GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation

*Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong,
Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu,
Jiafeng Xu, Yichu Yang, Hanbo Zhang, Minzhao Zhu

ByteDance Research

Abstract

We present GR-2, a state-of-the-art generalist robot agent for versatile and generalizable robot manipulation. GR-2 is first pre-trained on a vast number of Internet videos to capture the dynamics of the world. This large-scale pre-training, involving 38 million video clips and over 50 billion tokens, equips GR-2 with the ability to generalize across a wide range of robotic tasks and environments during subsequent policy learning. Following this, GR-2 is fine-tuned for both video generation and action prediction using robot trajectories. It exhibits impressive multi-task learning capabilities, achieving an average success rate of 97.7% across more than 100 tasks. Moreover, GR-2 demonstrates exceptional generalization to new, previously unseen scenarios, including novel backgrounds, environments, objects, and tasks. Notably, GR-2 scales effectively with model size, underscoring its potential for continued growth and application. Project page: `https://gr2-manipulation.github.io`.

1 Introduction

The rise of high-capacity foundation models has contributed significantly to the success of language [1], image [2], and video [3] processing tasks. These models are initially pre-trained on large-scale diverse datasets and can subsequently be adapted to specific downstream tasks, making them versatile in application. This paradigm allows these models to tackle a variety of tasks with a single generalist model when conditioned on different inputs (*e.g.*, language prompts [4]).

Following the foundation models established in other domains, our goal is to develop a foundation generalist manipulation agent via large-scale pre-training on a comprehensive dataset. This would enable rapid adaptation to a wide range of novel manipulation tasks via efficient fine-tuning. A generalist manipulation agent should be capable of executing a wide range of manipulation skills. And more importantly, it should exhibit strong performance in acquiring new skills and handling disturbances. Despite recent advances in AI and a shift towards data-driven learning, collecting large-scale robot data remains a significant challenge due to inefficient data collection methods and the limited scalability of real-robot systems. Research suggests that pre-training on video generation can effectively transfer valuable knowledge from videos to policy learning, thus improving the action prediction capability [5].

This report introduces GR-2, an evolution of our previous model [5], featuring improved performance and expanded capabilities. To achieve this, we pre-train GR-2 on an extensive video dataset encompassing diverse daily human activities across different contexts (household, outdoor, workplace, leisure, etc.). The primary pre-training objective is straightforward: given a textual description and a video frame, the model predicts subsequent frames based on the text. By mastering this auto-regressive prediction task, we anticipate the

*Authors are listed in alphabetical order. Contributions are listed at the end of the report. Corresponding email(s): `{kongtao,wuhongtao.123}@bytedance.com`

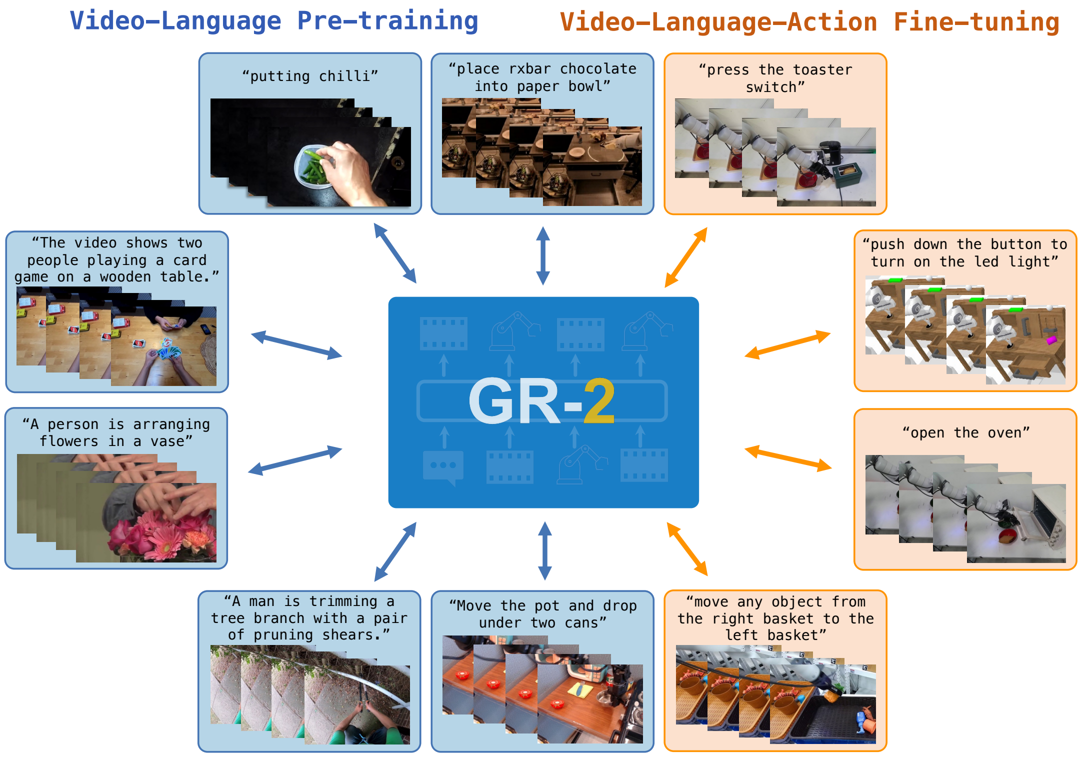

Figure 1: **Overview.** GR-2 undegoes two stages of training: video generation pre-training and robot data fine-tuning.

model to capture crucial temporal dynamics and semantic information which are essential for downstream policy learning. Through fine-tuning on robot trajectories, GR-2 demonstrates the capability to learn multiple manipulation tasks and adapt to novel scenarios, including novel backgrounds, environments, objects, and tasks. Notably, GR-2 efficiently learns over 100 tasks from a dataset with only 5,000 trajectories (an average of 50 trajectories per task). This significantly reduces the cost of acquiring new skills and adapting to new environments in application. Furthermore, GR-2 excels in generalizing to unseen objects in an end-to-end bin-picking setting, highlighting its strong potential for industrial application. Specifically, GR-2 builds upon GR-1 [5] with several key improvements:

- GR-2 is pre-trained on 38 million text-video data (amounting to over 50 billion tokens), and is capable of accomplishing over 100 manipulation tasks and performing bin-picking of over 100 objects. It significantly scales up the pre-training data and number of tasks.
- We develop a novel model architecture that allows the knowledge gathered from pre-training to seamlessly transfer to downstream fine-tuning in a lossless way. The model demonstrates strong scalability in handling multiple tasks in challenging generalization settings.
- For real-robot deployment, we introduce a whole-body control (WBC) algorithm that incorporates trajectory optimization and real-time motion tracking.

The remainder of this report is organized as follows. Sec. 2 provides a detailed description of GR-2, including its model architecture, training process, and real-world deployment. Sec. 3 outlines our experiment setups and results. Sec. 4 discusses the relation of GR-2 to existing works. Finally, Sec. 5 concludes the work and discusses future directions.

