VLA-OS: Structuring and Dissecting Planning Representations and Paradigms in Vision-Language-Action Models

Chongkai Gao[1] **Zixuan Liu**[1] **Zhenghao Chi**[1] **Junshan Huang**[2] **Xin Fei**[1]

Yiwen Hou[1] **Yuxuan Zhang**[1] **Yudi Lin**[1] **Zhirui Fang**[3] **Zeyu Jiang**[4]

Lin Shao[1†]

[1]National University of Singapore, [2]University of Science and Technology of China,
[3]Tsinghua University, [4]Nanyang Technological University
linshao@nus.edu.sg

Abstract: Recent studies on Vision-Language-Action (VLA) models have shifted from the end-to-end action-generation paradigm toward a pipeline involving task planning followed by action generation, demonstrating improved performance on various complex, long-horizon manipulation tasks. However, existing approaches vary significantly in terms of network architectures, planning paradigms, representations, and training data sources, making it challenging for researchers to identify the precise sources of performance gains and components to be further improved. To systematically investigate the impacts of different planning paradigms and representations isolating from network architectures and training data, in this paper, we introduce VLA-OS, a unified VLA architecture series capable of various task planning paradigms, and design a comprehensive suite of controlled experiments across diverse object categories (rigid and deformable), visual modalities (2D and 3D), environments (simulation and real-world), and end-effectors (grippers and dexterous hands). Our results demonstrate that: 1) visually grounded planning representations are generally better than language planning representations; 2) the Hierarchical-VLA paradigm generally achieves superior or comparable performance than other paradigms on task performance, pretraining, generalization ability, scalability, and continual learning ability, albeit at the cost of slower training and inference speeds. Video results are in https://nus-lins-lab.github.io/vlaos/.

Keywords: VLA Models, Task Planning, Foundation Models, Evaluation

1 Introduction

Building intelligent and generalizable robots capable of perceiving, reasoning about, and interacting with physical environments remains a persistent challenge in the robotics community [1, 2]. Recent studies have increasingly emphasized the development of foundational models for robot manipulation tasks by training large Vision-Language-Action models (VLAs) on extensive datasets [3, 4, 5, 6, 7, 8, 9, 10]. Different from end-to-end foundation models in computer vision [11, 12, 13] and natural language processing tasks [14, 15, 16], recent studies of VLAs have shifted toward a new paradigm capable of performing task planning and policy learning either simultaneously or sequentially [17, 18, 19, 20, 21, 22, 23, 4]. This shift arises from the inherent complexity of robotic manipulation tasks, which naturally exhibit hierarchical structures involving both high-level task planning and low-level physical interactions [24]. Compared to end-to-end VLAs that only generate actions, these methods demonstrate stronger capabilities in task reasoning and comprehension for long-horizon tasks [25, 4], better success rates [18, 20], and higher sample efficiency [26, 19, 27].

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

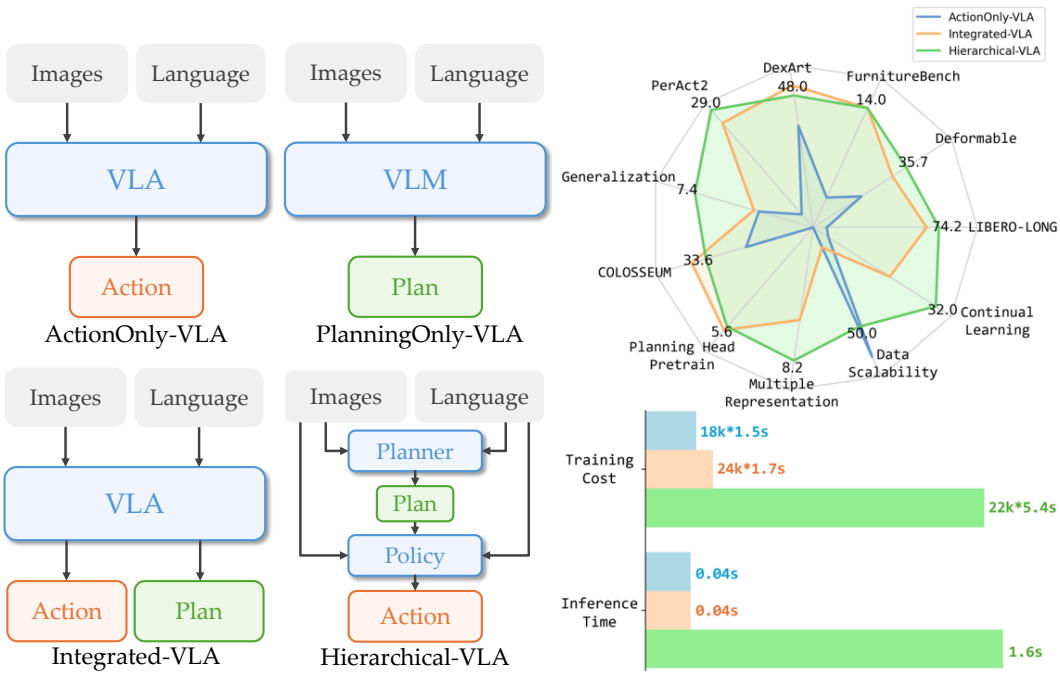


Figure 1: Left: four different VLA paradigms. Note that in this paper, we didn't explore PlanningOnly-VLA since they usually cannot be trained with the provided datasets and perform worse than others. Right: VLA paradigm comparison results. Hierarchical-VLA exhibits a generally better performance than ActionOnly-VLA and Integrated-VLA, while it incurs larger training and inference costs. This motivates future work on improving training and inference algorithms for Hierarchical-VLA models.

