→	task 1	task 2	task 3	task 4
Index 1: Safety→	unsafe	unsafe	unsafe	unsafe
Index 2: Success rate→	80%	80%	80%	80%
Index 3: Time consuming→	45s	45s	45s	45s

diffusion policy to generate safe robot trajectories and actions. This scheme improves the flexibility and safety of robot trajectory planning and control, and enhances the success rate and generalization of task execution. The experimental results have demonstrated the effectiveness and correctness of this method. In the future, we will apply this embodied intelligence solution with universality and generalization to humanoid robot manipulation tasks.

References

[1] R. Chrisley, "Embodied artificial intelligence," Artificial intelligence, vol. 149, no. 1, pp. 131–150, 2003.

[2] R. Pfeifer and F. Iida, "Embodied artificial intelligence: Trends and challenges," Lecture notes in computer science, pp. 1–26, 2004.

[3] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," arXiv preprint arXiv:2307.05973, 2023.

[4] 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," arXiv preprint arXiv:2303.03378, 2023.

[5] C. Zhao, S. Yuan, C. Jiang, J. Cai, H. Yu, M. Y. Wang, and Q. Chen, "Erra: An embodied representation and reasoning architecture for long-horizon language-conditioned manipulation tasks," IEEE Robotics and Automation Letters, vol. 8, no. 6, pp. 3230–3237, 2023.

[6] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "Chatgpt for robotics: Design principles and model abilities. 2023," Published by Microsoft, 2023.

[7] K. Lin, C. Agia, T. Migimatsu, M. Pavone, and J. Bohg, "Text2motion: From natural language instructions to feasible plans," Autonomous Robots, vol. 47, no. 8, pp. 1345–1365, 2023.

[8] K. Bousmalis, G. Vezzani, D. Rao, C. Devin, A. X. Lee, M. Bauza, T. Davchev, Y. Zhou, A. Gupta, A. Raju et al., "Robocat: A self-improving foundation agent for robotic manipulation," arXiv preprint arXiv:2306.11706, 2023.

[9] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn et al., "Rt-2: Vision-language-action models transfer web knowledge to robotic control," arXiv preprint arXiv:2307.15818, 2023.

[10] S. Belkhale, T. Ding, T. Xiao, P. Sermanet, Q. Vuong, J. Tompson, Y. Chebotar, D. Dwibedi, and D. Sadigh, "Rt-h: Action hierarchies using language," arXiv preprint arXiv:2403.01823, 2024.

[11] 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.

[12] 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," arXiv preprint arXiv:2403.09631, 2024.

[13] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," arXiv preprint arXiv:2402.10885, 2024.

[14] B. Kang, X. Ma, C. Du, T. Pang, and S. Yan, "Efficient diffusion policies for offline reinforcement learning," Advances in Neural Information Processing Systems, vol. 36, 2024.

[15] M. Janner, Y. Du, J. B. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," arXiv preprint arXiv:2205.09991, 2022.

[16] X. Ma, S. Patidar, I. Haughton, and S. James, "Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 18 081–18 090.

[17] Z. Wang, J. J. Hunt, and M. Zhou, "Diffusion policies as an expressive policy class for offline reinforcement learning," arXiv preprint arXiv:2208.06193, 2022.

[18] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," arXiv preprint arXiv:2303.04137, 2023.

[19] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, W. Ai, B. Martinez et al., "Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation," arXiv preprint arXiv:2403.09227, 2024.

[20] X. Kong, W. Zhang, J. Hong, and T. Braunl, "Embodied ai in mobile robots: Coverage path planning with large language models," arXiv preprint arXiv:2407.02220, 2024.

[21] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023, pp. 9493–9500.

[22] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen et al., "Simple open-vocabulary object detection," in European Conference on Computer Vision. Springer, 2022, pp. 728–755.

[23] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2020, pp. 1055–1059.

[24] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," IEEE transactions on medical imaging, vol. 37, no. 12, pp. 2663–2674, 2018.

[25] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in neural information processing systems, vol. 33, pp. 6840–6851, 2020.

[26] Y. Zheng, J. Li, D. Yu, Y. Yang, S. E. Li, X. Zhan, and J. Liu, "Safe offline reinforcement learning with feasibility-guided diffusion model," arXiv preprint arXiv:2401.10700, 2024.