2 Methods

We consider language-conditioned visual robot manipulation as our approach towards generalist robot manipulation, as language is one of the most flexible ways for a human to specify tasks for a robot. In this setting, a single robot policy must solve multiple complex manipulation tasks by understanding different

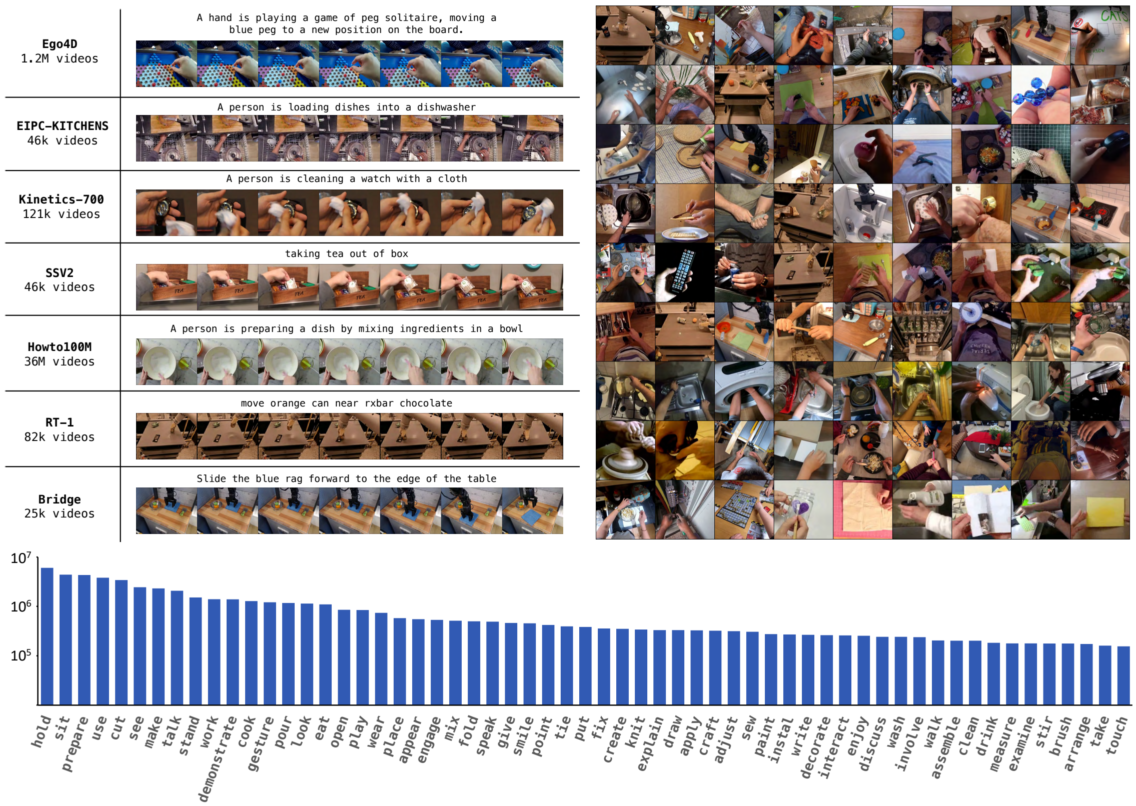

Figure 2: **Pre-training Dataset.** We show sample videos and the verb distribution of the pre-training dataset we curated. The y-axis of the bottom plot is the logarithm frequency of the top words.

unconstrained language instructions. Specifically, we want to train a universal policy π that takes as inputs a language instruction l, a sequence of environment observation $\mathbf{o}_{t-h:t}$, and a sequence of robot states $\mathbf{s}_{t-h:t}$. The policy outputs an action trajectory $\mathbf{a}_{t:t+k}$ in an end-to-end manner:

$$\mathbf{a}_{t:t+k} = \pi(l, \mathbf{o}_{t-h:t}, \mathbf{s}_{t-h:t}), \tag{1}$$

where h and k denote the length of the observation history and the action trajectory, respectively.

2.1 Model & Training

GR-2 is a language-conditioned GPT-style visual manipulation policy model (Fig. 1). The training undergoes two stages: video generative pre-training and robot data fine-tuning. During the pre-training stage, we train GR-2 on a curated large-scale video dataset. After that, we fine-tune GR-2 on robot data to predict action trajectories *and* videos in tandem:

$$\pi(l, \mathbf{o}_{t-h:t}, \mathbf{s}_{t-h:t}) \rightarrow \mathbf{o}_{t+1}, \mathbf{a}_{t:t+k} \tag{2}$$

The inputs to GR-2 contain a language instruction, a sequence of video frames, and a sequence of robot states.

We use a frozen text encoder [6] to tokenize the language instruction. For the image frames in the video, we employ a VQGAN [7] to convert each image into discrete tokens. The VQGAN is trained on a large corpus of Internet data as well as in-domain robot data and is kept frozen during the training process. This approach facilitates fast training and supports the generation of high-quality videos. Robot states contain the position and rotation of the end-effector, as well as the binary gripper state. The states are encoded via linear layers, which are trainable during the fine-tuning stage.

Our goal in the pre-training stage is to equip GR-2 with the capability to predict future videos. This enables the model to develop a strong prior for predicting future events, thereby enhancing its ability to make accurate action predictions. The model, built upon a GPT-style transformer, takes the tokenized text and image sequence as inputs and outputs the discrete tokens of future images. Future images are decoded from these tokens

Figure 3: **Multi-Task Learning.** We perform experiments in two basic settings (Simple and Distractor) and three generalization settings (Unseen Backgrounds, Unseen Environments, and Unseen Manipulation).

with the VQGAN decoder. We highlight that GR-2 is pre-trained on a significantly larger volume of video data compared to previous works that utilize video pre-training. The pre-training data includes commonly used public datasets of human activities, *e.g.*, Howto100M [8], Ego4D [9], Something-Something V2 [10], EPIC-KITCHENS [11], and Kinetics-700 [12]. To tailor the pre-training data for robot manipulation tasks, we carefully establish a data processing pipeline that includes hand filtering [13] and re-captioning [14]. In addition, we include publicly available robot datasets, *e.g.*, RT-1 [15] and Bridge [16]. In total, the number of video clips used for pre-training is 38 million, equivalent to approximately 50 billion tokens. The distribution of human activities and video samples are illustrated in Fig. 2.

GR-2 can be seamlessly fine-tuned on robot data after large-scale pre-training. Unlike the videos in pre-training data which only have a single camera view, robot data usually contain multiple views. GR-2 is designed to gracefully handle multiple views. It takes as inputs the tokenized language instruction, the image sequences captured from multiple views, and the robot state sequence. The outputs include future images of each view and an action trajectory. The action trajectory is generated with a conditional VAE (cVAE) [17, 18, 19]. We empirically found that generating action trajectories rather than single-step actions is crucial for both trajectory smoothing and real-time performance.

2.2 Real-Robot System & Deployment

Our real-robot system consists of a 7-DoF Kinova Gen3 robot arm paired with a Robotiq 2F-85 gripper. We utilize two cameras: a static head camera provides an overview of the workspace; another camera, which is mounted on the end-effector, offers a close-up view of interactions between the gripper and the environment.

GR-2 generates an action trajectory in Cartesian space. To ensure that the robot arm accurately follows this trajectory, we develop a Whole-Body Control (WBC) algorithm that employs trajectory optimization for motion tracking [20]. The generated trajectory first undergoes optimization to improve its smoothness and continuity. Subsequently, the WBC algorithm converts the Cartesian trajectory into low-level joint actions, which are executed on the real robot at a frequency of 200 Hz. This process integrates collision constraints and manipulability into the optimization framework.

3 Experiments

We perform large-scale real-robot experiments in two settings: multi-task learning (Fig. 3) and end-to-end bin picking (Fig. 7). In multi-task learning, we aim to evaluate the capability of GR-2 on learning multiple different tasks. We also evaluate in multiple challenging out-of-distribution settings to verify its generalization capabilities (Fig. 3). In end-to-end bin picking, our goal is to evaluate GR-2 in a more industrial setting. In this setting, the model is provided with a single text prompt and is required to perform the bin-picking task within an object cluster. Finally, we present a benchmark comparison with state-of-the-art methods on the challenging CALVIN benchmark [21]. If not specified otherwise, the default GR-2 model contains 230M parameters, of which 95M are trainable. We also show model scaling results in Sec 3.5.