However, current task-planning approaches in VLA are mainly based on intuitive designs and lack fair and systematic comparisons, as these methods vary along multiple dimensions, including network architectures, planning paradigms, data representations, and training data sources. For example, some works [20, 22, 19, 23] use a separate high-level task planning model to generate various task planning representations for a low-level VLA model, while others [4, 18, 17] use a single VLA to generate task planning representations and actions together. Consequently, substantial disagreement remains within the VLA community regarding the appropriate design and effective utilization of task planning. This makes it difficult for researchers to clearly identify which specific component contributes to performance gains or requires further improvement, hindering progress in the field.

Among these challenges, five core questions stand out: 1) **Representation**: What representation should we adopt for task planning and policy learning? Does using multiple representations yield better results, or could they conflict with one another? 2) **Paradigm**: Should we employ a monolithic model that jointly performs task planning and policy learning, or should we opt for a hierarchical paradigm where two separate models handle these tasks independently? 3) **Bottleneck**: Between task planning and policy learning, which presents a greater challenge for current manipulation tasks? 4) **Scalability and Pretraining**: Do VLAs that incorporate task planning preserve the advantageous properties of end-to-end foundation models, such as model and data scalability, as well as benefits derived from pretraining? and 5) **Performance**: Do VLAs employing task planning have better generalization and continual learning ability than end-to-end VLAs? Addressing these questions will provide the community with a clearer understanding of how task planning works in VLA models, and offer empirical evidence and guidance for future developments.

In this work, we aim to answer these questions with systematic and controllable experiments. First, to avoid biases introduced by specific neural network choices, we develop VLA-OS[1] model series: a unified and composable family of VLA models for general-purpose manipulation tasks capable

[1]"OS" stands for "Operating System" and designates that our model family provides unified and organized interfaces of advanced VLA architectures with various planning heads and different paradigms for users.

of different task planning paradigms. Concretely, we designed VLA-OS-A, VLA-OS-I, and VLA-OS-H that correspond to three mainstream VLA paradigms (ActionOnly-VLA, Integrated-VLA, Hierarchical-VLA), respectively, as illustrated in Figure 1. VLA-OS series features a unified, interchangeable VLM backbone that can be directly downloaded from HuggingFace, various plug-and-play planning heads for different representations, and two different action heads both supporting 2D/3D tasks, as shown in Figure 2. We show in Table 1 that VLA-OS exhibits superior performance compared to most existing VLA methods with fewer parameters and without pretraining.

Next, to answer the **representation** question, we annotate three kinds of task reasoning representations, including language reasoning, visual reasoning, and goal images, and conducted exhaustive combinatorial experiments with Integrated-VLA and Hierarchical-VLA models on LIBERO [28] benchmark to identify representations that yield optimal performance. Subsequently, employing the optimal representations identified, we conducted performance comparisons among three VLA paradigms on six benchmarks to answer the **paradigm** question, including rigid body manipulation tasks [28], visual generalization tasks [29], complex long-horizon tasks [30], real-world deformable manipulation tasks, dexterous manipulation tasks [31], and dual-arm manipulation tasks [32]. Furthermore, to answer the **bottleneck** question, we designed a novel set of evaluation metrics tailored to separately assess the performance of task planning and policy learning parts. To answer the **scalability** question, we use LIBERO [28] to test the model and data scalability as well as the effects of pretraining among different paradigms. And lastly, we test the generalization capabilities and continual learning ability of different VLA paradigms to answer the **performance** question.

Our experiments yield three primary findings: 1) Visually grounded planning representations (visual reasoning and image foresight planning) outperform language-based planning representations across multiple dimensions including task performance, generalization, training and speed, and low-level policy execution; 2) Hierarchical-VLA matches or exceeds the performance of Integrated-VLA and ActionOnly-VLA in terms of task performance, generalization, scalability, planning scores, continual learning, and gains from task-planning pretraining, albeit at the expense of increased training cost and slower inference; 3) On LIBERO [28] benchmark tasks, policy learning is consistently more challenging than task planning, regardless of which planning representation is used. We believe that our findings (as well as source codes, annotated datasets, and checkpoints) will provide significant help and guidance for future research within the VLA community and the broader robotics community.

2 VLA-OS Model Family Design

2.1 Preliminaries

We study imitation learning for robot manipulation tasks. Specifically, for each task $\mathcal{T}$, we assume a set of demonstrations $\mathcal{D}_{\mathcal{T}} = \{(o_i^1, a_i^1), (o_i^2, a_i^2), \cdots, (o_i^{T_i}, a_i^{T_i})\}_{i=1}^N$ and a language goal are given, where T_i is the episode length, o is the observation, a is the robot action, and N is the number of demonstrations. We use a history of multi-view images and proprioception information as observations. In this work, we set the image resolution as 224×224. For actions, we use a normalized continuous delta end-effector pose δ_p action space and gripper open/close action σ for training. We also let the policy generate action chunks, i.e., $a_t = ([\delta_p, \sigma]^t, \cdots, [\delta_p, \sigma]^{t+L-1})$. For dexterous hands, we use the delta joint values as the action space. We train the policy with either flow matching [33, 34] loss (for multi-modal demonstration datasets) or L1 loss (for simple and uni-modal demonstration datasets) under the suggestion of previous works [35, 3, 36, 37].

