3D-VLA: A 3D Vision-Language-Action Generative World Model

Haoyu Zhen [1 2] **Xiaowen Qiu** [1] **Peihao Chen** [3] **Jincheng Yang** [2] **Xin Yan** [4]
Yilun Du [5] **Yining Hong** [6] **Chuang Gan** [1 7]

https://vis-www.cs.umass.edu/3dvla

Abstract

Recent vision-language-action (VLA) models rely on 2D inputs, lacking integration with the broader realm of the 3D physical world. Furthermore, they perform action prediction by learning a direct mapping from perception to action, neglecting the vast dynamics of the world and the relations between actions and dynamics. In contrast, human beings are endowed with world models that depict imagination about future scenarios to plan actions accordingly. To this end, we propose 3D-VLA by introducing a new family of embodied foundation models that seamlessly link 3D perception, reasoning, and action through a generative world model. Specifically, 3D-VLA is built on top of a 3D-based large language model (LLM), and a set of interaction tokens is introduced to engage with the embodied environment. Furthermore, to inject generation abilities into the model, we train a series of embodied diffusion models and align them into the LLM for predicting the goal images and point clouds. To train our 3D-VLA, we curate a large-scale 3D embodied instruction dataset by extracting vast 3D-related information from existing robotics datasets. Our experiments on held-in datasets demonstrate that 3D-VLA significantly improves the reasoning, multimodal generation, and planning capabilities in embodied environments, showcasing its potential in real-world applications.

[1]University of Massachusetts Amherst [2]Shanghai Jiao Tong University [3]South China University of Technology [4]Wuhan University [5]Massachusetts Institute of Technology [6]University of California, Los Angeles [7]MIT-IBM Watson AI Lab. Correspondence to: Chuang Gan <ganchuang1990@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Nowadays, there has been a proliferation of vision-language models (Liu et al., 2023; Alayrac et al., 2022; Li et al., 2023b) that can take images as inputs and perform a series of reasoning tasks in the 2D space, mirroring the versatility of the human brain. Such 2D foundation models also lay the foundation for recent embodied foundation models such as RT-2 (Brohan et al., 2023) and PALM-E (Driess et al., 2023a) that could generate high-level plans or low-level actions contingent on the images. However, they neglect the fact that human beings are situated within a far richer 3D physical world beyond 2D images - they reason, plan, and act based on their 3D understanding of the environment (Palmer, 1975; Pylyshyn, 2003; Marr, 2010). It's crucial that human-like intelligent embodied agents are equipped with the same 3D understanding ability.

Taking a step forward, recent works (Huang et al., 2023b; Hong et al., 2024) develop embodied foundation models that could plan and act in the 3D environment. However, such models mainly learn a direct mapping from perception to action, devoid of a broader understanding of the dynamics of the world, and the relations between actions and world dynamics. On the other hand, human beings are blessed with world models that simulate future events based on 3D internal representations. By depicting the imagination and anticipation about the future states, one could better plan actions toward the predicted goals.

Challenges inevitably exist for building such human-like 3D world models. Firstly, existing foundation models focus on language generation, unable to imagine modalities beyond language and simulate future states to facilitate action generation, which is a crucial aspect of world models. Secondly, existing embodied datasets mainly contain 2D images or videos, lacking 3D-related annotations for reasoning and planning in the 3D space.

To this end, we propose 3D-VLA by introducing a new family of embodied foundation models that seamlessly link 3D perception, reasoning, and action through a generative world model. Specifically, we build our 3D-VLA on top of a 3D large language model (Hong et al., 2023) to equip

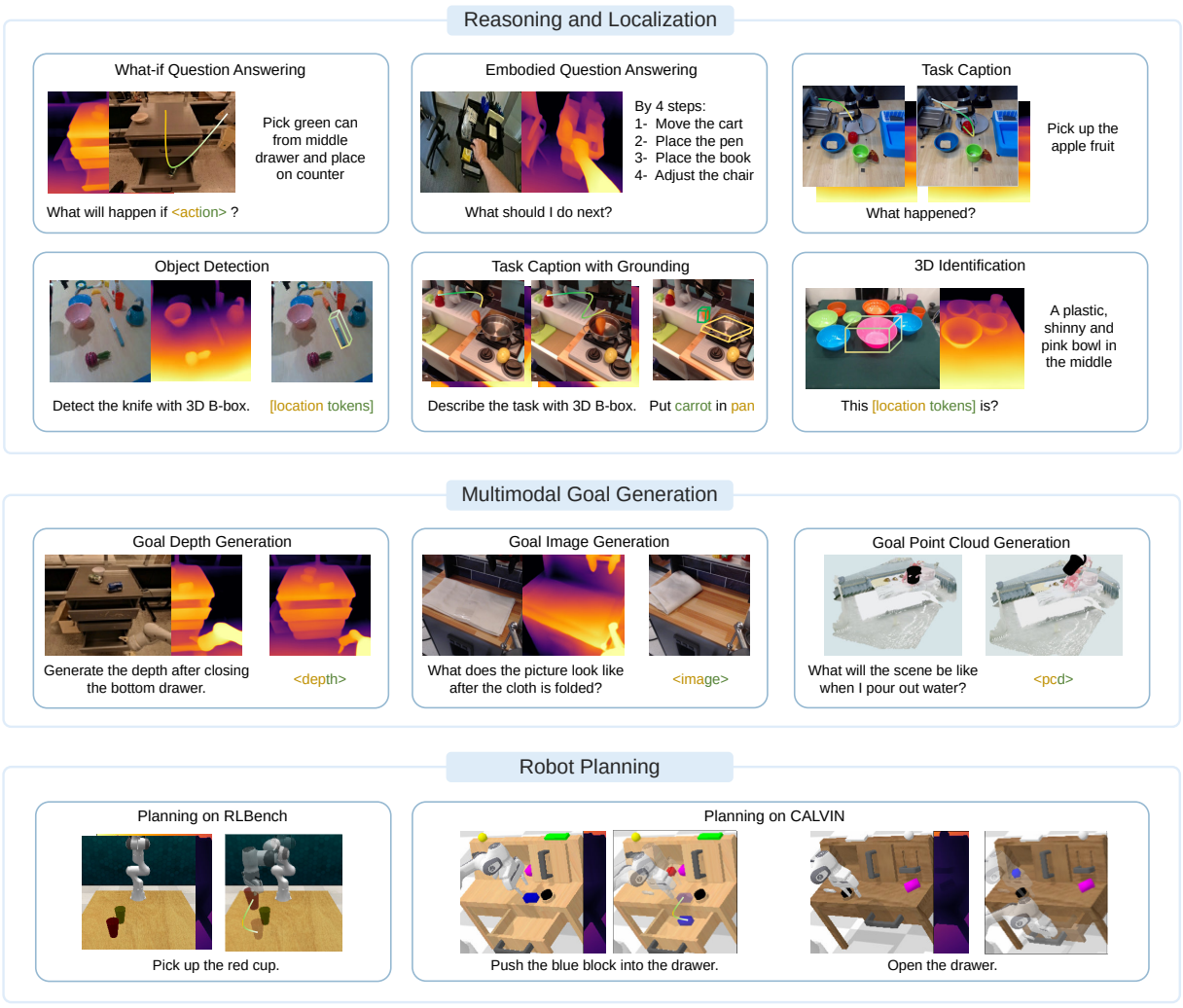

Figure 1. Examples from our 3D Embodied Instruction Tuning Dataset.

the model with 3D understanding ability. Since embodied tasks could not be accomplished via language generation solely and require deeper digging into the dynamic scenes, the manipulated objects as well as actions to interact with the scenes, we add special interactive tokens to the LLM vocabulary (*e.g.,* scene, object, and action tokens). These added tokens enable our model to perform a wider range of embodied tasks and support interleaved 3D-text data. Recognizing the inadequacy of multimodal generation ability in embodied foundation models, we propose to inject the goal generation ability into 3D-VLA. We first pretrain a set of embodied diffusion models for RGBD-to-RGBD and point-to-point generation respectively. To efficiently bridge between the diffusion decoders of various modalities and the LLM embedding space, we employ a projector that aligns multi-modal goal generation in 3D-VLA. It strategically incorporates multimodal signals to specify the type of modality for a generation.