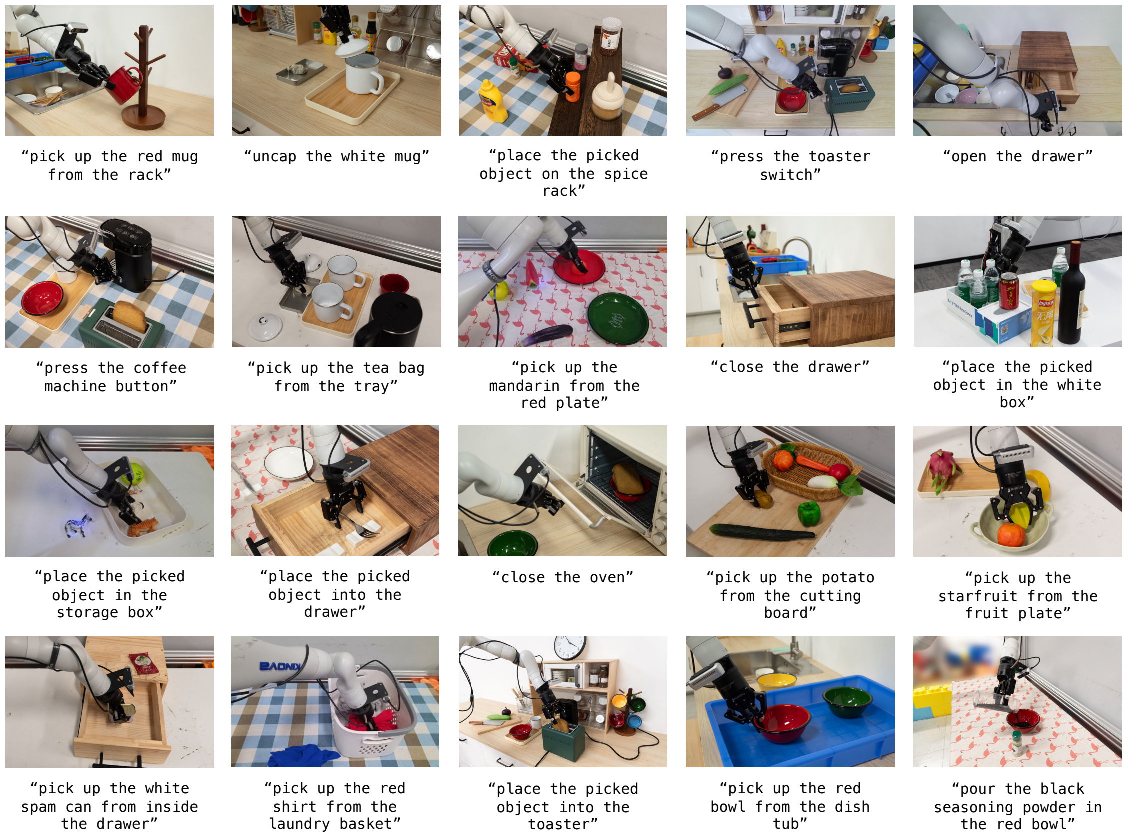

Figure 4: **Task Examples.** GR-2 is able to perform more than 100 tasks of 8 skills including picking, placing, uncapping, capping, opening, closing, pressing, and pouring.

3.1 Real-World Multi-Task Learning

We collected human demonstrations of 105 table-top tasks via teleoperation. These tasks cover 8 different skills, *i.e.*, picking, placing, uncapping, capping, opening, closing, pressing, and pouring (Fig. 4). In total, we collected about 40,000 trajectories, with an average of 400 trajectories per task. Based on the model pre-trained on the curated large-scale video dataset, we further fine-tune GR-2 using this dataset. Additionally, to evaluate the performance under the condition of data scarcity, we train GR-2 using approximately 1/8 of the full dataset, which corresponds to around 50 trajectories per task.

To enable better generalization to unseen scenarios, we perform data augmentation during fine-tuning by adding new objects into the scene and/or changing the background. To insert new objects into the scene, a diffusion model [22] is trained with a combination of a self-collected object dataset and the Open Images dataset [23]. This model enables us to insert a specific object in a designated region. For changing the background, we utilize Segment Anything Model (SAM) [24] to extract regions corresponding to the background. Finally, we employ a video generation model [25] that conditions on the original video and the inpainted frame to produce an augmented video while preserving the robot motion.

Basic Settings. We first evaluate GR-2 in two basic settings: Simple and Distractor. In Simple, the test environment is set similar to the training data. In Distractor, we add a few distractors to the scene. This becomes challenging for the reason that 1) distractors, especially those that share a similar color and/or shape with the target object, may confuse the robot and 2) the environment becomes more cluttered and sometimes requires collision avoidance to accomplish a task.

Generalization Settings. To further investigate the capability of GR-2 in unseen scenarios, we introduce three more challenging settings: Unseen Backgrounds, Unseen Environments, and Unseen Manipulation (Fig. 3). In Unseen Backgrounds, we change the background by adding two unseen tablecloths that are very different from the original plain background in the training dataset as shown in Fig 3. For Unseen Environments, we evaluate in two unseen kitchen environments. Besides changed backgrounds, these environments also contain

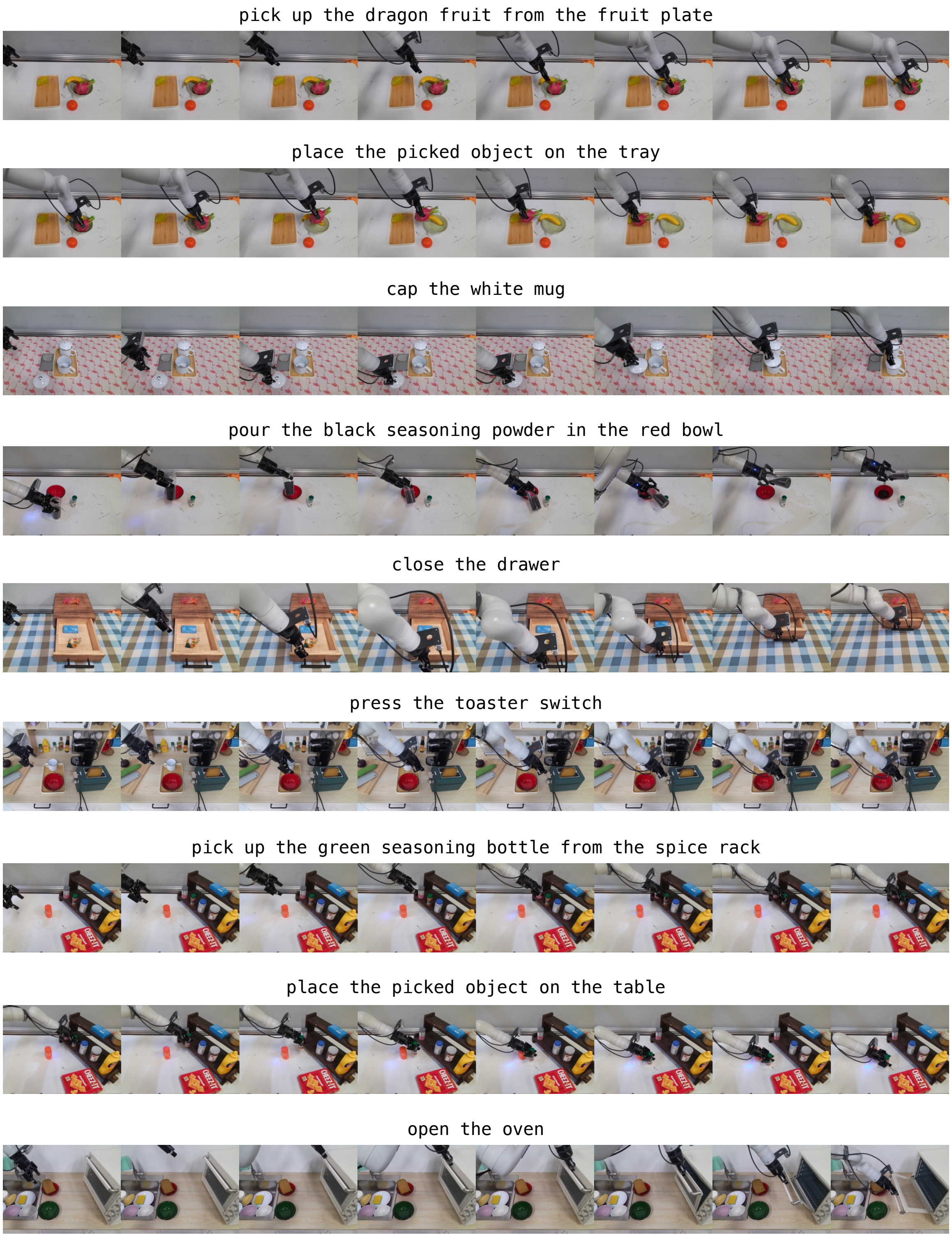

Figure 5: **Qualitative Results of Multi-Task Learning.** We show end-to-end rollouts of different tasks.