2.2 VLA-OS-A for ActionOnly-VLA Paradigm

VLA-OS-A model series directly generates actions without task planning stages. It is also used as the base model for other paradigms. We design a block-wise causal attention VLA drawing inspiration from [3], as shown in Figure 2. First, a VLM encodes the visual and language inputs, where the vision encoder will encode input image patches and project them into language embedding space with an MLP. Then, we use a separate set of weights as an action head for the robotics-specific tokens (action and proprioception states). The action head is a transformer decoder that has the same number of

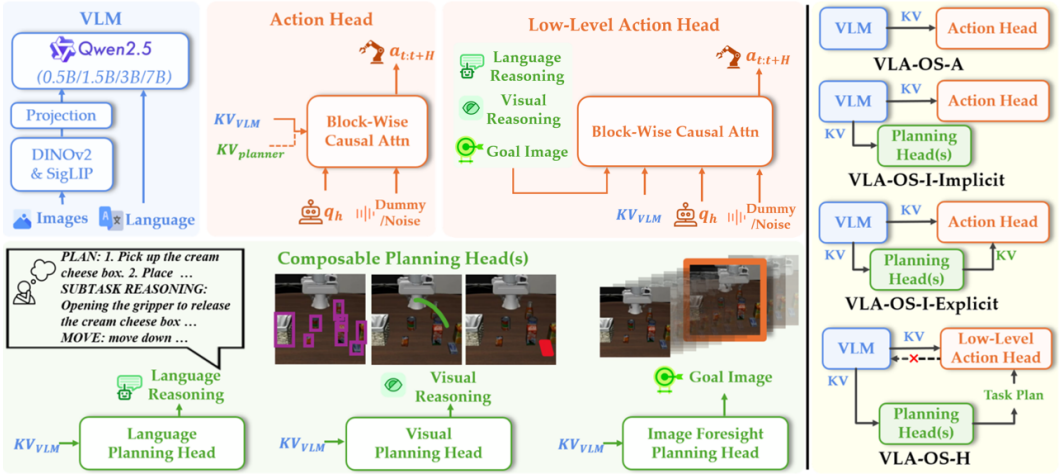


Figure 2: The VLA-OS model family. Left: the VLM and the composable heads. Our VLM has the same architecture with different numbers of parameters. Although we only draw Qwen2.5 here, our code supports any kind of LLM backbone from HuggingFace. Right: four VLA-OS architectures used in our experiments. To minimize the effects of different numbers of parameters in different models, we restrict the number of parameters of all heads to about 5% of the VLM.

layers as the LLM, and for each layer, the queries of the proprioception tokens can attend to both the keys and values from the LLM and the proprioception keys and values, and the queries of the action tokens can attend to the keys and values from the LLM, the proprioception tokens, and themselves.

Compared to π_0 [3], we make two changes in VLA-OS-A: 1) we use an ensemble of vision encoders (DINOV2 [11]+SigLIP [38]), which is proven to be better than using a single vision encoder [39]; 2) to support the model scalability experiments, we need a set of LLMs with the same structure but have different number of parameters. Thus, we choose Qwen2.5 [16] LLM series with 0.5B, 1.5B, 3B, 7B pretrained checkpoints rather than the original PaliGamma [40]. To make it a VLM, we finetune Qwen2.5 LLMs with the vision encoders and the projector on LLaVa v1.5 [41] data mixture by ourselves. We call our VLA family that uses 0.5B, 1.5B, 3B, 7B LLM backbones with suffixes of -S(mall), -B(ase), -M(iddle), and -L(arge). Detailed information can be found in Appendix E. Note, although we use Qwen2.5 in this work, our codes support any kind of LLM from HuggingFace, which makes VLA-OS highly flexible compared to [3] that is restricted to a specific backbone.

For 3D action head, we also take in multi-view depth images as input, and fuse the multi-view RGBD images to 3D point cloud using camera intrinsics and extrinsics, and inject additional CLIP features onto the point cloud, as in 3D diffuser actor [42]. We also downsample the point cloud with farthest point sampling. Each point from the downsampled point cloud will be seen as a token and these 3D tokens are sent to the action head as additional inputs.

2.3 VLA-OS-I for Integrated-VLA Paradigm

To perform task planning with different kinds of representations, we design three kinds of task planning heads for VLA-OS. We first annotate three kinds of task reasoning datasets corresponding to each planning representation, as shown in Figure 3. Here we only briefly introduce each of them. Details of the data annotation process can be found in Appendix D.

The **language reasoning** data contains 8 different keys [18] for each timestep, including `Task`, `Plan`, `Subtask`, `Subtask Reason`, `Move`, `Move Reason`, `Gripper Position`, and `Object Bounding Boxes`, containing the understanding of the scene and decomposition of the task. The **visual reasoning** data contains spatial semantic information and is more grounded in input images compared to language reasoning. We follow [43, 44] and use location tokens `<loc i>` to represent the i-th bin token from top-left to bottom-right. We use this kind of token to represent `object bounding boxes`, `end-effector flow`, and `target object affordance` as the visual planning representations. The **image foresight reasoning** data is a third-person view image at the K-th future step as the short-horizon goal image.