Another challenge for building such a generative world model lies in the lack of data. The embodied datasets in use (Padalkar et al., 2023; Brohan et al., 2022; Jang et al., 2022) mainly consist of 2D images, deficient in 3D-related information. Thus, we curate a large-scale 3D embodied instruction tuning dataset. Specifically, we first gather a diverse collection of datasets that includes real and synthetic data featuring robot manipulations and human-object interactions. For datasets lacking depth data, we utilize a depth estimator to append necessary 3D details and project them to 3D point clouds. Additionally, we design a pipeline to use the off-the-shelf models to extract 3D-related annotations and enrich the language descriptions. In this way, we collect 2M 3D-language-action data pairs, covering various tasks such as task captioning, action prediction, localization, multimodal goal generation, etc, as shown in Figure 1.

To sum up, we have the following contributions:

- We propose 3D-VLA, a new family of 3D vision-language-action embodied foundation models that unify 3D percep-

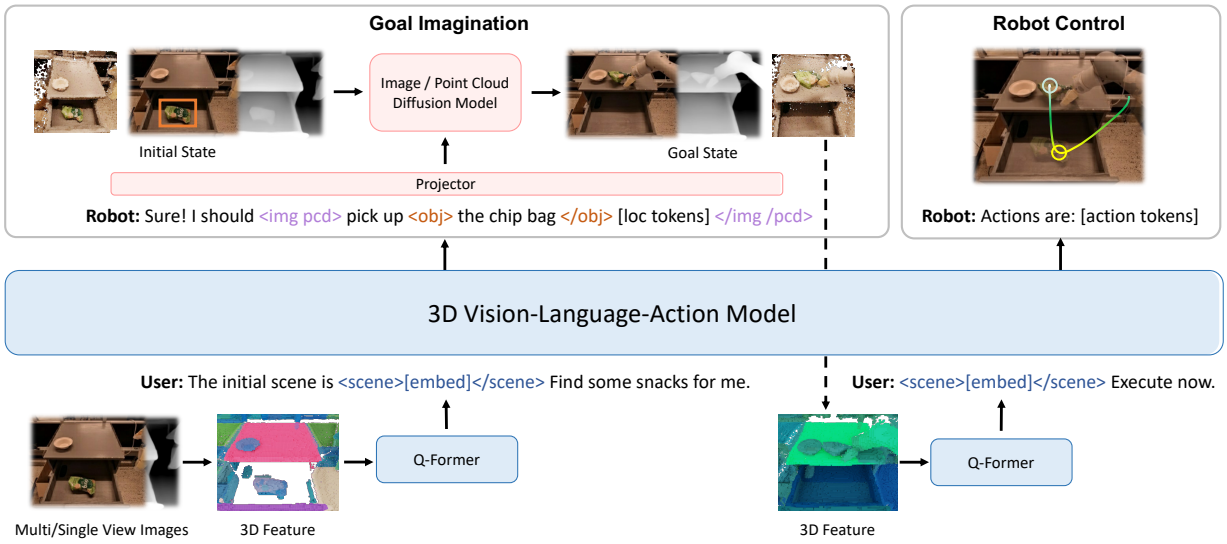

Figure 2. Overview of our 3D-VLA pipeline. The left part shows our goal-generation capability. Our model can imagine the final state image and point cloud based on the user's input. This generated goal state can then be fed back to our model to guide the robot control.

tion, reasoning, and action with a generative world model.

- We create a large-scale 3D embodied instruction tuning dataset addressing the absence of 3D-related information in existing embodied datasets.
- We add interaction tokens to better interact with the environment. We further train diffusion models for goal image and point cloud generation. We utilize a projector to efficiently align LLM output features and diffusion models.
- Our 3D-VLA can conduct a series of tasks, including goal generation (in terms of images, depths, and point clouds), goal-based planning, and embodiment action prediction. It outperforms baseline models by a large margin in these novel embodied tasks. It also outshines baseline models in traditional language-based tasks.

2. Related Works

Multimodal Language Models. Recent Multimodal Language Models have made remarkable advances in various domains, including vision and language understanding (Li et al., 2022; 2023b; Liu et al., 2023; Huang et al., 2023c; Peng et al., 2023; Zhu et al., 2023), interleaved image and text understanding (Alayrac et al., 2022), interleaved image and text generation (Dong et al., 2023). Some more unified models can perceive inputs and generate outputs in arbitrary combinations of text, images, videos, and audio (Wu et al., 2023; Lu et al., 2023). However, none of these models can perceive 3D inputs or output actions according to 3D input.

Vision-Language-Action Models. Previous vision-language models with action output have predominantly leveraged 2D features, thereby lacking the capability of 3D spatial understanding (Driess et al., 2023b; Brohan et al., 2022; 2023). In contrast, our model is guided by 3D features, which are predicted in alignment with goal objectives in our general world model. We are the first to leverage 3D features such as point clouds for action token generation, significantly improving action planning accuracy. Additionally, this pipeline possesses the potential to be extended for applications in real-world scenarios.

3D Foundation Models. Our paper is closely related to the 3D foundation models that integrate 3D features in MLLMs (Hong et al., 2023; Chen et al., 2023b; Qi et al., 2023; Xu et al., 2023; Huang et al., 2023a; Zhou et al., 2023; Guo et al., 2023; Li et al., 2024). These studies have successfully stepped forward to leverage foundation models to comprehend 3D features. However, they primarily focus on analyzing and reasoning in the current observable state of the 3D scenes, thereby revealing a limitation in predicting future features that extend beyond immediate perception. Contrasting with them, we aim to not only understand the perceivable scenes but also predict imperceptible multimodal features guided by specific goals. This capability enables our model to further generate action tokens to interact with the 3D world.

3. 3D Embodied Instruction Tuning Dataset

Recently, benefiting from billion-scale datasets on the internet, VLMs have demonstrated exceptional proficiency in various tasks. Similarly, million-level datasets comprising video-action pairs lay the foundation for embodied VLMs for robot control. However, they mostly don't provide depth or 3D annotations and precise control in robot operations

that necessitate the inclusion of 3D spatial reasoning and interaction. Without 3D information, it is challenging for a robot to comprehend and execute the commands that require 3D spatial reasoning, such as "place the farthest cup into the middle drawer".

To bridge this gap, we build a large-scale 3D embodied instruction tuning dataset that provides sufficient 3D-related information as well as paired text instructions to train our model. We design a pipeline to extract 3D-language-action pairs from existing embodied datasets, obtaining annotations for point clouds, depth maps, 3D bounding boxes, the robot's 7D actions, and textual descriptions. The details are outlined as follows.

3.1. Dataset Collection

Our data are curated from various sources. We provide an overview here, with details available in the Appendix:

Robot Datasets: We select 12 datasets (Brohan et al., 2022; Jang et al., 2022; Walke et al., 2023; Lynch et al., 2023; Feng et al., 2023; Chen et al., 2023a; Dass et al., 2023; Mandlekar et al., 2019; Mees et al., 2023; Shah et al., 2023; Sawhney et al., 2021; Sermanet et al., 2023) from the Open-X Embodiment Dataset (Padalkar et al., 2023). They have high-quality images with linguistic instructions in the real world but lack more in-depth information and 3D annotations. We also select datasets with excellent depth information, such as Dobb-E (Shafiullah et al., 2023) and RH20T (Fang et al., 2023). Additionally, we use datasets collected from two simulator environments, RLBench (James et al., 2020) and CALVIN (Mees et al., 2022).

Human Object Interaction Datasets: Human/hand-object interactions could provide demonstrations that benefit robot decision-making and imitation. Therefore, we utilize several human-object interaction datasets, including datasets without depth information, such as Epic-Kitchens (Damen et al., 2018), and datasets with better 3D annotations, such as HOI4D (Liu et al., 2022).

3.2. Visual Annotations

Estimating depths and optical flows. Given that over 95% of the video datasets for embodied tasks do not provide 3D information, we employ ZoeDepth (Bhat et al., 2023) on each frame of the video from these datasets. Additionally, to better utilize video data, we use RAFT (Teed & Deng, 2020) for optical flow estimation. Optical flow aids in refining the data we generate. Thus, for video segments where the camera pose does not change, we use optical flow to estimate which pixels are the unmoved background. We align the depth maps of these backgrounds across different frames of the same video, multiplying each frame's depth map by a coefficient to ensure depth consistency. After getting the depth maps, we can directly lift the RGB-D images into 3D point clouds using camera intrinsics and poses.