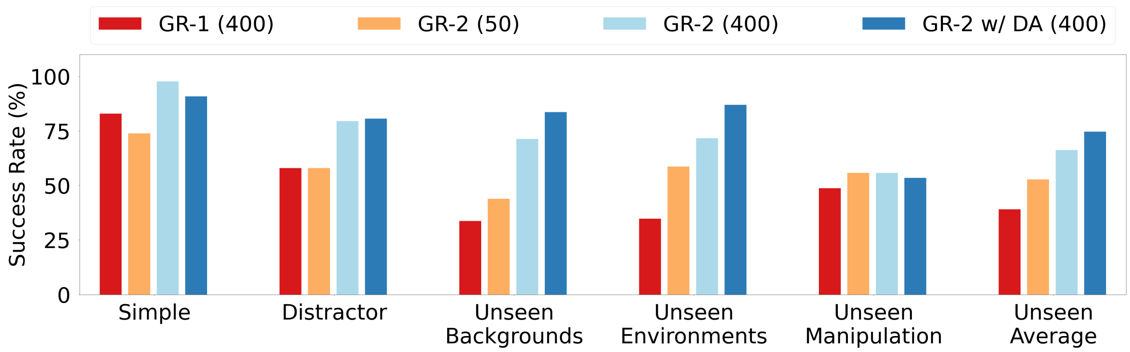

Figure 6: **Success Rates of Multi-Task Learning.** We show the success rates of four models across different evaluation settings. 400 (50) indicates that the model is trained on about 400 (50) trajectories per task on average. GR-2 w/ DA indicates that we perform data augmentation on the training data. See Sec. 3.1 for more details.

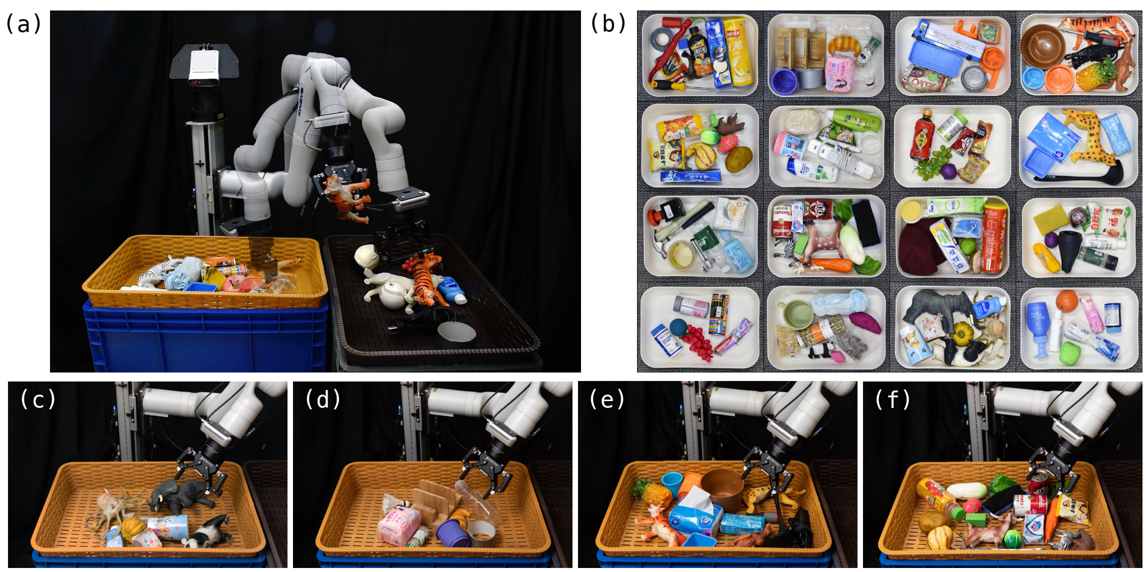

Figure 7: **End-to-End Bin Picking.** (a) Experiment setting. (b) Objects used in the experiments. We evaluate in four different settings: (c) Seen, (d) Unseen, (e) Cluttered Seen, and (f) Cluttered Unseen. Seen (Unseen) indicates that the objects are seen (unseen) during training. The two cluttered settings (e) and (f) have more objects compared to the training setting.

scene-related distractors. Finally, for Unseen Manipulation, we instruct the robot to perform manipulations that are unseen in the robot training data. It includes manipulating objects of unseen categories and unseen object instances. This setting is extremely challenging given the robot has never seen these objects in the training data. And the unseen instructions for manipulating objects of unseen categories further increase the difficulties.

Results. Qualitative results are shown in Fig. 5. Quantitative results are shown in Fig. 6. GR-2 achieves a success rate of 97.7% on 105 tasks in the Simple setting, showcasing its powerful multi-task learning capability. It can also robustly handle distractors and attend to target objects correctly. It improves the success rates of GR-1 in all settings. Notably, it achieves success rates of 71.4% and 71.7% in Unseen Backgrounds and Unseen Environments, respectively, doubling those of GR-1. By introducing data augmentation, GR-2 w/ DA is able to achieve even more competitive generalization performance, obtaining a success rate of 87.0% in Unseen Environments and an average success rate of 74.7% across all three generalization settings. When trained with only 50 trajectories per task, GR-2 is able to achieve a success rate of 73.9% in the Simple setting and outperforms GR-1 in all three generalization settings. This showcases the strong potential of GR-2 in efficiently adapting to new tasks and environments. Finally, GR-2 achieves a success rate of 55.8% in Unseen Manipulation. Typical failure cases include 1) failing to pick unseen objects of novel shapes and 2) mistakenly selecting the wrong object when instructed to pick an unseen one. Moving forward, we plan to explore techniques to further improve generalization for unseen manipulation tasks, including handling novel objects and executing new skills.

3.2 End-to-End Bin Picking of Different Objects

To further assess the capabilities of GR-2 in an industrial context, we conduct large-scale experiments on end-to-end bin picking. The experiment setup contains a source and a target basket (Fig. 7(a)). The robot is tasked with picking objects from the source basket and placing them into the target basket in a seamless and end-to-end manner. In total, we collected about 94,000 pick-and-place trajectories of 55 objects for training. The language instruction is very simple:

```
move any object from the right basket to the left basket.
```

Settings. We evaluate GR-2 in four different settings: Seen, Unseen, Cluttered Seen, and Cluttered Unseen (Fig. 7(c)(d)(e)(f)). In total, we perform experiments on 122 objects, among which 55 of them are seen and the other 67 are unseen during training (Fig. 7(b)). We transport 5-9 seen (unseen) objects from the source basket to the target one in the Seen (Unseen) setting. The number of objects in the source basket at the beginning is similar to those in the training data. For the Cluttered Seen (Unseen) setting, we increase the number of objects by twofold, *i.e.*, including 12-17 seen (unseen) objects in the source basket. And thus the two cluttered settings can be considered as unseen settings regardless of whether the objects are seen or unseen.

Results. Qualitative results are shown in Fig. 8. Quantitative results are shown in Fig. 9. GR-2 outperforms GR-1 by a large margin, improving the average success rate from 33.3% to 79.0%. GR-1 is not able to handle the Unseen and the two cluttered settings. The performance degrades largely from that of the Seen setting. On the other hand, we highlight that the success rates of GR-2 in the Unseen and the two cluttered settings are comparable to that of the Seen setting. These results showcase that GR-2 possesses powerful generalization capabilities for unseen objects and unseen scenarios, indicating great potential for industrial application. GR-2 is able to handle objects that may be challenging for model-based methods, including transparent, deformable, and reflective objects. See Fig. 8 for some examples.