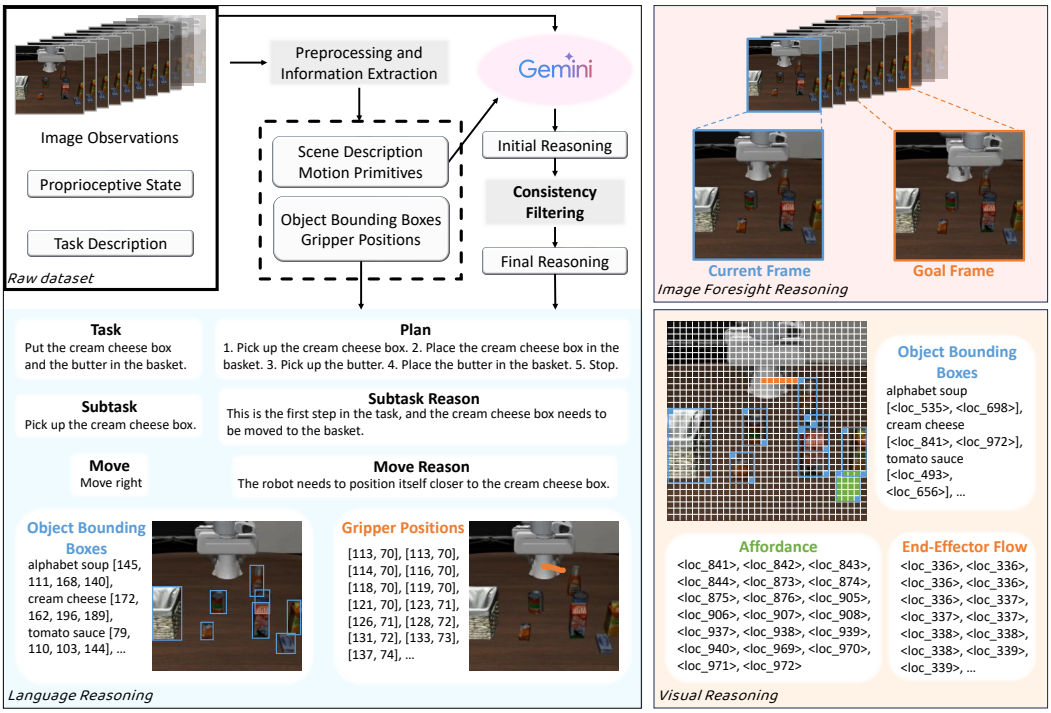


Figure 3: The formats and contents of the language reasoning dataset, the visual reasoning dataset, and the image foresight reasoning dataset in this work. We use various vision-language models for data annotation. We illustrate the language reasoning data annotation process on the top left part.

We then design language planning head, visual planning head, and image foresight planning head for each kind of representation, as shown in Figure 2. All of them are transformers that have the same number of layers with the LLM backbone, and use the block-wise causal attention mechanism to acquire the keys and values from each layer of the LLM backbone as conditions. The language planning head uses the LLM's tokenizer for decoding, whereas the visual planning head uses an extended tokenizer vocabulary to predict location tokens. The image foresight planning head is an autoregressive image generation model similar to the recent SOTA image generator [45]. It auto-regressively generates the image in a coarse-to-fine paradigm proposed by VAR [46]. The language and visual planning heads are trained with cross-entropy loss, while the image foresight planning head is trained with the special loss in [45].

For all three planning heads, there are two kinds of ways to use them: 1) implicit planning: the action head is independent of the planning heads, i.e., the planning heads serve as auxiliary losses for the VLA training and will not be executed during inference. This may help the model avoid planning accumulation error and improve the inference speed; 2) explicit planning: the action head also attends to the keys and values from each layer of the planning heads, and during inference, the VLA must first perform task planning before generating actions. This may help solve complex tasks in a chain-of-thought [47, 18, 17] manner.

2.4 VLA-OS-H for Hierarchical-VLA Paradigm

This model uses two networks for task planning and policy learning respectively. As shown in Figure 2, we use the VLM together with planning heads for task planning, and modify the action head to an encoder-decoder transformer for policy learning. This action head can take as input the images, proprioception observations, and the planning representations to generate actions. To keep the comparison fair, we make the layer of the encoder and decoder of the action head half of the other two VLA-OS paradigms. We also give frozen image features from AM-Radio [48] and language features from Qwen2.5 [16] for the inputs of the action head to compensate for deficiencies in visual and linguistic features not captured by the VLM. Training details are in Appendix E.

Figure 4: Benchmarks used in our evaluations, including LIBERO [28] and FurnitureBench [30] for 2D rigid body manipulation experiments, The COLOSSEUM [29] for 3D and generalization evaluation, real-world deformable object manipulation tasks (fold the handkerchief, unfold the jean, and straighten the rope), DexArt [31] for dexterous tasks, and PerAct2 [32] for dual-arm tasks.

3 Experiments and Findings

In this section, we perform systematic and controllable experiments with the VLA-OS model series on various manipulation tasks shown in Figure 4 to answer the research questions in Section 1. The task planning experiments, training costs, scalability experiments, and continual learning experiments are in Appendix B. Detailed experimental settings are in Appendix E. All models are trained on 8×NVIDIA A100 80G GPUs.

3.1 Sanity Check of VLA-OS

Before investigating different VLA paradigms for our research questions, we first verify the correctness and basic performance of our VLA-OS models to serve as a foundational sanity check. We train VLA-OS-A-S on four suites from LIBERO [28] (LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, LIBERO-Long) from scratch with L1 loss and compare them with Diffusion-Policy [49], fine-tuned OpenVLA [5], fine-tuned CoT-VLA [17], fine-tuned DiT Policy [50], and the state-of-the-art methods: fine-tuned π_0 [3] and its variant π_0-FAST [51]. Results are shown in Table 1.