Generating 3D annotations. We aim to generate several 3D-related annotations: 3D bounding boxes of the objects, goal images, depths, or point clouds as the imagination outcomes, as well as robot actions in the 3D space. We first extract the 3D bounding boxes of the objects in the scenes. Such information could benefit 3D models' ability to capture 3D information and attend to the manipulated object for better decision-making. The embodied datasets that serve as sources provide text instructions to describe the commands executed by the robots. We use spaCy (Honnibal & Montani, 2017) to parse the instructions to obtain all noun chunks, including the manipulated object. We utilize a pre-trained grounding model (e.g., Grounded-SAM (Ren et al., 2024)) to obtain the 2D mask of each object. These 2D masks, when lifted to 3D, correspond to parts of the point cloud, allowing us to obtain the 3D bounding boxes of all the objects in space. When selecting masks, the manipulated object is chosen based on the highest confidence value in areas of significant optical flow. Since we reconstruct the depths and point clouds, we could use images, depths, and point clouds in future frames as ground-truth goals. For actions, we use the 7 DoF actions from the provided datasets.

3.3. Language Annotations

Inspired by (Li et al., 2023a; Peng et al., 2023), we propose to generate dense language annotations consisting of tokens (*e.g.,* `<image></image>`; `<pcd></pcd>`) that encompass the 3D annotations (bounding boxes, goal images / depths / point clouds, actions) we generated before, as shown in the prompts in Figure 2.

We use pre-defined language templates with tokens to construct these 3D annotations into prompts and answers. Following (Hong et al., 2023), we use ChatGPT-based prompting to diversify prompts. Specifically, we provide instructions to ChatGPT, as well as our annotated objects and bounding boxes. We also give 2-3 few-shot human-written demonstrations to guide the GPT on the type of data it is instructed to generate. ChatGPT is asked to summarize the information and rewrite the template-generated prompts into more diverse forms. For tasks without pre-defined templates, ChatGPT is also asked to generate prompts and answers as language inputs and outputs of these tasks by itself. We show the detailed templates and prompts to generate all types of data in the Appendix.

4. Methods

4.1. Overview

In this section, we introduce 3D-VLA, a world model for 3D reasoning, goal generation, and decision-making in em-

bodied environments. As shown in Figure 2, we first build our backbone on top of 3D-LLM (Hong et al., 2023), and further enhance the model's capabilities to interact with the 3D world by adding a series of interaction tokens. Next, we inject goal generation ability into 3D-VLA by first pretraining the embodied diffusion models and employing a projector for aligning the LLM and the diffusion models.

4.2. 3D-VLA

4.2.1. Backbone

In the first stage, we develop the 3D-VLA base model following the methodology of 3D-LLM (Hong et al., 2023). Since the dataset we collected is not at the billion-level scale required for training a multi-modal LLM from scratch, we follow the approach of 3D-LLM by leveraging multi-view features to generate 3D scene features. This enables the seamless integration of visual features into a pre-trained VLM with no need for adaptation. Meanwhile, the training datasets for 3D-LLM mostly comprise objects (Deitke et al., 2022) and indoor scenes (Dai et al., 2017; Ramakrishnan et al., 2021), which do not directly align with our embodied setup. Therefore, we choose not to load the 3D-LLM pretrained model. Instead, we utilize BLIP2-FlanT5$_{\text{XL}}$ (Li et al., 2023b) as our pretrained model. During training, we unfreeze both the input and output embeddings for tokens, as well as the weights of the Q-Former.

4.2.2. Interaction Tokens

To enhance the model's comprehension of 3D scenes and facilitate interaction within these environments, we introduce a novel set of interaction tokens. Firstly, We incorporate object tokens `<obj> </obj>` that enclose the object nouns in the parsed sentences (*e.g.*, `<obj> a chocolate bar </obj> [loc tokens] on the table`) so that the model could better capture which objects are manipulated or referred to. Secondly, to better represent spatial information by language, we devise a set of location tokens `<loc0-255>` for grounding referred objects, which are represented by six tokens for the 3D bounding box in the form of AABB. Thirdly, to better encode dynamics with our framework, we introduce the `<scene> </scene>` tokens to enclose the embeddings of a static scene. By composing over the scene tokens, 3D-VLA could comprehend dynamic scenes and manage inputs that interleave 3D scenes and text.

We further enhance the architecture with an expanded set of specialized tokens that represent robotic actions. The robot's actions, with 7 degrees of freedom, are represented by discrete tokens such as `<aloc0-255>`, `<arot0-255>`, and `<gripper0/1>` to denote the arm's intended absolute location, rotation, gripper openness. These actions are separated by token `<ACT_SEP>`.

4.3. Injecting Goal Generation Ability into 3D-VLA

In this section, we introduce how our 3D-VLA performs goal generation in terms of images, depths, and point clouds.

Human beings pre-visualize the final states of the scenes to facilitate action prediction or decision making, which is a key aspect in building world models. Moreover, during preliminary experiments, we also discover that providing the ground-truth final states can enhance the model's reasoning and planning capabilities. However, training an MLLM to generate images, depths, and point clouds is non-trivial. Firstly, state-of-the-art video diffusion models are not tailored for embodied setups. For instance, when asking Runway (Esser et al., 2023) to generate future frames given the instruction "open the drawer", the entire scene is altered to a great extent with regard to view change, unexpected object deformation, and weird texture replacement, as well as layout distortion. Similarly, using the method of DreamLLM (Dong et al., 2023) to directly freeze the stable diffusion trained on internet data, can lead to collapsed outputs. Secondly, how to incorporate diffusion models of various modalities into a single foundation model remains a challenge. Therefore, we propose to inject the ability to generate images, depths and point clouds into 3D-VLA. We first pretrain the embodied diffusion models in terms of different modalities such as images, depths and point clouds, and then align the decoders of these diffusion models to the embedding space of 3D-VLA through an alignment stage.

4.3.1. Pretraining Embodied Diffusion Models for Goal Generation

To address the limitations of current diffusion models for goal generation in an embodied environment, we train RGB-D to RGB-D and point-cloud to point-cloud diffusion models. We utilize our curated 3D-language video data to train a conditional diffusion model that edits the initial state modality based on instructions to generate the corresponding final state modality. The specific training details for these models are as follows: For RGBD to RGBD generation, we employ Stable Diffusion V1.4 (Rombach et al., 2022) as our pretrained model due to the efficiency and quality of image generation by latent diffusion when operating in the latent space of a pretrained VAE (Kingma & Welling, 2013). We concatenate the RGB latent and depth latent as the image condition. Similarly, for point-to-point generation, we use Point-E (Nichol et al., 2022) as the pretrained model, to which we add a point cloud condition input.

4.3.2. Bridging LLM and Goal Generation

After pretraining the diffusion models, we are equipped with various decoders that could generate goals by conditioning the latent spaces in their modalities. Challenges remain as to how to seamlessly incorporate the pretrained

Tasks	Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH-L	EM@1
Embodied QA	3D-LLM*	1.05	0.38	0.15	0.02	12.96	0.91	0.00
	BLIP2 $OPT_{2.7B}$*	7.39	3.17	0.03	0.02	3.87	7.40	3.03
	BLIP2 $FlanT5_{XL}$*	22.84	16.17	12.50	10.11	11.41	32.01	10.31
	$OpenFlamingo_{4B}$*	9.50	6.51	5.14	4.29	6.84	10.40	1.21
	$LLaVA_{7B}$*	11.66	8.06	6.01	4.58	12.59	14.17	5.67
	BLIP2 $FlanT5_{XL}$	37.31	27.20	20.32	15.48	17.80	38.92	15.35
	3D-VLA	**48.34**	**38.55**	**31.72**	**26.80**	**23.72**	**49.33**	**24.53**
Task Caption	3D-LLM*	0.78	0.16	0.07	0.05	0.57	1.33	0.00
	BLIP2 $FlanT5_{XL}$*	8.50	2.07	0.35	0.00	3.40	8.45	0.00
	$OpenFlamingo_{4B}$*	7.61	1.64	0.37	0.00	4.74	9.36	0.00
	$LLaVA_{7B}$*	2.63	0.69	0.16	0.00	2.63	4.65	0.00
	BLIP2 $FlanT5_{XL}$	22.05	11.40	5.72	3.16	8.72	26.12	7.75
	3D-VLA	**55.69**	**45.88**	**39.39**	**34.88**	**27.57**	**62.01**	**29.34**
What-if QA	BLIP2 $FlanT5_{XL}$	28.23	11.47	4.49	0.06	8.27	28.41	5.85
	3D-VLA	**53.09**	**40.94**	**34.34**	**29.38**	**26.83**	**52.82**	**14.7**
Dense Caption	3D-LLM*	0.52	0.22	0.16	0.13	0.34	0.64	0.00
	BLIP2 $FlanT5_{XL}$	36.17	24.72	18.06	13.96	17.83	40.56	13.10
	3D-VLA	**51.90**	**42.83**	**38.11**	**34.62**	**25.25**	**55.91**	**39.49**

Table 1. Evaluation on reasoning ability using held-in data. $*$ denotes zero-shot transfer results without training on our pre-train datasets.

decoders into the LLMs so that 3D-VLA could generate goals with regard to any pretrained modalities conditioned on the input instructions. To bridge the gap between the LLM and the diffusion models of different modalities, we develop an alignment stage into our 3D-VLA. We first introduce additional special tokens such as `<image> </image>` and `<pcd> </pcd>`. These tokens are intricately designed to inform the decoder about the type of modal content to output. Between the enclosing tokens, we supervise the LLM in generating instructions for a robot to execute, which may include object tokens and location tokens, such as `<image> pick up the <obj> apple </obj> [loc tokens] </image>`. Based on this, we can apply a transformer-based projector, which is capable of mapping the decoder features and embeddings from the Large Language Model (LLM) into the space of the DM framework. It plays a crucial role in enhancing the model's capability to understand and generate multi-modal data, establishing a connection between high-level language understanding and multi-modal goal generation. To make training 3D-VLA more efficient and to avoid catastrophic forgetting, we utilize LoRA (Hu et al., 2021) to fine-tune different diffusion models. At the same time, we only train the newly introduced special tokens embeddings, the corresponding embedding output linear layer, and the entire projector. We minimize both the LLM and DM denoising loss.

5. Experiments

3D-VLA is a versatile 3D-based generative world model that can perform reasoning and grounding in the 3D world, imagine multi-modal goal content, and generate actions for robot manipulation. In this section, we evaluate 3D-VLA in three aspects: 3D reasoning and localization, multi-modal goal generation, and embodied action planning.

Methods	IoU	Acc@25	Acc@50
Kosmos-2 (w/ GT Depth)	10.92	12.73	3.85
CoVLM (w/ GT Depth)	19.81	25.39	16.61
3D-VLA	**29.33**	**42.26**	**27.09**

Table 2. Localization results on held-in robotics datasets.

5.1. 3D Reasoning and Localization

Tasks. Our primary focus is on scenes involving robots that are characterized by greater dynamism and a higher degree of interaction, which require a greater level of reasoning and localization abilities. We build several tasks on 3D embodied instruction tuning datasets for learning these abilities in the robotics domain. The tasks include 1) embodied QA on RoboVQA dataset (Sermanet et al., 2023); 2) task captioning on 11 Open-X datasets (Padalkar et al., 2023), where we input the initial and final scenes and ask the agent to reason what has happened; 3) what-if QA on RT-1 dataset (Brohan et al., 2022), where the agent is asked a question that what will happen if some specified actions (represented by action tokens) are executed; 4) dense captioning on 11 Open-X datasets, where the agent need to caption the content specified by a 3d bounding box; 5) localization on 11 Open-X datasets, where the agent is to localize the object mentioned in the robot manipulation instruction. We evaluate 3D-VLA on these tasks using held-in datasets.

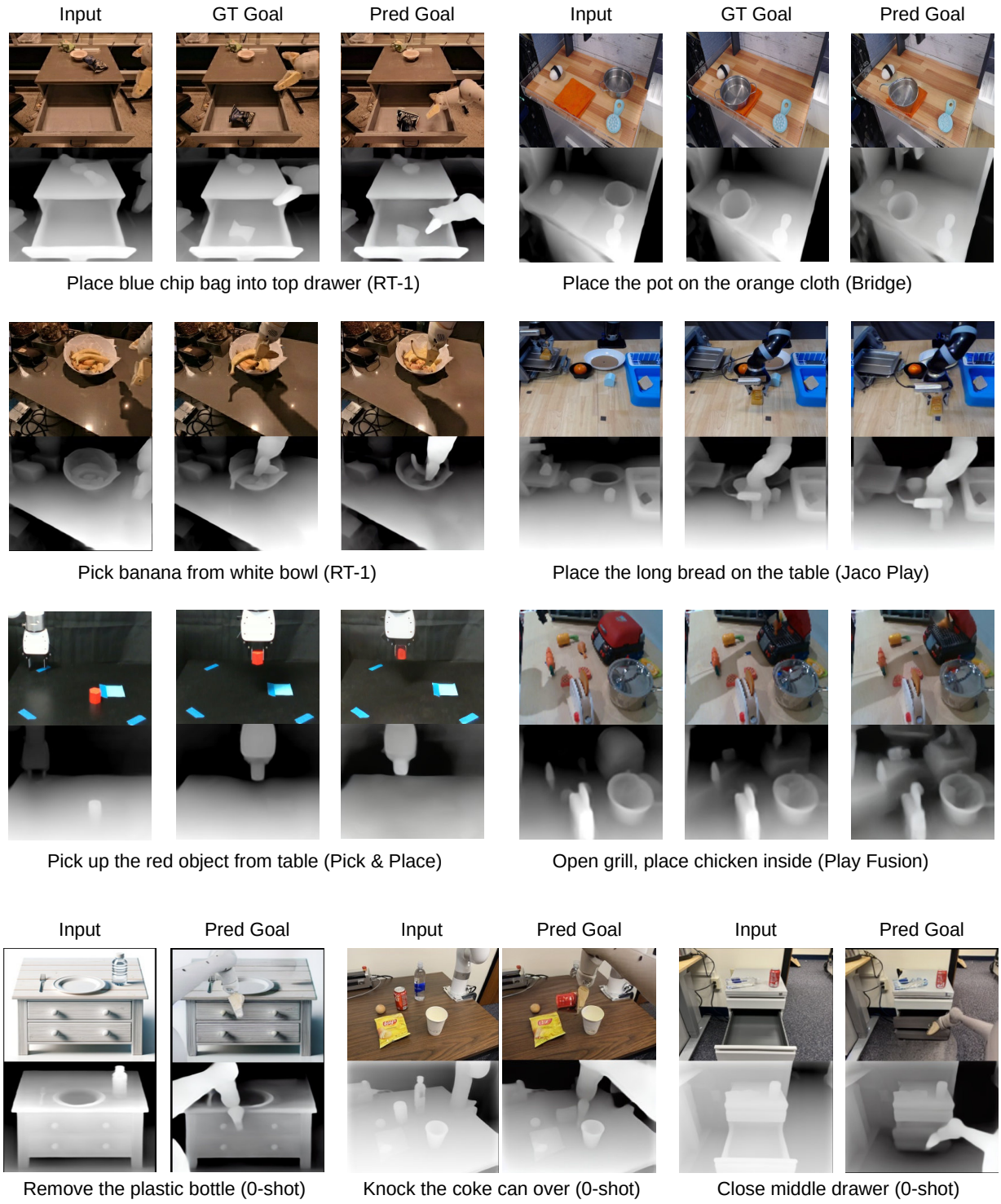

Figure 3. Visualization of generated RGB-D goal images. The results in the first row are sampled from the test set of held-in training data while the second row is the unseen environments gathered from the Internet or daily life.

Method	PSNR ↑	CLIP Sim ↑	SSIM ↑	FID ↓
Instruct-P2P	14.41	0.909	0.389	0.309
SuSIE	15.20	0.898	0.549	0.182
NeXT-GPT	8.86	0.199	0.153	0.432
Instruct-P2P*	16.67	**0.941**	0.628	0.178
3D-VLA w/o Pred BBox	17.02	0.919	0.632	**0.173**
3D-VLA	**17.21**	0.920	**0.636**	0.177

Table 3. RGB image goal generation results. $*$ denotes the model is trained on our pretrained dataset.

Models	P-FID ↓	Chamfer-L_1 ↓
Point-E*	5.241	0.159
3D-VLA w/o Pred BBox	4.914	0.143
3D-VLA	**4.796**	**0.139**

Table 4. Point Cloud goal generation results. $*$ denotes the model is trained on our pretrained dataset.

Baselines. We compare 3D-VLA with 3D-LLM (Hong et al., 2023) and 2D vision-language models, including BLIP2 (Li et al., 2023b), OpenFlamingo (Alayrac et al., 2022), and LLaVA (Liu et al., 2023). We implement these baselines in two ways: 1) zero-shot transfer where we test the released trained model on these new tasks; 2) held-in evaluation where we train the released model on 2D-image-action-language pairs (*i.e.,* , 11 datasets selected from Open-X and RoboVQA dataset). For the localization task, we compare with 2D grounding MLLM, namely Kosmos-2 (Peng et al., 2023) and CoVLM (Li et al., 2023a). Specifically, we use these models to detect 2D bounding boxes in a zero-shot manner and then transfer them to 3D bounding boxes using depth projection.

Result analysis. In Tables 1, 3D-VLA outperforms all 2D VLM methods on language reasoning tasks. We attribute it to the leverage of 3D information, which provides more accurate spatial information for reasoning. Besides, since our dataset contains a bunch of 3D localization annotations, 3D-VLA learns to localize the relevant objects, which helps the model focus more on key objects for reasoning. Moreover, we find that 3D-LLM performs poorly on these robotic reasoning tasks, which demonstrates the necessity of collecting and training on a robotics-related 3D dataset. In Table 2, 3D-VLA demonstrates a marked superiority over the 2D baseline methods in terms of localization performance. This finding serves as compelling evidence of the efficacy of our annotation process, which supplies a substantial quantity of 3D annotations, thereby facilitating the acquisition of robust 3D localization capabilities within our model.

5.2. Multi-modal Goal Generation

Tasks. We quantitatively evaluate the RGB goal and point cloud goal generation capability of 3D-VLA on Open-X test sets. We randomly sample 4000 episodes from the Open-X test set which 3D-VLA does not see in the training process.

Baselines. For image generation, we compare 3D-VLA with three types of image generation methods: 1) image-editing methods Instruct-P2P (Brooks et al., 2023); 2) goal image/video generation methods SuSIE (Black et al., 2023); 3) LLMs with image generation ability NeXT-GPT (Wu et al., 2023). For point cloud generation, we compare with text-to-3D diffusion model Point-E (Nichol et al., 2022).

Quantitative results. The image goal generation results are shown in Table 3. When compared with the existing generation methods that directly zero-shot transfers to the robotics domain (rows 1, 2, 3 in Table 3), 3D-VLA achieves a promising performance in terms of most metrics. This underscores the importance of training a world model using datasets specifically designed for robotics applications. Even in a direct comparison with Instruct-P2P*, which was trained on the same robotics datasets we employed (row 4 in the table), 3D-VLA consistently outperforms it. This highlights that the integration of a large language model into 3D-VLA results in a more comprehensive and insightful comprehension of robotics manipulation instructions, leading to better goal image generation performance. Furthermore, when we exclude the predicted bounding box from the input prompt (row 5), we observe a slight decrease in performance. This observation confirms the effectiveness of using these intermediate predicted bounding boxes as they assist the model in comprehending the overall scene, allowing the model to allocate more attention to the specific object mentioned in the given instruction, ultimately enhancing its ability to imagine the final goal images.

The point cloud generation results are presented in Table 4. 3D-VLA with intermediate predicted bounding boxes performs the best. This outcome reinforces the significance of incorporating large language models and precise object localization in the context of comprehending both the instruction and the scene.

Qualitative results. In the first row of Figure 3, we visualize the generated RGB-D goal images on the test set of RT-1 (Brohan et al., 2022) and Jaco Play (Dass et al., 2023) datasets. These samples are not seen in the training process. Given the initial scenes and instructions, the 3D-VLA model consistently exhibits the capability to maintain the background elements unchanged while accurately identifying the target object of interaction and correctly modifying the states of these identified objects following the provided instructions. The generated RGB-D goal images closely align both in terms of visual appearance and semantic content with the ground truth goal. In addition to our controlled experimental settings, we extended our testing to encompass scenes captured from the internet or everyday life. In these diverse and uncontrolled environments, our 3D-VLA model consistently and robustly demonstrated its efficacy.

	Put Knife	Take Umbrella	Pick up Cup	Pick up Cup (unseen)
LanCon-Learn	28.8	45.6	23.2	-
LanCon-Learn w/ His.	32.2	50.8	**44.2**	-
3D-VLA w/o Goal	58	68	34	24
3D-VLA	**68**	**80**	40	28

Table 5. Evaluation of action planning on RLBench dataset.

	Tasks completed in a row					
	1	2	3	4	5	Avg Len
MCIL	28.2	2.5	0.3	0.0	0.0	0.31
3D-VLA	**44.7**	**16.3**	**8.1**	**1.6**	0.0	**0.71**

Table 6. Evaluation of action planning on CALVIN dataset.

5.3. Embodied Action Planning

Tasks. We evaluate the ability of 3D-VLA for robot arm action prediction on two benchmarks, namely RLBench (James et al., 2020) and CALVIN (Mees et al., 2022). We select three tasks from RLBench for evaluation. Besides, we also select variation-1 from the pick-up-cup task as an unseen task to test the model's generalization ability. We show the performance of 3D-VLA in executing more complex tasks that require more 3D visual reasoning in the Appendix. For CALVIN, we evaluate our model under the long-horizon multi-task language control setting, where the agent is required to execute 5 tasks sequentially. We train the agent on scenes A, B, C, D and test on scene D.

Baselines. For RLBench, we compare our model 3D-VLA with LanCon-Learn (Silva et al., 2021), which is a multi-task approach that can predict actions based on instruction-conditioned inputs. For CALVIN, we compare with MCIL (Lynch & Sermanet, 2020), which is a conditional sequence-to-sequence variational autoencoder.

Result analysis. As shown in Table 5, 3D-VLA surpasses or matches the baseline performance in most tasks within the RLBench action prediction, showing its planning capability. It's worth noting that the baseline uses history observations, object states, and current state information, whereas we only execute via open-loop control. Additionally, our 3D-VLA model outperforms 3D-VLA w/o Goal by a lot on the Take Umbrella and Pick Up Cup tasks. This is because the imagined goal guides the robotic arm to move to the specific location or determine the color of the object. On the other hand, the performance on the task put the knife on the chopping board is same across both settings, as most failures might be due to object collisions. In Table 6, 3D-VLA also achieves promising results in CALVIN. We attribute the superiority to the ability to localize the objects of interest and imagine the goal state, which provides rich information for inferring actions.

6. Limitation

Difficulty in precise control. In RLBench, we are unable to successfully execute the task of picking up small cubes, as the 3D features and discrete action tokens make it difficult to accurately locate and manipulate these small objects.

Hallucination of the diffusion model. Occasionally, the output of the diffusion model is uncontrollable, with issues including 1) object consistency, where the texture and shape of the moved object change; 2) object disappearance, where objects may be removed during the denoising process; 3) and a certain probability of the goal image generating future frames of incorrect tasks or not changing at all. During training, we use the results predicted by the diffusion model, so that 3D-VLA could adapt to its own generated results.

Issues with depth in the real world. In the simulator, sensors could obtain good quality depth maps and point clouds. However, in real-world applications, depth does not have a unified scale across different scenes, making the predictions with stable diffusion imperfect, and the point clouds quite noisy. Future improvements include designing a better depth prediction decoder and using filters and other algorithms to optimize point clouds and enhance the model's robustness.

The long-tail distribution and datasets with high variance in quality. We find that in datasets such as RT1, and BridgeV2 (Walke et al., 2023), the scores of the goal generation and language-related tasks are higher due to their higher annotation quality and image quality; however, in datasets like BC_Z (Jang et al., 2022), and Roboturk (Mandlekar et al., 2019), the scores are lower.

7. Conclusion

In this paper, we introduce 3D-VLA, a generative world model that can reason, understand, generate, and plan in the embodied environment. We devise a novel data generation pipeline to construct a dataset including 2M 3D-Language-action data pairs to train our model. These data enable it to perform diverse tasks such as task caption, localization, goal image/point cloud generation, action prediction, etc. Our model uses 3D-LLM as the backbone and introduces interaction tokens to interact with the environment. We train a image to image and point to point diffusion model for embodied AI. They are further aligned by a projector with the LLM to enhance the LLM's multimodal generation capabilities. The experiment further shows that our 3D-VLA has stronger capabilities in embodied tasks than the 2D baseline.

Impact Statement

This paper introduces research aimed at pushing the boundaries of Machine Learning in the realm of robot manipulation. Given that robots operate in the physical world, the potential for collisions with objects and humans arises when the robot system is not adequately configured. To mitigate this issue, our approach involves initial training in a simulator environment followed by real-world deployment under human supervision, to minimize any adverse impacts.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Bhat, S. F., Birkl, R., Wofk, D., Wonka, P., and Müller, M. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Black, K., Nakamoto, M., Atreya, P., Walke, H., Finn, C., Kumar, A., and Levine, S. Zero-shot robotic manipulation with pretrained image-editing diffusion models, 2023.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

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

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions, 2023.

Chen, L., Bahl, S., and Pathak, D. Playfusion: Skill acquisition via diffusion from language-annotated play. In *Conference on Robot Learning*, pp. 2012–2029. PMLR, 2023a.

Chen, S., Chen, X., Zhang, C., Li, M., Yu, G., Fei, H., Zhu, H., Fan, J., and Chen, T. Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning, 2023b.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017.

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 720–736, 2018.

Dass, S., Yapeter, J., Zhang, J., Zhang, J., Pertsch, K., Nikolaidis, S., and Lim, J. J. Clvr jaco play dataset, 2023. URL https://github.com/clvrai/clvr_jaco_play_dataset.

Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., and Farhadi, A. Objaverse: A universe of annotated 3d objects, 2022.

Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., Kong, X., Zhang, X., Ma, K., and Yi, L. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023a.

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., and Florence, P. Palm-e: An embodied multimodal language model, 2023b.

Esser, P., Chiu, J., Atighehchian, P., Granskog, J., and Germanidis, A. Structure and content-guided video synthesis with diffusion models, 2023.

Fang, H.-S., Fang, H., Tang, Z., Liu, J., Wang, J., Zhu, H., and Lu, C. Rh20t: A robotic dataset for learning diverse skills in one-shot. *arXiv preprint arXiv:2307.00595*, 2023.

Feng, Y., Hansen, N., Xiong, Z., Rajagopalan, C., and Wang, X. Finetuning offline world models in the real world. *arXiv preprint arXiv:2310.16029*, 2023.

Guhur, P.-L., Chen, S., Pinel, R. G., Tapaswi, M., Laptev, I., and Schmid, C. Instruction-driven history-aware policies for robotic manipulations. In *Conference on Robot Learning*, pp. 175–187. PMLR, 2023.

Guo, Z., Zhang, R., Zhu, X., Tang, Y., Ma, X., Han, J., Chen, K., Gao, P., Li, X., Li, H., and Heng, P.-A. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following, 2023.

Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., and Gan, C. 3d-llm: Injecting the 3d world into

large language models. *arXiv preprint arXiv:2307.12981*, 2023.

Hong, Y., Zheng, Z., Chen, P., Wang, Y., Li, J., and Gan, C. Multiply: A multisensory object-centric embodied large language model in 3d world. *arXiv preprint arXiv:2401.08577*, 2024.

Honnibal, M. and Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Huang, H., Wang, Z., Huang, R., Liu, L., Cheng, X., Zhao, Y., Jin, T., and Zhao, Z. Chat-3d v2: Bridging 3d scene and large language models with object identifiers, 2023a.

Huang, J., Yong, S., Ma, X., Linghu, X., Li, P., Wang, Y., Li, Q., Zhu, S.-C., Jia, B., and Huang, S. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023b.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Liu, Q., et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023c.

James, S., Ma, Z., Arrojo, D. R., and Davison, A. J. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Jang, E., Irpan, A., Khansari, M., Kappler, D., Ebert, F., Lynch, C., Levine, S., and Finn, C. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pp. 991–1002. PMLR, 2022.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Li, J., Chen, D., Hong, Y., Chen, Z., Chen, P., Shen, Y., and Gan, C. Covlm: Composing visual entities and relationships in large language models via communicative decoding. *arXiv preprint arXiv:2311.03354*, 2023a.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023b.

Li, Z., Zhang, C., Wang, X., Ren, R., Xu, Y., Ma, R., and Liu, X. 3dmit: 3d multi-modal instruction tuning for scene understanding, 2024.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., and Yi, L. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21013–21022, 2022.

Lu, J., Clark, C., Zellers, R., Mottaghi, R., and Kembhavi, A. UNIFIED-IO: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=E01k9048soZ.

Lynch, C. and Sermanet, P. Language conditioned imitation learning over unstructured data. *arXiv preprint arXiv:2005.07648*, 2020.

Lynch, C., Wahid, A., Tompson, J., Ding, T., Betker, J., Baruch, R., Armstrong, T., and Florence, P. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.

Mandlekar, A., Booher, J., Spero, M., Tung, A., Gupta, A., Zhu, Y., Garg, A., Savarese, S., and Fei-Fei, L. Scaling robot supervision to hundreds of hours with roboturk: Robotic manipulation dataset through human reasoning and dexterity. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1048–1055. IEEE, 2019.

Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press, 07 2010. ISBN 9780262514620. doi: 10.7551/mitpress/9780262514620.001.0001. URL https://doi.org/10.7551/mitpress/9780262514620.001.0001.

Mees, O., Hermann, L., Rosete-Beas, E., and Burgard, W. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7327–7334, 2022.

Mees, O., Borja-Diaz, J., and Burgard, W. Grounding language with visual affordances over unstructured data. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.

Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., and Chen, M. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

Padalkar, A., Pooley, A., Jain, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Rai, A., Singh, A., Brohan, A., et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

Palmer, S. The effects of contextual scenes on the identification of objects. *Memory & Cognition*, 3:519–526, 01 1975.

Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., and Wei, F. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Pylyshyn, Z. *Seeing and Visualizing: It's Not What You Think*. 01 2003. ISBN 9780262316316. doi: 10.7551/mitpress/6137.001.0001.

Qi, Z., Fang, Y., Sun, Z., Wu, X., Wu, T., Wang, J., Lin, D., and Zhao, H. Gpt4point: A unified framework for point-language understanding and generation, 2023.

Ramakrishnan, S. K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A. X., Savva, M., Zhao, Y., and Batra, D. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai, 2021.

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., and Zhang, L. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Sawhney, A., Lee, S., Zhang, K., Veloso, M., and Kroemer, O. Playing with food: Learning food item representations through interactive exploration. In *Experimental Robotics: The 17th International Symposium*, pp. 309–322. Springer, 2021.

Sermanet, P., Ding, T., Zhao, J., Xia, F., Dwibedi, D., Gopalakrishnan, K., Chan, C., Dulac-Arnold, G., Maddineni, S., Joshi, N. J., Florence, P., Han, W., Baruch, R., Lu, Y., Mirchandani, S., Xu, P., Sanketi, P., Hausman, K., Shafran, I., Ichter, B., and Cao, Y. Robovqa: Multimodal long-horizon reasoning for robotics. In *arXiv preprint arXiv:2311.00899*, 2023.

Shafiullah, N. M. M., Rai, A., Etukuru, H., Liu, Y., Misra, I., Chintala, S., and Pinto, L. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.

Shah, R., Martín-Martín, R., and Zhu, Y. MUTEX: Learning unified policies from multimodal task specifications. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=PwqiqaaEzJ.

Silva, A., Moorman, N., Silva, W., Zaidi, Z., Gopalan, N., and Gombolay, M. Lancon-learn: Learning with language to enable generalization in multi-task manipulation. *IEEE Robotics and Automation Letters*, 7(2):1635–1642, 2021.

Teed, Z. and Deng, J. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 402–419. Springer, 2020.

Walke, H. R., Black, K., Zhao, T. Z., Vuong, Q., Zheng, C., Hansen-Estruch, P., He, A. W., Myers, V., Kim, M. J., Du, M., et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023.

Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., and Lin, D. Pointllm: Empowering large language models to understand point clouds, 2023.

Zhou, J., Wang, J., Ma, B., Liu, Y.-S., Huang, T., and Wang, X. Uni3d: Exploring unified 3d representation at scale, 2023.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A. Implementation Details

Our training process is divided into three steps: 1) we first finetune the diffusion models on our collected robotics datasets to bridge the domain gap; 2) we then train a multimodal LLM on language-related tasks; 3) we finally update the LoRA parameters in the diffusion models, the proposed special tokens and the projectors to align the diffusion models and the multimodal LLM. Details are shown as follows:

Step1: Adapting Diffusion Models to Robotics Data Domain. We train both the image and point cloud diffusion model, which takes the noisy goal state (latent), the initial state (latent), the time step t, and the CLIP embeddings of the instruction as inputs. The objective is to minimize the denoising loss. We train our RGB-D editing diffusion model on 6×16 V100 GPUs. We train at 256×256 resolution with batch size of 32 on each GPU. We use a learning rate of 10^{-4}. We apply random horizontal flip augmentation. For the point cloud diffusion model, we train it on 6 × 16 V100 GPUs with a learning rate of 10^{-4}, and the batch size per GPU is 2.

Step2: Training Multimodal LLM. We train M-LLM on all language-related tasks, including QA, captioning, localization, action prediction, etc. For tasks such as action prediction, we use either the ground truth goal or the diffusion model predicted goal in input prompts for training. The objective is to minimize the LLM cross entropy loss. We use pretrained BLIP-2 FlanT5 as LLM backbone. We train 3D-VLAs on 6×32 V100s. The batch size is set to 4 on each node during training. Additionally, we apply a linear warmup of the learning rate during the initial 1K steps, increasing from 10^{-8} to 10^{-5}, followed by a cosine decay with a minimum learning rate of 10^{-6}.

Step3: Alignment between Diffusion Models and M-LLM We only update the newly added special tokens (`<img>`, `</img>`, `<pcd>` and `</pcd>`) in the input and output embeddings of the LLM, as well as the projector and the parameters fine-tuned in the DM using LoRA. The objective for this stage is to minimize a combination of the LLM token loss and the diffusion model loss. We train 3D-VLAs for a maximum of epochs of 30 on 6×64 V100s. The batch size is set to 2 on each node for training. The AdamW optimizer is used, with $beta_1 = 0.9, beta_2 = 0.999$, and a weight decay of 0.05.

B. Datasets Details

B.1. Details on Question Templates

In this section, we show the question templates for data generation in Table 7. We designed corresponding templates for six tasks. We design the templates for six tasks, and we replace the `INSTRUCTION`, `OBJECT`, `LOCATION`, and `ACTION` in each template with the information processed from each sample.

Tasks	Templates
Verification	The initial scene is <scene></scene> and the current scene is <scene></scene>. Instruction: `INSTRUCTION`. Finished? Answer: [yes/no]
Task Caption	The initial scene is <scene></scene> and the final scene is <scene></scene>. Describe the task. Answer: `INSTRUCTION`.
Localization	The scene is <scene></scene>. Locate: `OBJECT`. Answer: `LOCATION`
Dense Caption	The scene is <scene></scene>. What is located at `LOCATION`? Answer: `OBJECT`
Image or Point Cloud Generation	The initial scene is <scene></scene>. Instruction: `INSTRUCTION`. Generate the goal image (or point cloud). Answer: <image> (<pcd>) `INSTRUCTION` </image> (</pcd>)
Action Prediction	<scene></scene>. `INSTRUCTION`. Predict {`key/dense`} actions. Answer: `ACTION`.

Table 7. Detailed on Question Templates.

B.2. Details on ChatGPT-based Prompting

In this section, we show the prompt used in ChatGPT-based data generation in Figure 4. The ChatGPT version used in our paper is GPT-3.5-turbo-0125. We generate data for all seven tasks, and we provide all the information in the form of text, such as the instructions performed by the robot, total execution time, objects and their locations in the scene, etc. Additionally, for each prompt, we provide two manually written samples as guidance to direct ChatGPT towards more natural data generation.

```
messages=[{"role": "system" , "content": "

You are an AI visual assistant and a question-answering generator capable of analyzing dynamic 3D scenes.

Suppose you have observed a robotic arm successfully executing an instruction: [instruction].
The scene's initial state is <initial scene> and <final scene>, where the final scene is the [num frame] frame, and we assume that the task was definitely not completed in the first 2/3 of the time.
You have the action sequence <action> of the robot arm.
In this instruction, the initial positions of these objects are [object + location]. Note that the location is the center points of objects represented by a 3D coordinate (x, y, z) with units of meters.

Utilizing all the information above, you can choose to rewrite the instruction while retaining its original meaning.
Further, you need to generate multiple rounds of dialogue or a question answer pair, which should correspond to one of the following tasks:
        1. Verification: Given the initial state and a mid-state frame, ask if the robot has completed the instruction.
        2. Task Caption: Given the initial and final states, ask what task the robot performed.
        3. Embodied QA: Please conduct some questions and answers about the current dynamic scene.
        4. Localization: Detect where objects are, answer the location of the objects.
        5. Dense Caption: Given the location of objects, answer with a description of those objects.
        6. Image or Point Cloud Generation: Given the initial scene and instruction, generate an image or point cloud of the final state.
           If choosing this task, enclose the instruction with the <image> </image> or <pcd> </pcd> token to represent generation.
        7. Action Prediction: Given the initial scene, or having both initial and final scenes, predict actions. You can include a simple task decomposition, but the length of the decomposition must not exceed 3.

"}]
```

Figure 4. Prompt for ChatGPT-based data generation.

B.3. Details on Dataset Construction

We show the number of the episodes and how we use them in table Table 8. We utilize two main categories of datasets, namely robotics datasets and human object interaction (HOI) datasets. For the former, we filtered out complex scene datasets to prevent the Grounded-SAM from detecting incorrect object locations. However, within the same robotics dataset, the background settings are largely the same. Therefore, in the Goal Generation tasks, we included HOI datasets to better allow the diffusion model to learn diverse scenes, object interaction methods, etc.

C. More Complex Tasks about Embodied Action Planning

We expand the set of tasks evaluated from RLBench, focusing on three main categories (Guhur et al., 2023): Tool Use, Visual Occlusion, and Screw. Tool Use requires the robot to interact with objects to perform tasks and we select sweep dirt to dustpan to represent this task category. For the Screw category, it requires precise rotational movements, and thus we choose the change clock task. Visual Occlusion tasks typically involve interactions with large items, where we choose open/close drawer as a representative task. The accuracies are shown in Table 9.

	Sweep Dirt	Open Drawer	Close Drawer	Change Clock
3D-VLA	86	72	96	18

Table 9. Evaluation of action planning on RLBench dataset.

D. More Visualization Results about Goal Generation

We show more qualitative examples in Figure 5, 6.

Dataset	# of Used Episodes	Reasoning and Perception					Goal Generation			Decision Making
		Embodied QA What-if QA	Task Caption (w/ Object Grounding)	Dense Caption	Verification	Detection	Image	Depth	Point Cloud	Action Prediction
Robotics Datasets	305k	✓	✓	✓	✓	✓	✓	✓	✓	✓
BC-Z	40k	✓	✓	✓	✓	✓	✓	✓	✓	✓
Bridge	25k	✓	✓	✓	✓	✓	✓	✓	✓	✓
CALVIN	10k	-	-	-	✓	✓	✓	✓	✓	✓
Dobb-E	20k	✓	✓	-	✓	-	✓	✓	✓	✓
Fractal	70k	✓	✓	✓	✓	✓	✓	✓	✓	✓
Jaco Play	0.9k	✓	✓	✓	✓	✓	✓	✓	✓	✓
Lang Table	13k	✓	✓	-	-	-	✓	✓	✓	✓
Mutex	1.5k	✓	✓	✓	✓	✓	✓	✓	✓	✓
Pick&Place	1.3k	✓	✓	✓	✓	✓	✓	✓	✓	✓
Play Fusion	0.5k	✓	✓	✓	✓	✓	✓	✓	✓	✓
Playing Food	4.2k	✓	✓	-	✓	-	✓	✓	✓	✓
RH20T	2.0k	✓	✓	✓	✓	✓	✓	✓	✓	✓
RLBench	50k	-	-	-	✓	✓	✓	✓	✓	✓
Roboturk	2.0k	-	-	-	✓	-	✓	✓	✓	✓
RoboVQA	61k	✓	-	-	-	-	-	-	-	-
Taco Play	3.2k	✓	✓	✓	✓	✓	✓	✓	✓	✓
HOI Datasets	11k	-	-	-	-	-	✓	✓	✓	-
Epic Kitchen	6k	-	-	-	-	-	✓	✓	✓	-
HOI4D	5k	-	-	-	-	-	✓	✓	✓	-
All Datasets	316k	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 8. Datasets used in our paper. We categorize them into four categories: Robotics, HOI, and Room datasets.

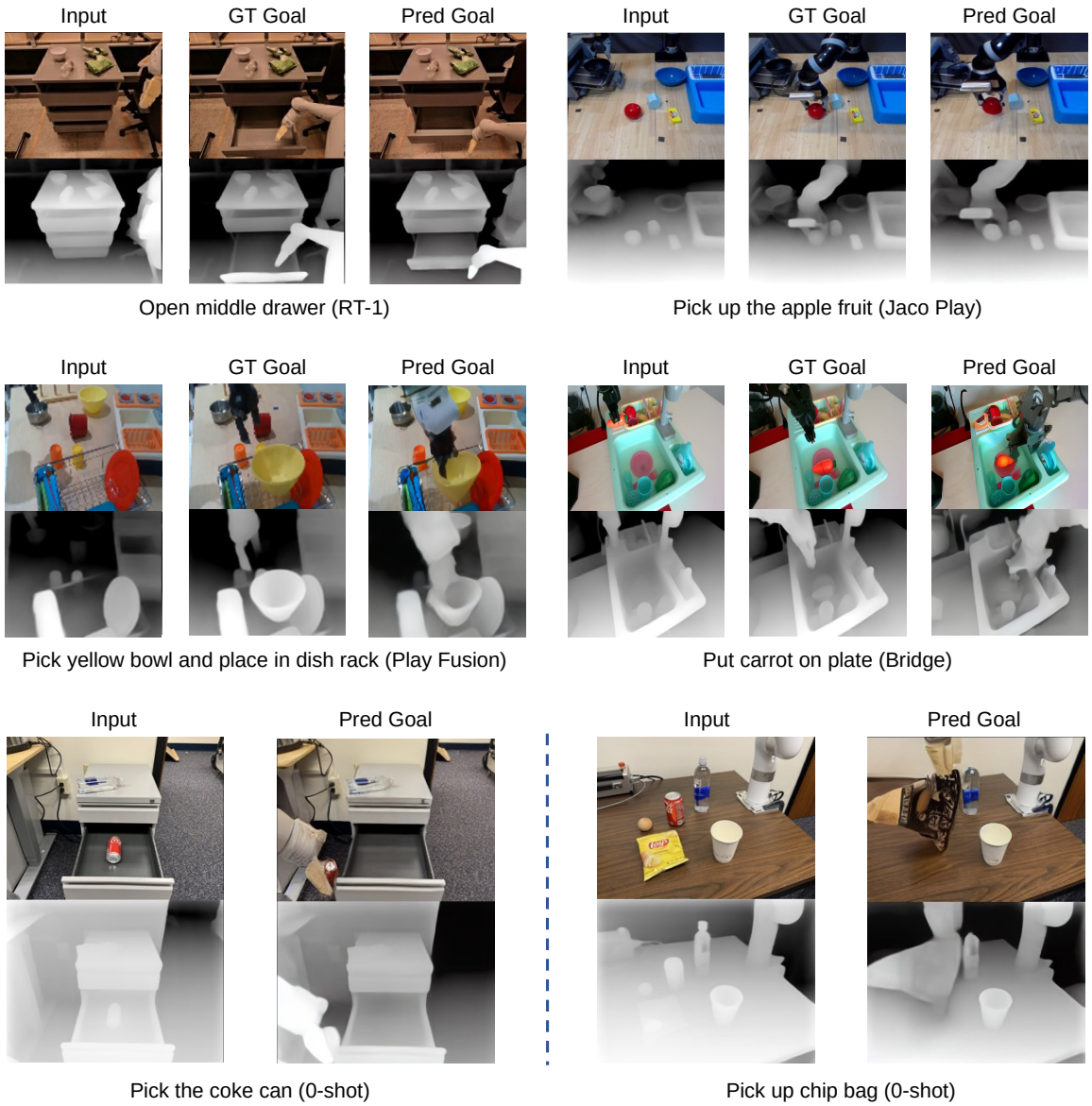

Figure 5. Visualization of generated RGBD goal images. The results in the first row are sampled from the test set of held-in training data while the second row are the unseen environments gathered from daily life.

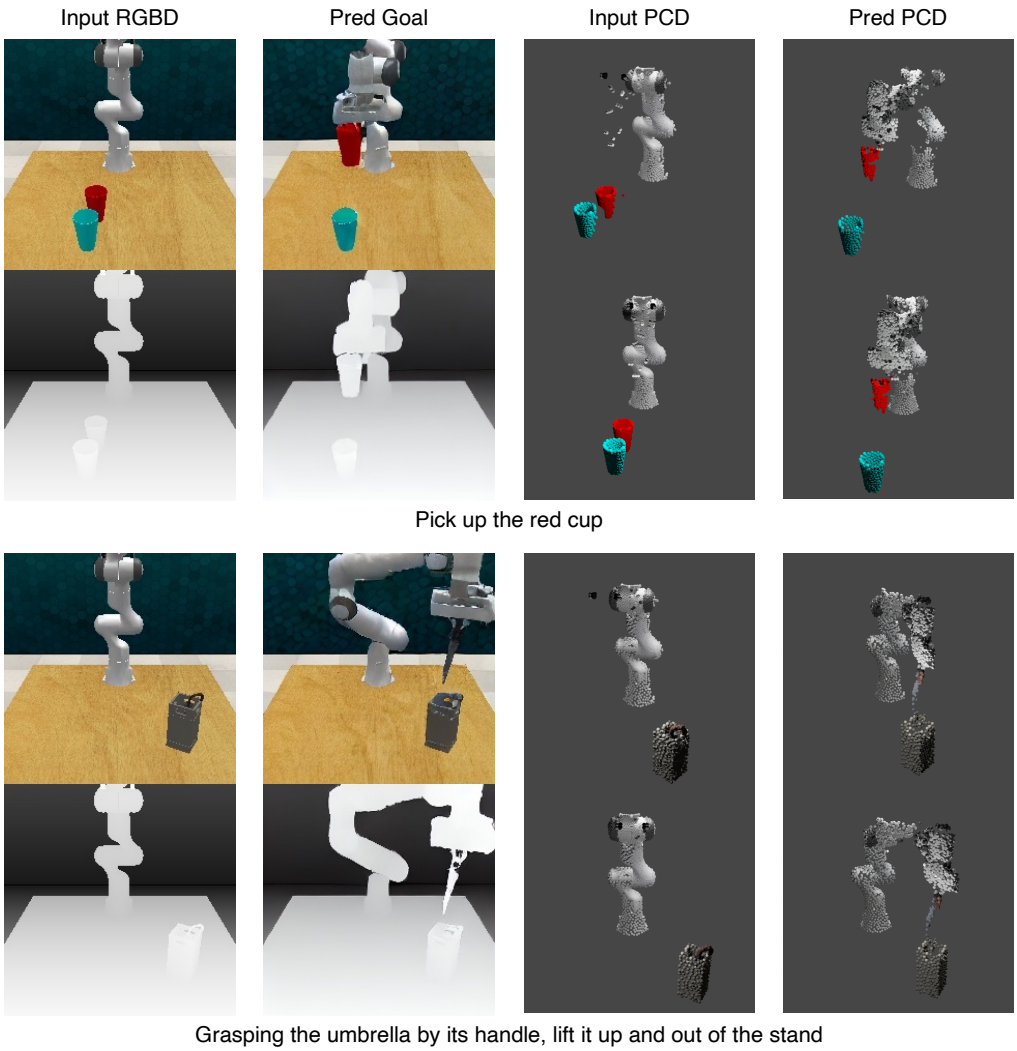

Figure 6. Visualization of generated RGB-D goal images and goal point cloud. (RLBench)