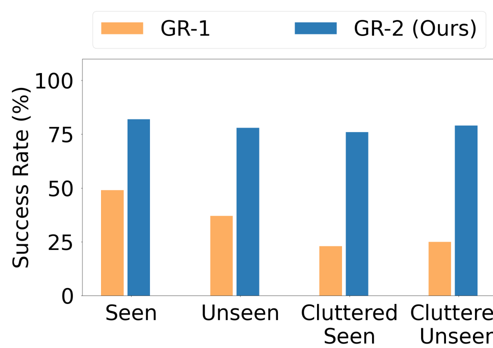

Figure 9: **Success Rates of End-to-End Bin Picking.**

3.3 CALVIN Benchmark

CALVIN is a simulated benchmark which targets long-horizon language-conditioned robot manipulation [21]. It includes 34 tasks and incorporates unconstrained language instructions. We perform experiments on the ABCD-D split which includes more than 20,000 expert demonstrations of 34 different manipulation tasks. Following [21], we perform evaluation on 1,000 unique sequences of instruction chains. For each sequence, GR-2 is instructed to perform 5 tasks in a row.

Fig. 10 shows the success rates of completing 1, 2, 3, 4, and 5 tasks in a row and the average length. The average length is a comprehensive evaluation metric which shows the average number of tasks the robot is able to accomplish in a sequence across the 1,000 evaluated sequences. We compare with five state-of-the-art baseline methods: RT-1 [15], MT-ACT [26], HULC [27], RoboFlamingo [28], and GR-1 [5]. RT-1 [15] is a language-conditioned multi-task policy that encodes the language condition via FiLM layers. MT-ACT [26] similarly uses FiLM layers to inject the language condition and leverages an action-chunking transformer to address the multi-modality in the action data. HULC [27] is a hierarchical method which first predicts a plan in a latent space and uses the predicted plan for generating actions. RoboFlamingo [28] fine-tunes a large pre-trained vision-language model on robotics data to perform language-conditioned manipulation. GR-2 establishes a new state of the art. It outperforms all the comparing baseline methods in terms of success rates and the average length. It improves the success rate of GR-1 from 94.9% to 98.6% for 1 task and from 73.1% to 85.9% for 5 tasks. The average length is increased from 4.21 to 4.64.

Figure 8: **Qualitative Results of End-to-End Bin Picking.** We show end-to-end picking of different objects, including objects that are transparent, deformable, or reflective.

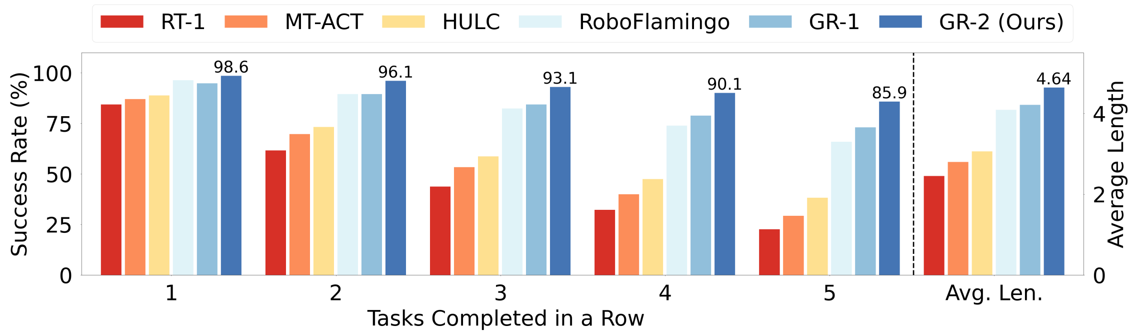

Figure 10: **CALVIN Benchmark Results.** We show the success rates of completing 1, 2, 3, 4, and 5 tasks in a row and the average length. The average length shows the average number of tasks the robot is able to accomplish when instructed to perform 5 tasks in a row.

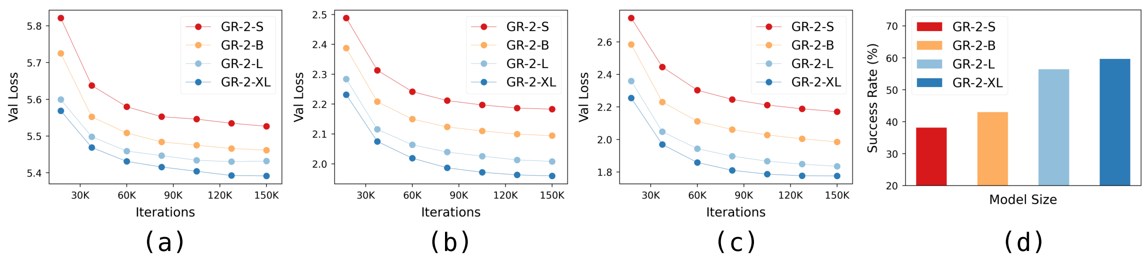

Figure 11: **Scaling Experiments.** We show the validation loss of video generation during pre-training on the validation sets of (a) Ego4D [9], (b) RT-1 [15], and (c) our robot data. (d) shows success rates in real-robot experiments. See Sec. 3.5 for more details.

3.4 Autoregressive Video Generation

GR-2 is pre-trained on a vast number of diverse videos, enabling it to predict future states within the image space. As a result, this video generation capability can effectively act as a planner for action generation. That is, after generating the visual trajectory, an action trajectory can be subsequently inferred based on the visual trajectory. To further investigate the effectiveness of this design, we visualize the video prediction result and compare it with the corresponding real rollout. We show the visualization of rollouts from multi-task learning (Fig. 12 13), end-to-end bin picking (Fig. 14 15), and CALVIN (Fig. 16 17).

GR-2 is able to generate high-quality videos alongside actions. We highlight that the generated videos align with the real-world rollouts faithfully. This indicates that the predicted action is trying to "replay" the trajectory in the predicted video. This property brings about a simple approach to continuously improving action prediction by iteratively improving video generation.

3.5 Scaling

We investigate how scaling up the model size can help GR-2 in pre-training and fine-tuning. In particular, we pre-train GR-2 of four sizes. The number of trainable parameters is 30M (GR-2-S), 95M (GR-2-B), 312M (GR-2-L), and 719M (GR-2-XL), respectively. The validation loss of video prediction is shown in Fig. 11(a)(b)(c). The validation loss decreases with the increase of the model size, showing scalability in terms of video generation. We incorporate videos of in-domain robot data during the pre-training stage and keep the pre-trained parameters frozen while fine-tuning on robot trajectories. After fine-tuning, we evaluate different models on a subset of settings in Sec. 3.1. Results are shown in Fig. 11(d). The success rate scales well with the model size. This result highlights the strong potential of GR-2 for continuous performance improvement through increasing the model size.

Figure 12: **Video Prediction (Pred) and Ground-Truth (GT) Rollouts of Multi-Task Learning (I).** We show autoregressive video predictions alongside the corresponding ground-truth videos captured from real-world rollouts.

Figure 13: **Video Prediction (Pred) and Ground-Truth (GT) Rollouts of Multi-Task Learning (II).** We show autoregressive video predictions alongside the corresponding ground-truth videos captured from real-world rollouts.

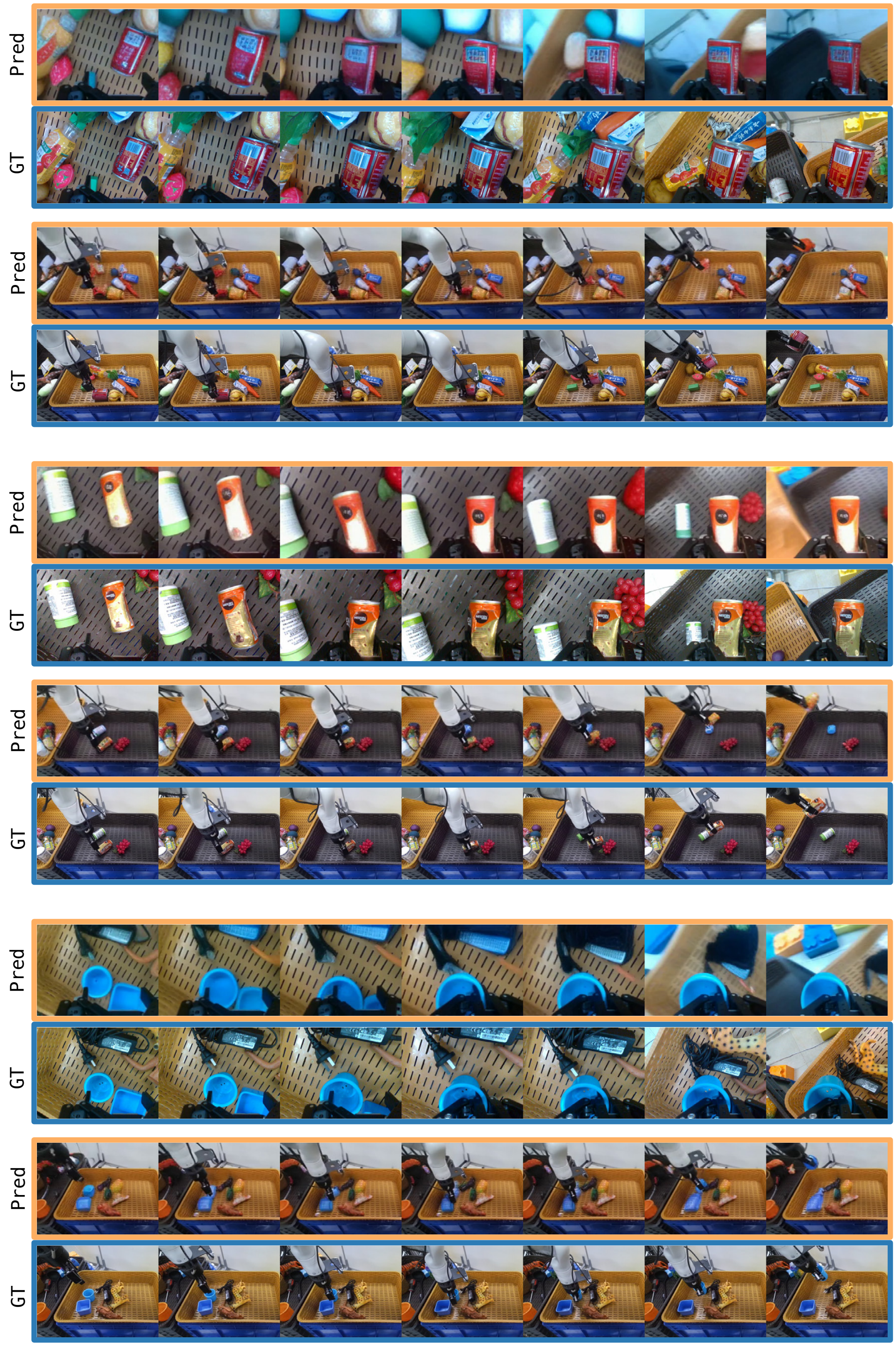

Figure 14: **Video Prediction (Pred) and Ground-Truth (GT) Rollouts of End-to-End Bin Picking (I).** We show autoregressive video predictions alongside the corresponding ground-truth videos captured from real-world rollouts. Both the views from the hand camera and the static head camera are shown.

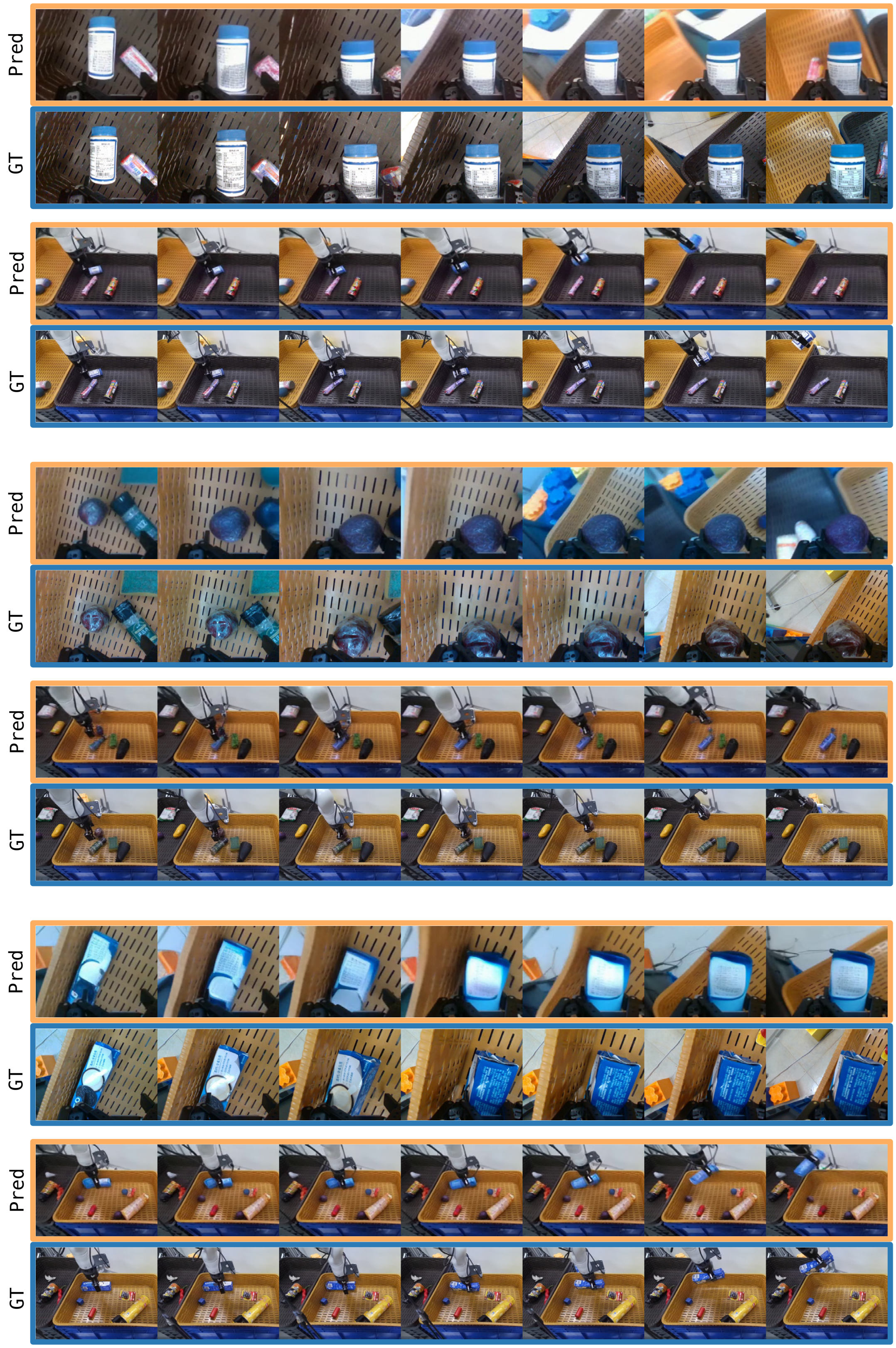

Figure 15: **Video Prediction (Pred) and Ground-Truth (GT) Rollouts of End-to-End Bin Picking (II).** We show autoregressive video predictions alongside the corresponding ground-truth videos captured from real-world rollouts. Both the views from the hand camera and the static head camera are shown.