Table 1: Sanity check. Success rates on four LIBERO benchmarks. Baseline results are from their papers [5, 3, 35]. Our results are the average of top-3 checkpoints averaged over 20 rollouts for each task suite. **Bold** indicates the best result except SOTA, and underline indicates comparable result.

	LIBERO-Spatial	LIBERO-Object	LIBERO-Goal	LIBERO-Long	Average
Diffusion Policy [49] (scratch)	78.3	92.5	68.3	50.5	72.4
OpenVLA [5] (fine-tuned)	84.7	88.4	79.2	53.7	76.5
CoT-VLA [17] (fine-tuned)	87.5	91.6	87.6	**69.0**	81.1
DiT Policy [50] (fine-tuned)	84.2	96.3	85.4	63.8	82.4
π_0-FAST [51] (fine-tuned)	**96.4**	**96.8**	88.6	60.2	85.5
VLA-OS-A-S (scratch, ours)	87.0	96.5	**92.7**	66.0	**85.6**
π_0 [3] (fine-tuned, SOTA)	96.8	98.8	95.8	85.2	94.2

We can see that VLA-OS-A-S performs better (+13.2%) than Diffusion Policy (trained from scratch) and the fine-tuned OpenVLA model (+9.1%), CoT-VLA (+4.5%), and DiT Policy (+3.2%), and is comparable to fine-tuned π_0-FAST (+0.1%). Although our model is worse than the SOTA method, these results sufficiently demonstrate that our model design is excellent and competitive. Note that VLA-OS-A-S is **trained from scratch** and utilizes **only a 0.5B LLM backbone**.

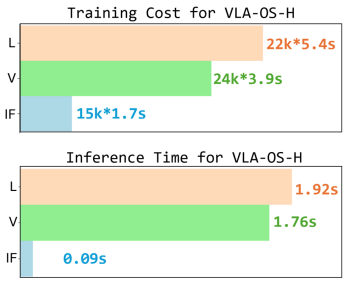
Finding 1: For downstream tasks, larger VLA models trained on large-scale datasets do not necessarily outperform smaller models that are trained from scratch. Model architectures and algorithmic designs are still important at the current moment.

3.2 Planning Representation Experiments

To explore which representation is better for task planning and policy learning, we perform comprehensive experiments with language planning (**L**), visual planning (**V**), image foresight planning (**IF**),

	VLA-OS-A	VLA-OS-I	VLA-OS-H
LIBERO-LONG (2D)	66.0	73.3 (↑ 7.3)	**74.2** (↑ 8.2)
The COLOSSEUM (3D)	34.4	**35.7** (↑ 1.3)	35.3 (↑ 0.9)
Deformable (Real-World)	28.5	**35.4** (↑ 6.9)	33.6 (↑ 5.1)
FurnitureBench	11.0	**14.0** (↑ 3.0)	**14.0** (↑ 3.0)
DexArt	45.0	**49.0** (↑ 4.0)	48.0 (↑ 3.0)
PerAct2	21.0	28.0 (↑ 7.0)	**29.0** (↑ 8.0)
Generalization	6.1	6.2 (↑ 0.1)	**7.4** (↑ 1.3)
Planning Head Pretraining	–	79.1 (↑**5.8**)	79.8 (↑5.6)

(a) Success rates of different VLA paradigms on more benchmarks, as well as the generalization and task planning pretraining experiments. All results are averaged over 20 rollouts among 3 best checkpoints.



(b) Training cost and inference time for different representations.

Figure 6: More results for different paradigms and inference time and training cost for different representations. Results of the right figure are calculated from the LIBERO-LONG benchmark.

and their combinations on LIBERO-LONG [28] benchmark that contains 10 long-horizon tasks with 50 demonstrations in each task for VLA-OS-I and VLA-OS-H. The best representation will be used as the default representation for all later experiments. Table 2 shows the results.

Table 2: Different planning representation comparison on LIBERO-Long. All results are the average of top-3 checkpoints averaged over 20 rollouts. Numbers in parentheses indicate the change relative to the result of VLA-OS-A in Table 1.

	L	V	IF	L+V	L+IF	V+IF	L+V+IF
VLA-OS-I-I	68.0 (↑2.0)	71.0 (↑5.0)	72.5 (↑6.5)	66.7 (↑0.7)	**73.3** (↑7.3)	71.0 (↑5.0)	71.7 (↑5.7)
VLA-OS-I-E	60.5 (↓5.5)	52.5 (↓13.5)	**67.5** (↑1.5)	42.7 (↓23.3)	56.7 (↓9.3)	56.7 (↓9.3)	50.7 (↓15.3)
VLA-OS-H	63.5 (↓2.5)	69.0 (↑3.0)	71.7 (↑5.7)	71.5 (↑5.5)	72.0 (↑6.0)	73.7 (↑7.7)	**74.2** (↑8.2)

Finding 2: For Integrated-VLA paradigm, implicit planning can yield a positive performance gain, whereas explicit planning incurs a significant performance degradation when trained from scratch.

For qualitative comparisons, we show in Figure 5 an example that when VLA-OS-H uses the same planning heads as VLA-OS-I-E where there are some planning errors, it can correct the behavior while VLA-OS-I-E cannot.

Finding 3: Visually grounded planning representations work better than language planning representations, and also have faster inference speed and smaller training cost.

From the results in Table 2, we can see that visual planning and image foresight planning are better than language planning (↑5.75 v.s. ↑2.0 for VLA-OS-I-I and ↑4.35 v.s. ↓2.5 for VLA-OS-H). We also illustrate the inference speed and training cost in Figure 6b (introduced in Section B.2) to show the speed and cost advantages of visually grounded planning representations.

Finding 4: When employing multiple planning representations concurrently, Hierarchical-VLA outperforms Integrated-VLA paradigms.

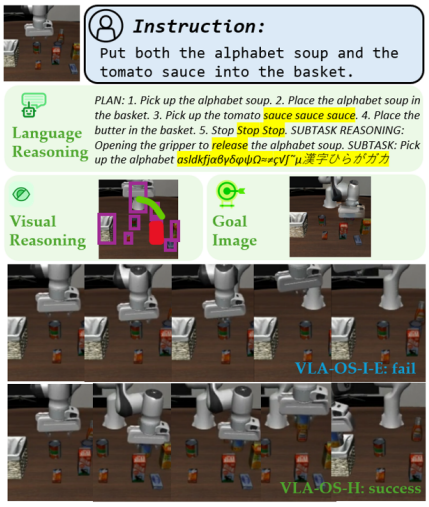


Figure 5: Comparison between VLA-OS-I-E and VLA-OS-H with the same planning errors. The three planning representations shown in this figure all have small planning errors (highlighted).

3.3 More Performance, Generalization, and Benefit from Planning Head Pretraining

To further compare different planning paradigms, we perform additional experiments to explore their performance on: 1) more manipulation benchmarks including 3D manipulation tasks [29], real-world deformable tasks, furniture assembly tasks [30], dexterous manipulation tasks [31], and dual-arm manipulation tasks [32]; 2) generalization ability; and 3) benefits from planning head pretraining. For 1), in COLOSSEUM, we train and test on the *No-Perturbation* setting. For 2), we use THE

COLOSSEUM and train on *No-Perturbation* but test on *ALL-Perturbation* setting, including changes in color, texture, size of objects, table-tops, backgrounds, lighting, distractors, physical properties, and camera poses. For 3), a lot of literature [19, 21, 52, 20, 53, 54] claim that the primary advantage of using task planning in VLA rather than ActionOnly-VLA is that their task-planning components can be trained on large-scale task-agnostic planning data without costly action annotations. Here, we train them on LIBERO-90, a larger dataset with 90 manipulation tasks and 50 demonstrations for each task. We only train the planning components, i.e., the VLM and planning heads. Then we fine-tune the pretrained VLM and planning heads together with the action head on LIBERO-LONG with both the task reasoning and policy learning losses. Results are in Table 6a.

Finding 5: *Integrated-VLA and Hierarchical-VLA outperform ActionOnly-VLA across a broad spectrum of tasks (2D, 3D, simulation, and real-world), with their performances largely comparable.*

Finding 6: *Both Integrated-VLA and Hierarchical-VLA benefit similarly from task-planning pretraining, exhibiting analogous gains in task success rate.*

Finding 7: *Hierarchical-VLA demonstrates the best generalization ability.*

4 Conclusion and Limitation

We provide a systematic investigation across different VLA paradigms and task planning representations through various kinds of manipulation tasks. Experiments show the superiority of visually grounded planning representations and the Hierarchical-VLA paradigm. The limitations of this paper are in Appendix F. Our findings can be summarized as follows:

1. The time has not yet come to scale up VLA model sizes.
2. Visually grounded representations (visual and image foresight) are better than language planning representations in terms of success rates, low-level following, and continual learning.
3. Integrated-VLA and Hierarchical-VLA outperform ActionOnly-VLA on task performance and generalization ability, but incur faster forgetting.
4. Integrated-VLA and Hierarchical-VLA perform comparably on task performance and Planning Head Pretraining, but Hierarchical-VLA generalizes better and has better task-planning performance.
5. All VLA paradigms have the data scalability. For tasks trained from scratch with roughly 5,000 demonstrations, the LLM backbone should be limited to 0.5B parameters, or keeping the total model size under 1B parameters.

We believe our findings offer meaningful insights that can inform future research in VLA and the broader robotics community. We recommend the following research directions for the community based on our findings:

1. Why are visually grounded representations better than language?
2. How to avoid gradient conflict between planning head losses and action head losses on the VLM backbone? This is because that in both explicit v.s. implicit and Hierarchical v.s. Integrated comparisons, reducing the influence of action head training on VLM improves the performance.
3. How to design network architectures to effectively extract information from VLM? There could be better mechanism than the current KV extraction method.
4. How to design faster planning heads for autoregressive planning heads?
5. How to design better low-level action heads with better planning-following ability?
6. How to construct large-scale task planning datasets for VLA? How to transfer current datasets to task planning datasets? This is because that our finding 6 shows that task planning pretraining is useful.

Acknowledgments

We thank Zhixuan Xu for his valuable discussion and his guidance for drawing the pictures.

References

[1] Y. Hu, Q. Xie, V. Jain, J. Francis, J. Patrikar, N. Keetha, S. Kim, Y. Xie, T. Zhang, H.-S. Fang, et al. Toward general-purpose robots via foundation models: A survey and meta-analysis. *arXiv preprint arXiv:2312.08782*, 2023.

[2] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, page 02783649241281508, 2023.

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

[4] J. Wen, M. Zhu, Y. Zhu, Z. Tang, J. Li, Z. Zhou, C. Li, X. Liu, Y. Peng, C. Shen, et al. Diffusion-vla: Scaling robot foundation models via unified diffusion and autoregression. *arXiv preprint arXiv:2412.03293*, 2024.

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

[6] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

[7] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

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

[9] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[10] FigureAI. Helix: A vision-language-action model for generalist humanoid control. https://www.figure.ai/news/helix, 2025. Accessed: 2025-02-20.

[11] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[13] N. Karaev, I. Makarov, J. Wang, N. Neverova, A. Vedaldi, and C. Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024.

[14] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[15] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[16] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[17] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, A. Handa, M.-Y. Liu, D. Xiang, G. Wetzstein, and T.-Y. Lin. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. *arXiv preprint arXiv:2503.22020*, 2025.

[18] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

[19] C. Gao, H. Zhang, Z. Xu, Z. Cai, and L. Shao. Flip: Flow-centric generative planning for general-purpose manipulation tasks. *arXiv preprint arXiv:2412.08261*, 2024.

[20] L. X. Shi, B. Ichter, M. Equi, L. Ke, K. Pertsch, Q. Vuong, J. Tanner, A. Walling, H. Wang, N. Fusai, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025.

[21] Y. Li, Y. Deng, J. Zhang, J. Jang, M. Memmel, R. Yu, C. R. Garrett, F. Ramos, D. Fox, A. Li, et al. Hamster: Hierarchical action models for open-world robot manipulation. *arXiv preprint arXiv:2502.05485*, 2025.

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

[23] Y. Tian, S. Yang, J. Zeng, P. Wang, D. Lin, H. Dong, and J. Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. *arXiv preprint arXiv:2412.15109*, 2024.

[24] M. Brady. *Robot motion: Planning and control*. MIT press, 1982.

[25] Z. Zhou, Y. Zhu, M. Zhu, J. Wen, N. Liu, Z. Xu, W. Meng, R. Cheng, Y. Peng, C. Shen, et al. Chatvla: Unified multimodal understanding and robot control with vision-language-action model. *arXiv preprint arXiv:2502.14420*, 2025.

[26] Y. Ji, H. Tan, J. Shi, X. Hao, Y. Zhang, H. Zhang, P. Wang, M. Zhao, Y. Mu, P. An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:2502.21257*, 2025.

[27] Z. Xu, C. Gao, Z. Liu, G. Yang, C. Tie, H. Zheng, H. Zhou, W. Peng, D. Wang, T. Hu, et al. Manifoundation model for general-purpose robotic manipulation of contact synthesis with arbitrary objects and robots. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10905–10912. IEEE, 2024.

[28] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

[29] W. Pumacay, I. Singh, J. Duan, R. Krishna, J. Thomason, and D. Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation. *arXiv preprint arXiv:2402.08191*, 2024.

[30] M. Heo, Y. Lee, D. Lee, and J. J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. *The International Journal of Robotics Research*, page 02783649241304789, 2023.

[31] C. Bao, H. Xu, Y. Qin, and X. Wang. Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21190–21200, 2023.

[32] M. Grotz, M. Shridhar, Y.-W. Chao, T. Asfour, and D. Fox. Peract2: Benchmarking and learning for robotic bimanual manipulation tasks. In *CoRL 2024 Workshop on Whole-body Control and Bimanual Manipulation: Applications in Humanoids and Beyond*, 2024.

[33] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[34] Q. Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

[35] M. J. Kim, C. Finn, and P. Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

[36] S. Belkhale and D. Sadigh. Minivla: A better vla with a smaller footprint. https://ai.stanford.edu/blog/minivla/, 2024.

[37] X. Li, P. Li, M. Liu, D. Wang, J. Liu, B. Kang, X. Ma, T. Kong, H. Zhang, and H. Liu. Towards generalist robot policies: What matters in building vision-language-action models. *arXiv preprint arXiv:2412.14058*, 2024.

[38] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

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

[40] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[41] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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

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

[44] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024.

[45] J. Han, J. Liu, Y. Jiang, B. Yan, Y. Zhang, Z. Yuan, B. Peng, and X. Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024.

[46] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

[47] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[48] M. Ranzinger, G. Heinrich, J. Kautz, and P. Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12490–12500, 2024.

[49] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[50] Z. Hou, T. Zhang, Y. Xiong, H. Duan, H. Pu, R. Tong, C. Zhao, X. Zhu, Y. Qiao, J. Dai, et al. Dita: Scaling diffusion transformer for generalist vision-language-action policy. *arXiv preprint arXiv:2503.19757*, 2025.

[51] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

[52] Q. Bu, Y. Yang, J. Cai, S. Gao, G. Ren, M. Yao, P. Luo, and H. Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.xxxxx*, 2025.

[53] W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

[54] M. Yang, Y. Du, K. Ghasemipour, J. Tompson, D. Schuurmans, and P. Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023.

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

[56] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR, 2022.

[57] D. Qiu, W. Ma, Z. Pan, H. Xiong, and J. Liang. Open-vocabulary mobile manipulation in unseen dynamic environments with 3d semantic maps. *arXiv preprint arXiv:2406.18115*, 2024.

[58] R. Shah, A. Yu, Y. Zhu, Y. Zhu, and R. Martín-Martín. Bumble: Unifying reasoning and acting with vision-language models for building-wide mobile manipulation. *arXiv preprint arXiv:2410.06237*, 2024.

[59] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, et al. Palm-e: An embodied multimodal language model. 2023.

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

[61] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE, 2023.

[62] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024.

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

[64] S. Wang, M. Han, Z. Jiao, Z. Zhang, Y. N. Wu, S.-C. Zhu, and H. Liu. Llm^ 3: Large language model-based task and motion planning with motion failure reasoning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12086–12092. IEEE, 2024.

[65] S. Nasiriany, F. Xia, W. Yu, T. Xiao, J. Liang, I. Dasgupta, A. Xie, D. Driess, A. Wahid, Z. Xu, et al. Pivot: Iterative visual prompting elicits actionable knowledge for vlms. *arXiv preprint arXiv:2402.07872*, 2024.

[66] C. Gao, Y. Jiang, and F. Chen. Transferring hierarchical structures with dual meta imitation learning. In *Conference on Robot Learning*, pages 762–773. PMLR, 2023.

[67] C. Gao, Z. Li, H. Gao, and F. Chen. Iterative interactive modeling for knotting plastic bags. In *Conference on Robot Learning*, pages 571–582. PMLR, 2023.

[68] S. Zhang, Z. Xu, P. Liu, X. Yu, Y. Li, Q. Gao, Z. Fei, Z. Yin, Z. Wu, Y.-G. Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. *arXiv preprint arXiv:2412.18194*, 2024.

[69] R. Yang, H. Chen, J. Zhang, M. Zhao, C. Qian, K. Wang, Q. Wang, T. V. Koripella, M. Movahedi, M. Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025.

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

[71] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[72] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[73] J. Wen, Y. Zhu, J. Li, M. Zhu, Z. Tang, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*, 2025.

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

[75] J. Zheng, J. Li, D. Liu, Y. Zheng, Z. Wang, Z. Ou, Y. Liu, J. Liu, Y.-Q. Zhang, and X. Zhan. Universal actions for enhanced embodied foundation models. *arXiv preprint arXiv:2501.10105*, 2025.

[76] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[77] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, K. Ghasemipour, C. Finn, and A. Wahid. Aloha unleashed: A simple recipe for robot dexterity. *arXiv preprint arXiv:2410.13126*, 2024.

[78] R. Zheng, Y. Liang, S. Huang, J. Gao, H. Daumé III, A. Kolobov, F. Huang, and J. Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024.

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

[80] M. Reuss, Ö. E. Yağmurlu, F. Wenzel, and R. Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. *arXiv preprint arXiv:2407.05996*, 2024.

[81] S. Ye, J. Jang, B. Jeon, S. Joo, J. Yang, B. Peng, A. Mandlekar, R. Tan, Y.-W. Chao, B. Y. Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.

[82] D. Schmidt and M. Jiang. Learning to act without actions. *arXiv preprint arXiv:2312.10812*, 2023.

[83] X. Chen, J. Guo, T. He, C. Zhang, P. Zhang, D. C. Yang, L. Zhao, and J. Bian. Igor: Image-goal representations are the atomic control units for foundation models in embodied ai. *arXiv preprint arXiv:2411.00785*, 2024.

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

[85] H. Chen, J. Li, R. Wu, Y. Liu, Y. Hou, Z. Xu, J. Guo, C. Gao, Z. Wei, S. Xu, et al. Metafold: Language-guided multi-category garment folding framework via trajectory generation and foundation model. *arXiv preprint arXiv:2503.08372*, 2025.

[86] Z. Wei, Z. Xu, J. Guo, Y. Hou, C. Gao, Z. Cai, J. Luo, and L. Shao. D (r, o) grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping. *arXiv preprint arXiv:2410.01702*, 2024.

[87] J. Wen, Y. Zhu, J. Li, Z. Tang, C. Shen, and F. Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*, 2025.

[88] J. Gu, S. Kirmani, P. Wohlhart, Y. Lu, M. G. Arenas, K. Rao, W. Yu, C. Fu, K. Gopalakrishnan, Z. Xu, et al. Rt-trajectory: Robotic task generalization via hindsight trajectory sketches. *arXiv preprint arXiv:2311.01977*, 2023.

[89] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

[90] R. Mendonca, S. Bahl, and D. Pathak. Structured world models from human videos. *arXiv preprint arXiv:2308.10901*, 2023.

[91] S. Nasiriany, S. Kirmani, T. Ding, L. Smith, Y. Zhu, D. Driess, D. Sadigh, and T. Xiao. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. *arXiv preprint arXiv:2411.02704*, 2024.

[92] Y. Chen, Z. Chen, J. Yin, J. Huo, P. Tian, J. Shi, and Y. Gao. Gravmad: Grounded spatial value maps guided action diffusion for generalized 3d manipulation. *arXiv preprint arXiv:2409.20154*, 2024.

[93] F. Yan, F. Liu, L. Zheng, Y. Zhong, Y. Huang, Z. Guan, C. Feng, and L. Ma. Robomm: All-in-one multimodal large model for robotic manipulation. *arXiv preprint arXiv:2412.07215*, 2024.

[94] C. Gao, H. Gao, S. Guo, T. Zhang, and F. Chen. Cril: Continual robot imitation learning via generative and prediction model. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6747–5754. IEEE, 2021.

[95] Y. Zhao, A. Gu, R. Varma, L. Luo, C.-C. Huang, M. