

# AR-VRM: Imitating Human Motions for Visual Robot Manipulation with Analogical Reasoning

Dejie Yang<sup>1</sup>, Zijing Zhao<sup>1</sup>, Yang Liu<sup>1,2\*</sup>

<sup>1</sup> Wangxuan Institute of Computer Technology, Peking University

<sup>2</sup> State Key Laboratory of General Artificial Intelligence, Peking University  
{ydj,zijingzhao}@stu.pku.edu.cn, yangliu@pku.edu.cn

## Abstract

Visual Robot Manipulation (VRM) aims to enable a robot to follow natural language instructions based on robot states and visual observations, and therefore requires costly multi-modal data. To compensate for the deficiency of robot data, existing approaches have employed vision-language pre-training with large-scale data. However, they either utilize web data that differs from robotic tasks, or train the model in an implicit way (e.g., predicting future frames at the pixel level), thus showing limited generalization ability under insufficient robot data. In this paper, we propose to learn from large-scale human action video datasets in an explicit way (i.e., imitating human actions from hand keypoints), introducing Visual Robot Manipulation with Analogical Reasoning (AR-VRM). To acquire action knowledge explicitly from human action videos, we propose a keypoint Vision-Language Model (VLM) pretraining scheme, enabling the VLM to learn human action knowledge and directly predict human hand keypoints. During fine-tuning on robot data, to facilitate the robotic arm in imitating the action patterns of human motions, we first retrieve human action videos that perform similar manipulation tasks and have similar historical observations, and then learn the Analogical Reasoning (AR) map between human hand keypoints and robot components. Taking advantage of focusing on action keypoints instead of irrelevant visual cues, our method achieves leading performance on the CALVIN benchmark and real-world experiments. In few-shot scenarios, our AR-VRM outperforms previous methods by large margins, underscoring the effectiveness of explicitly imitating human actions under data scarcity. Code available at <https://github.com/idejie/ar>.

## 1. Introduction

Visual robot manipulation (VRM) is an essential task in the robotics field [3, 5, 11]. Based on current state and vi-

\*Corresponding author

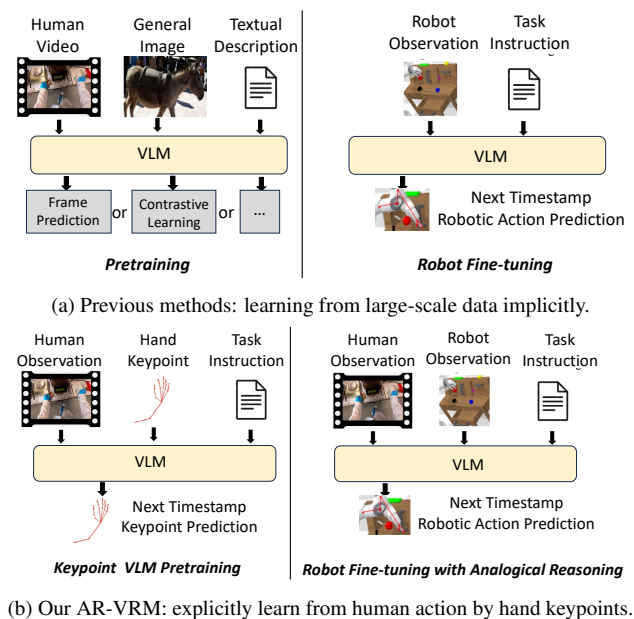


Figure 1. **Demonstration of the differences between our framework and previous methods:** we propose to learn from human actions explicitly by hand keypoints with analogical reasoning.

sual inputs, the robot is required to execute some actions according to human natural language instructions, including tasks such as object grasping, placement, and assembly [1, 2, 23, 31]. Training such task requires multi-modal data, including paired images, natural language instructions, and robot action states under specific scenarios where robots work, which is costly and requires time-consuming human demonstrations to manipulate robots. Therefore, the performance of VRM is often limited by the **scarcity of robot trajectories and annotations for training** [1, 31].

To compensate for the deficiency of robot data, existing approaches utilize large-scale vision-language data pretraining and fine-tune the model under robot scenarios. But,

these pretraining data do not directly reflect object manipulation tasks, such as visual-question-and-answer datasets about animals and diets, and it is hard to provide relevant knowledge guidance for robot manipulation tasks. Some studies [20, 31] have utilized human action video datasets that are more similar to robotic manipulation tasks for pre-training, but they train the model in implicit ways, either by contrastive learning in feature space or by predicting future frames with pixel-level generative models. Though acquiring human action knowledge, these approaches inevitably introduce irrelevant background information or pixel-level noise, which limits their performance on data-scarce VRM tasks.

In this paper, we propose to imitate human motions explicitly from a large-scale human action video dataset, training the model with human hand keypoints to learn from the motion itself, ignoring the irrelevant visual information. To achieve this goal, two main challenges exist: (1) How to extract human action knowledge in terms of hand keypoints from large-scale human video datasets? (2) How to train the robot, which differs from the human arm, to imitate human actions, i.e., building the correlation between the robotic components and human keypoints for the manipulation task?

To address these challenges in learning from human actions, we propose Visual Robot Manipulation with Analogical Reasoning (AR-VRM). Human hand keypoints and robotic actions share underlying similarities, especially when it comes to manipulating objects. Specifically, to extract action knowledge explicitly from a large-scale human video dataset, we propose a keypoint Visual-Language Model (VLM) pretraining scheme. We use a large-scale egocentric human action video dataset Ego4D[7] with common operation videos by human hands, which have similar manipulation tasks and environment views to robotic applications. We detect human hand keypoints in instruction videos, and pre-train a VLM to predict the keypoints of future actions based on the current visual inputs and language instructions. During the fine-tuning stage, to provide demonstrations under manipulation tasks, we first retrieve relevant human action videos that have similar history observations to the robot situation. Moreover, to bridge the gap between robotic arm and human, we propose an Analogical Reasoning (AR) map between robotic arm components and human hand keypoints to learn geometric and functional correlation. Learning this correspondence helps to guide the robot to imitate human actions explicitly, e.g., the process of approaching the object and the grasping operation, to achieve manipulation tasks.

We perform extensive experiments on the challenging CALVIN benchmark. Our AR-VRM achieves state-of-the-art performance under every experimental setting. Specifically, under 10% robot training data as well as unseen scenarios, our method outperforms previous approaches by large margins, improving success rate from 40.0% to 45.6% under

10% training data, and from 61.2% to 65.9% under unseen scenes. These results demonstrate that our idea of explicitly imitating human actions by keypoints has stronger generalization ability under limited robot data and under new scenarios. In addition, our model also outperforms previous SOTA methods on real robot experiments.

We conclude the contributions of our paper as follows: (1) We propose AR-VRM, the first visual robot manipulation approach that explicitly imitates human actions in terms of hand keypoints from large-scale human action video datasets. (2) We propose the keypoint vision language model pre-training scheme to extract action knowledge from human videos, and the analogical reasoning module to align robotic arms with human hand keypoints. (3) Our method achieves state-of-the-art performance on CALVIN benchmark and real-world robot experiments, especially under limited robot training data, demonstrating the effectiveness and generalization ability.

## 2. Related Work

### 2.1. Visual Robot Manipulation

The visual robot manipulation task aims to train the robot to follow natural language instructions. It is a flexible and intuitive way for non-expert humans to instruct the robot to execute tasks such as object grasping, placement, and assembly [3, 5, 11]. Some pioneer approaches delved into diverse visual encoder pre-training methods, aiming to learn useful visual representations by masked image modeling [23, 24] and contrastive learning [10, 20], and reward signals for reinforcement learning [4]. In recent years, some studies introduced large language models as the action planner [6, 28]. ATM[30] proposes to imitate human actions for robot manipulation tasks, but requires paired human-robot action data under specific operation scenarios, limiting its generalization potential. These methods all have limited performance due to the insufficient multimodal data, which is difficult to obtain and costly to collect. In this paper, we employ large-scale human action video data for pretraining to compensate for the insufficiency of robot data.

### 2.2. Vision Language Pretraining for Robotics

Vision-Language Models (VLMs) are pretrained on large-scale datasets with texts and images or videos, thus showing strong generalization ability on multiple downstream tasks, like video understanding[29, 34, 35], interaction grounding [12, 13] and generation[14, 32], and 3D scene understanding[18, 33]. In the robotics field, some approaches introduce large-scale vision language pretraining to enhance robot intelligence. [2] employs Internet-scale web data with images and texts to pretrain the model for basic visual-language understanding, and fine-tune the model on robot data. But the web data, for example, vision question

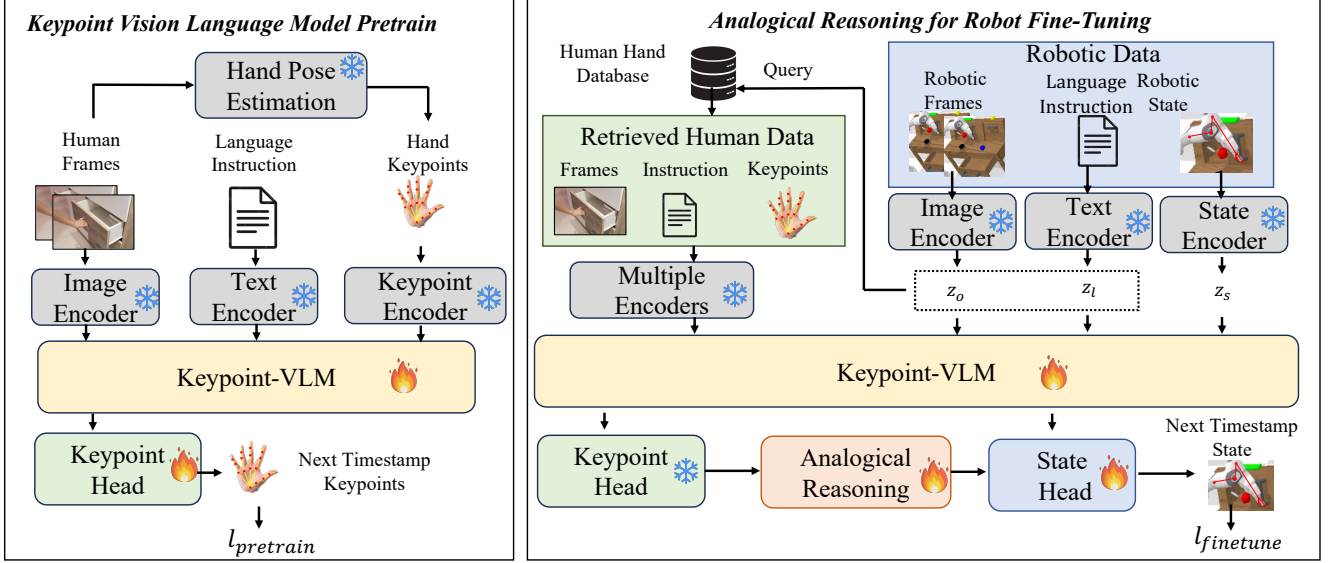


Figure 2. AR-VRM: Visual Robot Manipulation with Analogical Reasoning.

answering data, differs from robot manipulation tasks, leading to limited performance gains. [20, 31] utilize human action video dataset, i.e., Ego4D[7] which is closely related to manipulation task. However, they train the robotic model in an implicit way, either by representation contrastive learning or by predicting pixel-level future frames, introducing irrelevant background information, and are unable to focus on the action itself. In this paper, we propose to explicitly learn from human actions in terms of hand keypoints, enhancing the generalization ability under insufficient robot data.

### 3. Method

#### 3.1. Problem Formulation and Method Overview

We formulate the Visual Robot Manipulation (VRM) task as follows: At timestep  $t$ , a robot model  $R$  maps a language instruction  $l$  and a sequence of history visual observations  $o_{1:t}$  and robot states  $s_{1:t}$  from the starting timestep to current timestep  $t$  to a robot action  $a_t$ :

$$R(l, o_{1:t}, s_{1:t}) \rightarrow a_t \quad (1)$$

where  $o_t \in \mathbb{R}^{H \times W \times 3}$  is the input image and  $s_t = \{s_t^e, s_t^g\}$  is the corresponding robot state with the 6D pose of the robot end-effector  $s_t^e \in \mathbb{R}^6$  and a binary state of the gripper  $s_t^g \in \{0, 1\}$  at timestep  $t, t \in [1, T]$ . Action  $a_t = \Delta(s_t) = s_{t+1} - s_t$  is the variation of state parameters. The complete robot dataset  $D_R = \{\tau_i^R\}_{i=1}^{|D_R|}$  contains paired language instructions, visual inputs, and robot states:

$$\tau^R = \{l, o_1, s_1, o_2, s_2, \dots, o_T, s_T\} \quad (2)$$

These paired data with ground truth robot actions are difficult to obtain and costly to collect, thus have limited scale.

In this paper, we propose to employ vision language pre-training using large-scale human action videos and fine-tune the model on robot data, learning from explicit human action knowledge in terms of hand keypoints. As overviewed in Figure 2, our framework AR-VRM contains a keypoint Vision-Language Model (VLM) pretraining scheme, and an Analogical Reasoning (AR) module during fine-tuning.

In the pretraining stage, we first extract human hand keypoints in the large-scale human action videos, formulating sequences of human action data, and then introduce a keypoint head for the VLM to directly predict future human body keypoints, acquiring human action knowledge and focusing on the explicit action keypoints. In the fine-tuning stage, given the limited robot data, we first retrieve human actions that have a similar manipulation task as well as history observations from the human instructional video database, letting the pretrained keypoint VLM predict future actions, and introduce analogical reasoning to map the correspondence between the human hand keypoints and robot components. In this way, the model learns from explicit human actions and can generalize well under insufficient robot data.

#### 3.2. Keypoint Vision-Language Model Pretraining

We employ a large-scale human egocentric video dataset Ego4D[7]. The video samples in Ego4D are human operations of massive-scale manipulation tasks, with similar environments and views to robotic applications. In the human action video dataset  $D_H = \{\tau_i^H\}_{i=1}^{|D_H|}$ , a video of human

hands accomplishing a task is provided:

$$\tau^H = \{l, o_1, o_2, \dots, o_T\} \quad (3)$$

where  $l$  is the language description of the task, and  $o_t$  is the captured image at timestep  $t$ . Unlike previous methods that learn implicitly from visual information, which inevitably contains irrelevant background information, we propose to explicitly learn from the human action itself with hand keypoints. We employ an offline hand pose estimation model InterHand[19] to extract 3D hand keypoints in each video frame  $k_t = \text{InterHand}(o_t)$ ,  $k_t \in \mathbb{R}^K$ ,  $K$  is the number of keypoints in human hands. The preprocessed human action video dataset  $D_{H^*} = \{\tau_i^{H^*}\}_{i=1}^{|D_{H^*}|}$  is:

$$\tau^{H^*} = \{l, o_1, k_1, o_2, k_2, \dots, o_T, k_T\} \quad (4)$$

The keypoints extracted by [19] are under image coordinate system with 3D coordinates, which could be imitated by robotic arms, for example, mimicking the process of approaching an object and performing operations.

Given the data from three different modalities, we pre-train a keypoint Vision-Language Model (VLM), enabling the VLM to understand and directly predict hand keypoints in a human action sequence. For language instruction  $l$ , following [26, 27, 31], we use CLIP[22] text encoder to extract language embeddings, followed by a multi-layer perceptron (MLP) to project the embedding to dimension  $d$ :

$$z_l = \text{MLP}(\text{CLIP}_{\text{text}}(l)), z_l \in \mathbb{R}^d \quad (5)$$

For each visual input  $o$ , we employ a vision transformer (ViT) pretrained with MAE[8] as the image encoder. We take the output CLS token  $z^{\text{CLS}}$  as the global representation and output patch tokens  $z^{p1:v}$  as local representations, resampled by a preceiver resampler (PR) [9] to reduce token numbers to  $M$ . All the outputs are projected by MLP to dimension  $d$ :

$$z^{\text{CLS}}, z^{p1:v} = \text{ViT}(o_t) \quad (6)$$

$$z_{o_t}^{\text{CLS}} = \text{MLP}(z^{\text{CLS}}), z_{o_t}^{\text{CLS}} \in \mathbb{R}^d \quad (7)$$

$$z_{o_t}^{p1:M} = \text{MLP}(\text{PR}(z^{p1:v})), z_{o_t}^{p1:M} \in \mathbb{R}^d \quad (8)$$

For hand keypoints  $k$ , we employ HandFormer[25] as the keypoint encoder, and also project the embedding to dimension  $d$ :

$$z_{k_t} = \text{MLP}(\text{HandFormer}(k_t)), z_{k_t} \in \mathbb{R}^d \quad (9)$$

With the aligned dimension of tokens from the three modalities  $z_l, z_o = (z_{o_t}^{\text{CLS}}, z_{o_t}^{p1:M}), z_k$ , we concatenate them into a sequence of tokens and feed them to Transformer layers and perform next token prediction for pretraining. Specifically, at timestep  $t$ , the model with self-attention layers  $h^{\text{Atten}}$  predict the keypoint token  $\hat{z}_k$  based on the previous token sequences, and a keypoint prediction head (MLP)

project the keypoint token back to keypoint vectors, and finally compute the mean square error (MSE) with the ground truth keypoint. The pretraining loss can be formulated as:

$$\hat{z}_{k_t} = h^{\text{Atten}}(z_l, z_{o_1}, z_{k_1}, z_{o_2}, z_{k_2}, \dots, z_{o_{t-1}}, z_{k_{t-1}}) \quad (10)$$

$$\mathcal{L}_{\text{pretrain}} = \sum_{t=1}^T \mathcal{L}_{\text{MSE}}(\text{KeypointHead}(\hat{z}_{k_t}), k_t) \quad (11)$$

During training we fix the parameters of the CLIP text encoder and pretrained image encoder. With the keypoint VLM pretraining, we acquire human action knowledge from a large-scale human action video dataset and enable the VLM to explicitly predict action keypoints.

### 3.3. Robot Fine-tuning with Analogical Reasoning

With the keypoint VLM that understands action sequences by large-scale human data pretraining, we could then fine-tune the model on the robot data to accomplish the visual robot manipulation task. In the robot dataset  $D_R$ , we extract the language and vision embeddings and produce tokens  $z_l, z_o$  same as the pretraining stage. For robot state  $s = \{s^e, s^g\}$ , we employ two MLPs  $E^e, E^g$  to encode  $s^e$  and  $s^g$  respectively, and then project the encoded vectors into dimension  $d$  by another MLP:

$$z_s = \text{MLP}(E^e(s^e), E^g(s^g)), z_s \in \mathbb{R}^d \quad (12)$$

With the aligned dimension of tokens  $z_l, z_o, z_s$ , we feed them to the pretrained VLM to predict the state token  $\hat{z}_s$ . Introducing the robot state head (MLP), we project the state token back to robot states and compute the MSE loss with the ground truth states:

$$z_{s_T} = h^{\text{Atten}}(z_l, z_{o_1}, z_{s_1}, z_{o_2}, z_{s_2}, \dots, z_{o_{T-1}}, z_{s_{T-1}}) \quad (13)$$

$$\mathcal{L}_{\text{state}} = \mathcal{L}_{\text{MSE}}(\text{StateHead}(z_{s_T}), s_T) \quad (14)$$

In addition to utilizing the knowledge from pretrained weights of the VLM, we propose to explicitly learn from the human actions in terms of hand keypoints. Specifically, given the robot data sample  $\tau_R$ , we first retrieve the human action videos from the large-scale database based on the language introduction and the visual frame features, with the similarity:

$$\text{sim}(\tau^R, \tau^H) = \cos(z_l^R, z_l^H) + \sum_t \cos(z_{o_t}^R, z_{o_t}^H) \quad (15)$$

where  $\cos(\cdot, \cdot)$  is the cosine similarity. In this way, we could retrieve human action videos that have both similar manipulation tasks and similar visual observations with the robotic data sample. We retrieve the top  $J$  similar samples  $\{\tau_j^H\}_{j=1}^J$ .

After the forward pass of the human action samples  $\tau_j^H$  and robot data sample  $\tau^R$ , the features of the last layer of keypoint head and robot state head  $f_{k_j} \in \mathbb{R}^{K \times d}$ ,  $f_s \in \mathbb{R}^{S \times d}$

serve as the representation of each keypoint node and robot state node, where  $K$  is the number of keypoints of human hands and  $S$  is the number of robot arms components. We introduce a learnable analogical map matrix  $m \in \mathbb{R}^{S \times K}$  to represent the mapping between the human hand keypoints and robot arm components, where each element represents the influence of the corresponding hand keypoint on the robot arm component. We compute the imitated robot state features  $f_{s_j}^*$  as follow:

$$f_{s_j}^* = (1 - \alpha) \cdot m \cdot f_{k_j} + \alpha \cdot f_s \quad (16)$$

where  $\alpha \in \mathbb{R}$  is also a learnable parameter to weight the overall influence of keypoint features. We then employ a new linear layer for new robot states, and compute MSE as the analogical reasoning loss:

$$\mathcal{L}_{AR} = \sum_{j=1}^J \mathcal{L}_{MSE}(\text{Linear}(f_{s_j, T}^*), s_T) \quad (17)$$

The overall fine-tuning loss is the weighted sum of the robot state loss and the analogical reasoning loss:

$$\mathcal{L}_{finetune} = \mathcal{L}_{state} + \beta \cdot \mathcal{L}_{AR} \quad (18)$$

where  $\beta$  is the hyper-parameter.

**Discussion on the fine-tuning method design:** Note that during fine-tuning, we fix the parameters of the keypoint encoder and keypoint head, and fine-tune the Transformer layers of the VLM. Fixing the parameters of the keypoint encoder and keypoint head helps maintain the ability to encode and predict keypoint parameters that have already been accomplished in the pretraining stage. By fine-tuning the Transformers of the VLM with human action video samples, we not only allow the keypoint features to guide the training of the robot state head, but also act as a data replay operation to prevent the VLM from over-fitting to the limited robotic data and forgetting the vision-language understanding and action prediction knowledge gained from pretraining. Detailed experimental discussions will be in Section 4.

## 4. Experiment

### 4.1. Dataset and Benchmark

We conduct experiment on the challenging *long-horizon tasks benchmark* CALVIN [17]. CALVIN is a simulated language-conditioned robot manipulation benchmark that combines natural language conditioning, multi-modal high-dimensional inputs, 7-DOF continuous control, and long-horizon robotic object manipulation in both seen and unseen environments. It contains 34 subtasks and 5 horizon evaluation sequences, providing a range of sensors commonly utilized for visuomotor control, and allows testing zero-shot generalization by leveraging a range of 4 manipulation environments and unseen language instructions.

Robot manipulation dataset requires costly paired multi-modal data, and thus has limited scale. We employ a large-scale human action video dataset Ego4D [7] for pretraining. Ego4D contains massive-scale human-object interaction videos of 3,500 hours, each video containing a natural language annotation describing the human action. Following [31], we use a total of 800,000 video clips containing 8M frames for pretraining and as the human action video database for retrieval during fine-tuning. For real-world experiments, we use a plate with three objects: an orange, an apple, and a green jujube. We collected 1,200 moving-object demonstrations and 1,400 drawer-opening/closing trajectories.

### 4.2. Implementation Details

For network architecture, we use a pretrained CLIP[22] text encoder and a pretrained ViT-Base image encoder [8] to extract language instruction tokens and visual input tokens. For local image patch tokens, we use [9] to reduce the token numbers. For human keypoints, we use the pretrained HandFormer[25] as the encoder. For robot states, we follow [31] to encode the robot arm and grasper parameters by MLPs. We use a 12-layer transformer with causal attention mechanism and checkpoints loaded from [21]. During the pretraining stage, we sample uniformly spaced frames with 3 fps from video clips in the Ego4D dataset. And we set the learning rate to  $1e - 4$  with AdamW, training the model with a batch size of 512 on NVIDIA A800 GPUs with 100 epochs. During the fine-tuning phase, freezing multiple encoders that process human data and the keypoint prediction head serves to stabilize the model’s existing performance while focusing updates on robotic manipulation prediction. And we adjust the learning rate to  $1e - 5$  and execute 50 epochs.

### 4.3. Results Comparing with Other Methods

**Experiment setup:** We evaluate our method on the four different settings of the CALVIN benchmark, including full dataset multitask learning, unseen scene generalization, data-efficient few-shot learning, and unseen language generalization. The training data includes “A”, “B”, “C” and “D” scenes with 34 specific instruction languages. We use the success rate of sequentially completing 1, 2, 3, 4, and 5 tasks, and the average length of successfully completed tasks as the evaluation metrics following previous methods.

**Full dataset multitask learning:** As shown in Table 1 “Experiment ABCD  $\rightarrow$  D” following [31], we train the model on under 4 different scenes and evaluate it on a certain scene. Our method outperforms all the baseline methods, raising the average success rate improvement by +1.2%, demonstrating the effectiveness of our design of introducing explicit human hand keypoints for robot manipulation.

**Unseen scene generalization:** As shown in Table 1 “Experiment ABC  $\rightarrow$  D”, we train the model on under 3 scenes

Method	Experiment	Success rate of tasks completed in a row					Avg.Len.	Avg. Rate
		1	2	3	4	5		
MCIL[15]	ABCD→D	0.373	0.027	0.002	0.000	0.000	0.40	8.0%
RT-1[1]		0.844	0.617	0.438	0.323	0.227	2.45	49.0%
HULC[16]		0.889	0.733	0.587	0.475	0.383	3.07	61.3%
MT-R3M[25]		0.752	0.527	0.375	0.258	0.163	2.07	41.5%
GR-1[31]		0.949	0.896	0.844	0.789	0.731	4.21	84.2%
Ours		<b>0.951</b>	<b>0.915</b>	<b>0.855</b>	<b>0.800</b>	<b>0.751</b>	<b>4.27</b>	<b>85.4% (+1.2%)</b>
MCIL[15]	ABC→D	0.304	0.013	0.002	0.000	0.000	0.31	6.4%
RT-1[15]		0.533	0.222	0.094	0.038	0.013	0.90	18.0%
HULC[16]		0.418	0.165	0.057	0.019	0.011	0.67	13.4%
MT-R3M[25]		0.529	0.234	0.105	0.043	0.018	0.93	18.6%
GR-1[31]		0.854	0.712	0.596	0.497	0.401	3.06	61.2%
Ours		<b>0.901</b>	<b>0.759</b>	<b>0.642</b>	<b>0.531</b>	<b>0.461</b>	<b>3.29</b>	<b>65.9% (+4.7%)</b>

Table 1. Performance comparisons on CALVIN. Bold represents the best results and underline represents the second-best.

(“A”, “B”, and “C”) and evaluate it on scene “D”. Thanks to the replaying design of human videos during fine-tuning, our method further improves the performance by a large margin (Avg. Rate from 61.2% to 65.9%) compared with previous SOTA, indicating that our keypoint VLM pretraining and analogical reasoning have strong generalization ability across different scenes.

Approach	Transportation			Articulated Manipulation
	Seen Obj.	Unseen Ins.	Unseen Cate.	
RT-1[1]	0.27	0.13	0.00	0.35
MT-R3M [25]	0.15	0.13	0.10	0.30
GR-1[31]	0.79	0.73	0.30	0.75
AR-VRM(Ours)	<b>0.95</b>	<b>0.91</b>	<b>0.53</b>	<b>0.82</b>

Table 2. Real robot experiment success rates

**Real-robot experiments.** (1)Object Transportation: *Seen Objects.* The robot transported the three trained objects (orange, apple, jujube) in two disturbed scenes: with distractors (tomato, corn, yellow peach), and with an altered background (wooden board, bowl). *Unseen Instances.* Evaluated generalization to new instances of the same object categories (different orange, apple, jujube). *Unseen Categories.* Tested generalization to untrained categories (tomato, yellow peach). Quantitative results (Table 2) show AR-VRM outperformed baselines (RT-1, MT-R3M, GR-1). Baselines often failed due to incorrect object selection/placement or collisions (e.g., with plate/desk). AR-VRM maintained high success rates for seen objects and showed minimal performance drop for unseen instances, highlighting strong in-category generalization. (2)Articulated Manipulation: It surpassed baselines significantly but exhibited two failure modes: in-

complete drawer closure during closing tasks and missed handle engagement during opening tasks. Despite these, AR-VRM demonstrated superior robustness in articulated manipulation.

#### 4.4. Ablation Studies and Qualitative Analysis

**Effectiveness of our proposed modules:** In Table 3, we show the results of ablating each of our proposed modules. Line 1 represents training the VLM directly by robot data without any pretraining and guidance from human action videos. Due to the lack of large-scale data, it has limited performance. In line 2, with the help of human action video pretraining with hand keypoints, the VLM gains vision-language understanding knowledge, and the performance shows a significant gain. Line 3 stands for the introduction of human action videos during fine-tuning by retrieval, but only training the separate keypoint head instead of the analogical mapping to guide the robot state prediction. This method helps to maintain the knowledge learned by pretraining and prevents the over-fitting to small-scale robot data, thus showing performance gain compared with line 3, but still has room for improvement. Line 4 stands for our complete design of keypoint VLM pretraining with large-scale data and fine-tuning with retrieved human action videos and analogical reasoning. It achieves the best result, demonstrating the effectiveness of each of our proposed modules.

**Analysis of design choices of robot fine-tuning with analogical reasoning:** As noted in Section 3, we freeze the keypoint encoder/head and fine-tune the VLM during analogical reasoning adaptation. Results in Table 4 show: Freezing the VLM (lines 1 vs 4, 2 vs 3) causes significant performance drops, confirming that VLM fine-tuning is critical for transferring human knowledge to robots. Freezing keypoints (lines 1 vs 2, 3 vs 4) boosts performance, likely

Pretrain	Retrieval	AR	Success rate of tasks completed in a row					Avg.Len.	Avg. Rate
			1	2	3	4	5		
✗	✗	✗	0.824	0.701	0.595	0.472	0.410	3.00	60.0%
✓	✗	✗	0.892	0.871	0.810	0.781	0.710	4.06	81.3%
✓	✓	✗	0.939	0.897	0.843	0.797	0.738	4.21	84.3%
✓	✓	✓	<b>0.951</b>	<b>0.915</b>	<b>0.855</b>	<b>0.800</b>	<b>0.751</b>	<b>4.27</b>	<b>85.4%</b>

Table 3. **Ablation studies on the effectiveness of our proposed modules.** “Pretrain”, “Retrieval” and “AR” denotes keypoint VLM pretraining with large-scale human data, retrieving human action videos for fine-tuning and analogical reasoning, respectively.

Keypoint params	VLM params	Success rate of tasks completed in a row					Avg.Len.	Avg. Rate
		1	2	3	4	5		
frozen	frozen	0.942	0.845	0.803	0.741	0.598	3.93	78.6%
trainable	frozen	0.931	0.832	0.788	0.726	0.576	3.85	77.1%
trainable	trainable	0.939	0.847	0.821	0.772	0.730	4.11	82.2%
frozen	trainable	<b>0.951</b>	<b>0.915</b>	<b>0.855</b>	<b>0.800</b>	<b>0.751</b>	<b>4.27</b>	<b>85.4%</b>

Table 4. **Ablation study on the design choices of robot fine-tuning with analogical reasoning.**

Method	Data	Success rate of tasks completed in a row					Avg.Len.	Avg. Rate
		1	2	3	4	5		
GR-1	100%	0.949	0.896	0.844	0.789	0.731	4.21	84.2%
Ours		<b>0.951</b>	<b>0.915</b>	<b>0.855</b>	<b>0.800</b>	<b>0.751</b>	<b>4.27</b>	<b>85.4%</b>
GR-1	10%	0.778	0.533	0.332	0.218	0.139	2.00	40.0%
Ours		<b>0.809</b>	<b>0.589</b>	<b>0.381</b>	<b>0.287</b>	<b>0.213</b>	<b>2.28</b>	<b>45.6%</b>

Table 5. **Data efficient few-shot learning with only 10% train data on ABCD→D.**

Method	Lang	Success rate of tasks completed in a row					Avg.Len.	Avg. Rate
		1	2	3	4	5		
GR-1	Seen	0.949	0.896	0.844	0.789	0.731	4.21	84.2%
Ours		<b>0.951</b>	<b>0.915</b>	<b>0.855</b>	<b>0.800</b>	<b>0.751</b>	<b>4.27</b>	<b>85.4%</b>
GR-1	Unseen	0.764	0.555	0.381	0.270	0.196	2.17	43.3%
Ours		<b>0.802</b>	<b>0.575</b>	<b>0.430</b>	<b>0.313</b>	<b>0.209</b>	<b>2.33</b>	<b>46.6%</b>

Table 6. **Unseen instruction language generalization on ABCD→D.**

because stabilized keypoint features aid convergence of the analogical map, improving human-robot component alignment.

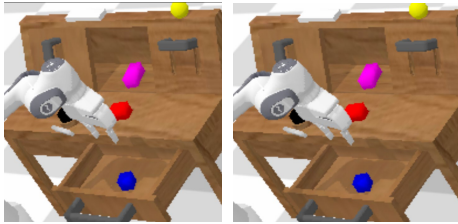
**Date efficient few-shot learning:** As shown in Table 5, we train the model with only 10% of the already small dataset to train our model. Our method outperforms all the baseline methods, raising the success rate of completing 5 tasks to 45.6% and the average length of completed tasks to 2.28. The results show that our analogical reasoning can provide a map from the diversity and complexity of actions in human keypoints to insufficient robot action, which can enable the robot to quickly learn how to cope with different tasks.

**Unseen instruction language generalization:** As shown in Table 6, we evaluate the model on the unseen instruction languages during training. Our method consistently achieves state-of-the-art, indicating that the relevant human action demonstration significantly help the robot generalize to multiple new instructional tasks.

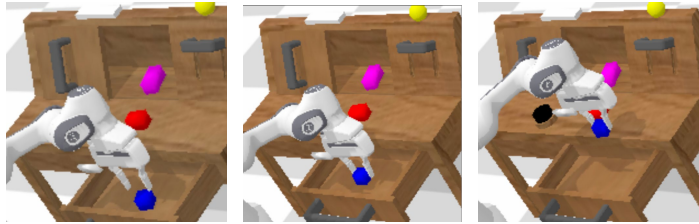
**Visualization of our prediction results:** In Figure 3, we provide an example of robot manipulation task of “grasp the blue block in drawer”. Our method retrieves relevant human action videos such as “pick up a tool/receipt/cloth/knife” from the database. By imitating the human action of picking up objects from the drawer, which contains process actions

Language Instruction: grasp the blue block in the drawer

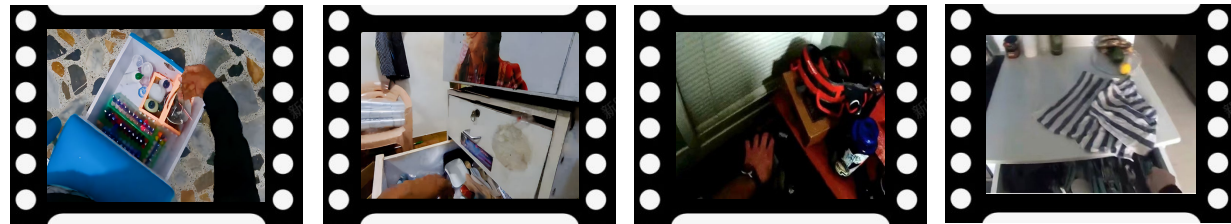
Visual Observations:



Successful action:



Retrieved human action videos:



opens a drawer and picks up a tool

picks a receipt from the drawer

picks cloth from the drawer

picks a knife in the drawer

Figure 3. Example of robot manipulation with our retrieved human action videos and action prediction result.

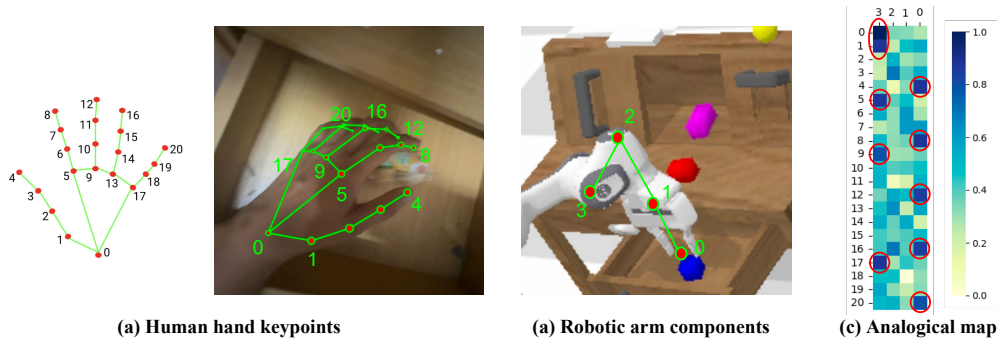


Figure 4. **Visualization of analogical mapping.** (a) Human hand keypoints code 0-20. (b) Robotic visible components code 0-3. (c) The learned analogical map. The highlighted map elements by red circles indicate a strong relationship between the two nodes.

of approaching the drawer and grasping the object, the robot could successfully follow the language instruction to pick up the blue block in the drawer.

**Visualization of analogical map:** In Figure 4, we show an analogical map partial example from hand keypoints to a robotic arm, and the map is normalized in a column-wise manner and shows the correlation between the estimated hand keypoints and visible robotic arm components. The coding of keypoint nodes and robotic arm component nodes is illustrated in (a) and (b), and in (c), the map elements that are highlighted by red circles indicate a strong relationship between the two nodes. The grasper of the robot (code 0) is strongly related to the fingertips of human hands (code 4, 8, 12, 20) which operates the grasping in manipulation tasks, and the root segment of the robotic arm (code 3) relates to the human palm (code 0, 1, 5, 9, 17) deciding the

direction of the movement. This visualization indicates that our method learns a reasonable relationship by mapping the human action to guide the robotic arm.

## 5. Conclusion

To explore key information in human interaction processes from videos to guide robot manipulation, we introduce a Visual Robot Manipulation with Analogical Reasoning (AR-VRM), leveraging large-scale human action video datasets to explicitly learn action knowledge through hand keypoints. We use a keypoint pretraining scheme to enhance generalization and performance. With Analogical Reasoning to map human hand keypoints to robot components, our method achieves SOTA results on the CALVIN and real robot experiments, especially in few-shot data scenarios, demonstrating our effectiveness and generalization.

**Acknowledgements.** This work was supported by the grants from the National Natural Science Foundation of China 62372014, Beijing Nova Program and Beijing Natural Science Foundation 4252040.

## References

- [1] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 1, 6
- [2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 1, 2
- [3] Yang Cong, Ronghan Chen, Bingtao Ma, Hongsen Liu, Dongdong Hou, and Chenguang Yang. A comprehensive study of 3-d vision-based robot manipulation. *IEEE Transactions on Cybernetics*, 53(3):1682–1698, 2021. 1, 2
- [4] Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Danijar Hafner, and Pieter Abbeel. Video prediction models as rewards for reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [5] Bin Fang, Shidong Jia, Di Guo, Muhua Xu, Shuhuan Wen, and Fuchun Sun. Survey of imitation learning for robotic manipulation. *International Journal of Intelligent Robotics and Applications*, 3:362–369, 2019. 1, 2
- [6] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12462–12469. IEEE, 2024. 2
- [7] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2, 3, 5
- [8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 4, 5
- [9] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 4, 5
- [10] Ya Jing, Xuelin Zhu, Xingbin Liu, Qie Sima, Taozheng Yang, Yunhai Feng, and Tao Kong. Exploring visual pre-training for robot manipulation: Datasets, models and methods. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11390–11395. IEEE, 2023. 2
- [11] Oliver Kroemer, Scott Niekum, and George Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *Journal of machine learning research*, 22(30):1–82, 2021. 1, 2
- [12] Ting Lei, Shaofeng Yin, and Yang Liu. Exploring the potential of large foundation models for open-vocabulary hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16657–16667, 2024. 2
- [13] Ting Lei, Shaofeng Yin, Yuxin Peng, and Yang Liu. Exploring conditional multi-modal prompts for zero-shot hoi detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [14] Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1769–1782, 2025. 2
- [15] Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data. *Robotics: Science and Systems*, 2021. 6
- [16] Oier Mees, Lukas Hermann, and Wolfram Burgard. What matters in language conditioned robotic imitation learning over unstructured data. *IEEE Robotics and Automation Letters (RA-L)*, 7(4):11205–11212, 2022. 6
- [17] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022. 5
- [18] Wentao Mo and Yang Liu. Bridging the gap between 2d and 3d visual question answering: A fusion approach for 3d vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 2
- [19] Gyeongsik Moon, Shou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020. 4
- [20] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022. 2, 3
- [21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 5
- [22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 5
- [23] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pages 416–426. PMLR, 2023. 1, 2
- [24] Younggyo Seo, Junsu Kim, Stephen James, Kimin Lee, Jinwoo Shin, and Pieter Abbeel. Multi-view masked world models for visual robotic manipulation. In *International Con-*

- ference on Machine Learning*, pages 30613–30632. PMLR, 2023. [2](#)
- [25] Md Salman Shamil, Dibyadip Chatterjee, Fadime Sener, Shugao Ma, and Angela Yao. On the utility of 3d hand poses for action recognition. In *European Conference on Computer Vision*, pages 436–454. Springer, 2025. [4](#), [5](#), [6](#)
- [26] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, pages 894–906. PMLR, 2022. [4](#)
- [27] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023. [4](#)
- [28] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. [2](#)
- [29] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. [2](#)
- [30] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023. [2](#)
- [31] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)
- [32] Zhu Xu, Qingchao Chen, Yuxin Peng, and Yang Liu. Semantic-aware human object interaction image generation. In *Forty-first International Conference on Machine Learning*, 2024. [2](#)
- [33] Dejie Yang, Zhu Xu, Wentao Mo, Qingchao Chen, Siyuan Huang, and Yang Liu. 3d vision and language pretraining with large-scale synthetic data. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1552–1560, 2024. [2](#)
- [34] Dejie Yang, Zijing Zhao, and Yang Liu. Planllm: Video procedure planning with refinable large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9166–9174, 2025. [2](#)
- [35] Minghang Zheng, Sizhe Li, Qingchao Chen, and Yuxin Peng and Yang Liu. Phrase-level temporal relationship mining for temporal sentence localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. [2](#)