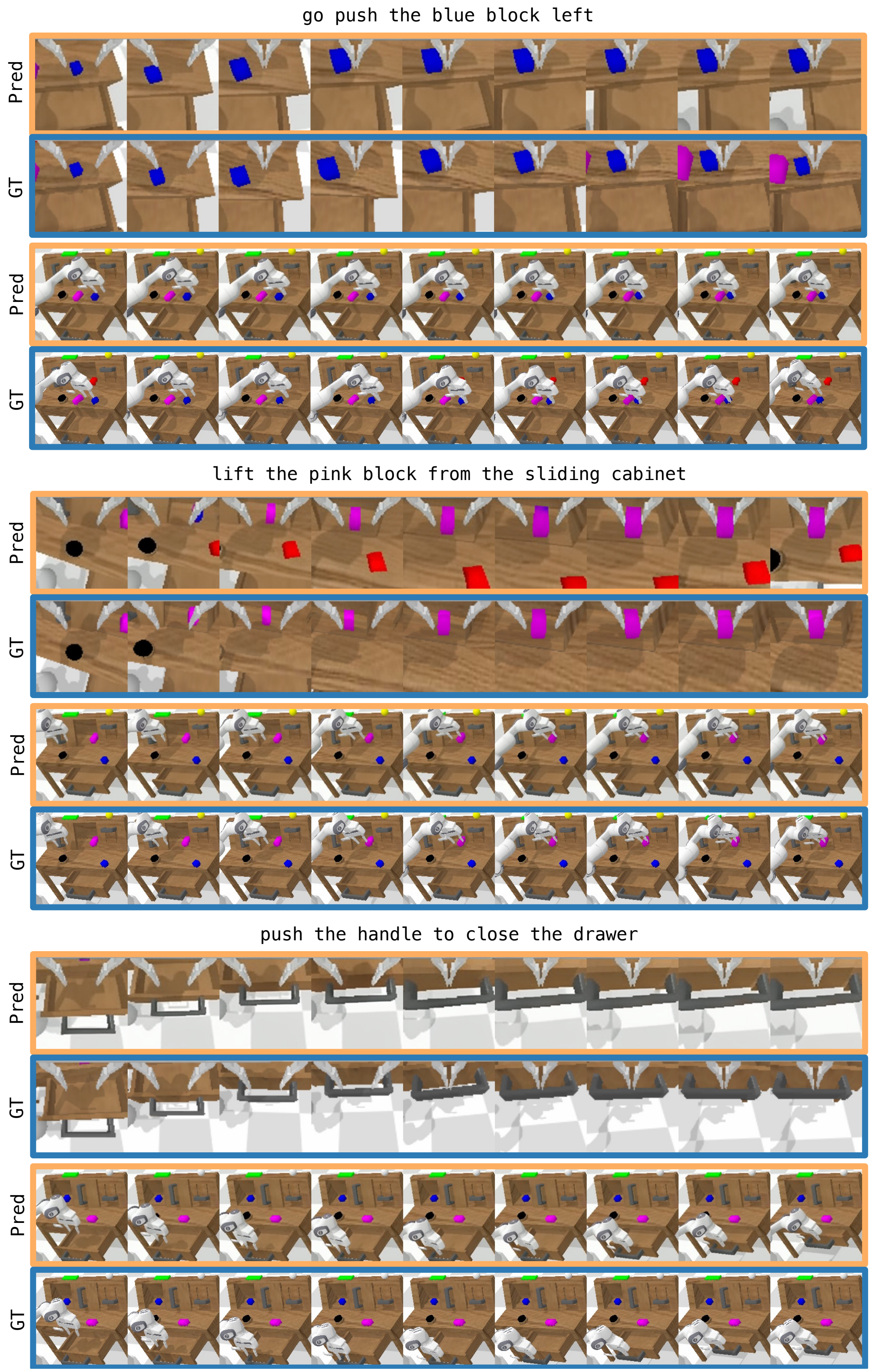

Figure 16: **Video Prediction (Pred) and Ground-Truth (GT) Rollouts of CALVIN Benchmark (I).** We show autoregressive video predictions alongside the corresponding ground-truth videos captured from the rollouts. Both the views from the hand camera and the static camera are shown.

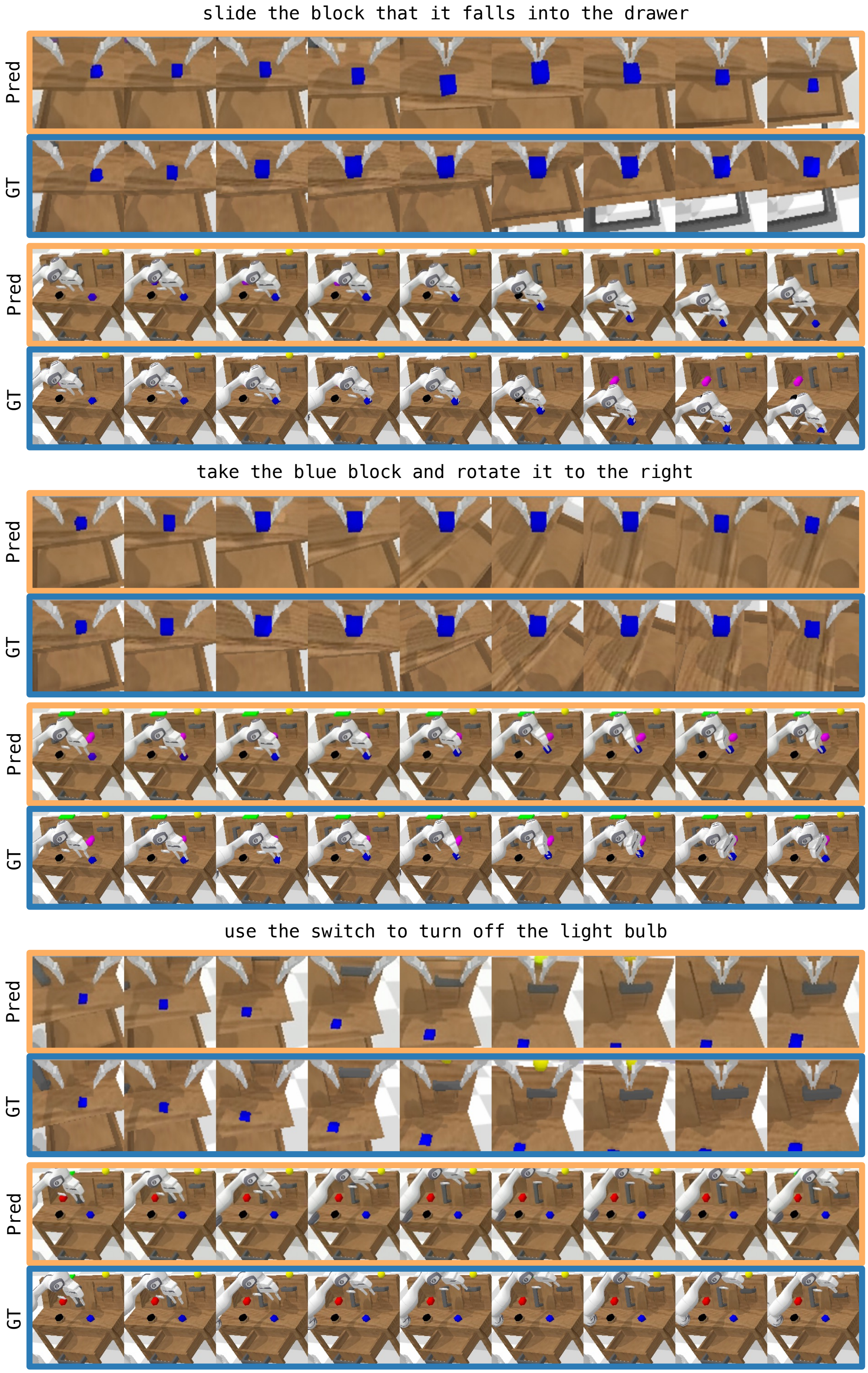

Figure 17: **Video Prediction (Pred) and Ground-Truth (GT) Rollouts of CALVIN Benchmark (II).** We show autoregressive video predictions alongside the corresponding ground-truth videos captured from the rollouts. Both the views from the hand camera and the static camera are shown.

4 Related Work

Generalist Robot Manipulation. A long-standing goal in robotics research is to develop a generalist robot agent that is able to accomplish a wide range of tasks in diverse environments. One of the most flexible ways to specify tasks is through natural languages [15, 29, 5, 28, 30, 31, 27, 26, 32, 33, 34, 35, 36, 37]. Pioneering studies explored using large-scale robot datasets to learn generalist policies which are able to accomplish a variety of tasks [32, 15, 29, 38]. To achieve generalization in unseen scenarios, some existing works combined data from other domains with robot data in policy training [39, 5, 29]. Recently, a number of works proposed to fine-tune a vision-language model which has been pre-trained on Internet-scaled data to obtain robust and generalizable robot policies [28, 29, 40]. In addition, some recent works resort to 3D information [41, 42, 43, 44] to achieve efficient policy learning by leveraging the geometry information contained in 3D data. Another line of works proposed to condition the policy with a goal image instead of a language [45, 46, 47, 48, 49]. And previous methods have also explored aligning the latent space of goal images and languages to enable both goal image condition and language condition during training [27, 31, 33]. GR-2 is a language-conditioned generalist robot manipulation agent. Unlike most previous works, it is first pre-trained on video generation with Internet-scale video datasets and then fine-tuned on robot data to predict both actions and videos.

Robot Learning with Pre-training. Inspired by the success in the field of vision [50] and languages [51], pre-training has gained increasing popularity in robot learning as it enhances the generalization capabilities and robustness of policies [52, 53, 54, 55, 56, 57, 58, 40, 5]. A popular approach is to first learn useful visual representations via masked modeling [53, 59, 60, 52] or contrastive learning [54, 61, 62, 63]. The learned representations are then used for downstream policy learning. RPT [55] performed self-supervised pre-training and showcased that pre-training with large robot datasets consistently surpasses training from scratch. In reinforcement learning (RL), previous works proposed to first train a world model to obtain latent state representations and then use them to train an RL agent [64, 56, 59]. VIPER [65] trained a video prediction model with expert data and utilized it as an action-free reward signal to train RL policies. Some model-based methods trained a video prediction model and combined it with an inverse dynamics model [66, 58, 67] or model predictive control [68, 69] to perform robot manipulation. VPT [70] first trained an inverse dynamics model with a small amount of data labeled with actions and used it to label a large amount of unlabeled data gathered from the web for policy training in Minecraft. Recent works trained policies based on models that have been pre-trained on Internet-scale data via end-to-end fine-tuning [28], co-training with robot data [29, 40], or a two-stream architecture [71]. The policy can make use of the web-scale knowledge obtained from pre-training in policy learning and showcases powerful generalization capabilities in unseen scenarios. Inspired by these works, we propose to leverage large-scale text-video data to perform video generative pre-training in our previous work GR-1 [5]. The motivation is that we believe videos contain valuable information on the dynamics of the environment and how the environment should evolve according to the text description. This information can facilitate action prediction during downstream policy learning. In comparison to GR-1 [5], GR-2 scales the number of pre-training videos from 0.8 million to 38 million, boosting the generalization capabilities in various unseen scenarios. In addition, the innovative model architecture facilitates more seamless knowledge transfer between pre-training and fine-tuning, leading to a policy that is more generalizable and robust.

5 Conclusions

We present GR-2, a generative robotic video-language-action model that is able to effectively learn a wide variety of tasks and generalize to unseen scenarios. GR-2 is first pre-trained on video generation with 38 million Internet videos. It is then fine-tuned on robot data to predict action trajectories and videos in tandem. It showcases strong multi-task learning capabilities, successfully completing more than 100 different manipulation tasks in the real world with a high success rate. It generalizes well to novel scenarios, including unseen backgrounds, environments, objects, and tasks. Moreover, GR-2 can perform bin-picking manipulation with over 100 objects in an end-to-end manner and handle unseen objects with remarkable robustness. We observe a strong correlation between the generated video and the action predicted alongside. In the future, we plan to enhance the generalization capabilities and robustness of action prediction, with a particular focus on improving the performance on unseen manipulation.

Contributions & Acknowledgements

- **Evaluation**: Chi-Lam Cheang, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Minzhao Zhu

- **Model & Training**: Ya Jing, Tao Kong, Yuxiao Liu, Hongtao Wu, Yichu Yang, Hanbo Zhang, Minzhao Zhu
- **Data Collection & Curation**: Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Yifeng Li, Hongtao Wu, Jiafeng Xu, Minzhao Zhu
- **Paper Writing**: Ya Jing, Tao Kong, Hang Li, Hongtao Wu, Yichu Yang

We thank the engineering team at ByteDance for their outstanding technical skills, and the data team for their meticulous work on data collection, annotation, and processing. We thank Xiao Ma for his helpful discussion and valuable advice on paper writing. Their commitment and expertise are crucial to the success of this project.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[5] 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.

[6] 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.

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[8] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[9] 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.

[10] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.

[12] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[13] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[14] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all, March 2024.

[15] 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.

[16] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[17] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[19] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[20] Taozheng Yang, Ya Jing, Hongtao Wu, Jiafeng Xu, Kuankuan Sima, Guangzeng Chen, Qie Sima, and Tao Kong. MOMA-Force: Visual-force imitation for real-world mobile manipulation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6847–6852. IEEE, 2023.

[21] 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 (RA-L)*, 7(3):7327–7334, 2022.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[23] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[25] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

[26] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. RoboAgent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. *arXiv preprint arXiv:2309.01918*, 2023.

[27] Oier Mees, Lukas Hermann, and Wolfram Burgard. What matters in language conditioned robotic imitation learning over unstructured data. *IEEE Robotics and Automation Letters*, 7(4):11205–11212, 2022.

[28] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023.

[29] 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.

[30] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2(3):6, 2022.

[31] Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data. *arXiv preprint arXiv:2005.07648*, 2020.

[32] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-Z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022.

[33] Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. In *Robotics: Science and Systems*, 2024.

[34] Huy Ha, Pete Florence, and Shuran Song. Scaling up and distilling down: Language-guided robot skill acquisition. In *Conference on Robot Learning*, pages 3766–3777. PMLR, 2023.

[35] 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.

[36] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[37] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023.

[38] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[39] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

[40] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[41] 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.

[42] Zhou Xian, Nikolaos Gkanatsios, Theophile Gervet, Tsung-Wei Ke, and Katerina Fragkiadaki. Chained-diffuser: Unifying trajectory diffusion and keypose prediction for robotic manipulation. In *7th Annual Conference on Robot Learning*, 2023.

[43] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*, 2024.

[44] Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3D: 3d feature field transformers for multi-task robotic manipulation. In *7th Annual Conference on Robot Learning*, 2023.

[45] Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X Lee, Maria Bauza, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, et al. RoboCat: A self-improving foundation agent for robotic manipulation. *arXiv preprint arXiv:2306.11706*, 2023.

[46] Hongtao Wu, Jikai Ye, Xin Meng, Chris Paxton, and Gregory S Chirikjian. Transporters with visual foresight for solving unseen rearrangement tasks. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10756–10763. IEEE, 2022.

[47] Daniel Seita, Pete Florence, Jonathan Tompson, Erwin Coumans, Vikas Sindhwani, Ken Goldberg, and Andy Zeng. Learning to rearrange deformable cables, fabrics, and bags with goal-conditioned transporter networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4568–4575. IEEE, 2021.

[48] Oliver Groth, Chia-Man Hung, Andrea Vedaldi, and Ingmar Posner. Goal-conditioned end-to-end visuomotor control for versatile skill primitives. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1319–1325. IEEE, 2021.

[49] Todor Davchev, Oleg Sushkov, Jean-Baptiste Regli, Stefan Schaal, Yusuf Aytar, Markus Wulfmeier, and Jon Scholz. Wish you were here: Hindsight goal selection for long-horizon dexterous manipulation. *arXiv preprint arXiv:2112.00597*, 2021.

[50] 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.

[51] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners, 2020.

[52] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

[53] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023.

[54] 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.

[55] Ilija Radosavovic, Baifeng Shi, Letian Fu, Ken Goldberg, Trevor Darrell, and Jitendra Malik. Robot learning with sensorimotor pre-training. In *Conference on Robot Learning*, pages 683–693. PMLR, 2023.

[56] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[57] 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.

[58] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

[59] Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked world models for visual control. In *Conference on Robot Learning*, pages 1332–1344. PMLR, 2023.

[60] 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.

[61] 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.

[62] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International conference on machine learning*, pages 5639–5650. PMLR, 2020.

[63] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018.

[64] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[65] 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.

[66] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[67] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023.

[68] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2786–2793. IEEE, 2017.

[69] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. MaskViT: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*, 2022.

[70] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

[71] Xingyu Lin, John So, Sashwat Mahalingam, Fangchen Liu, and Pieter Abbeel. SpawnNet: Learning generalizable visuomotor skills from pre-trained network. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4781–4787. IEEE, 2024.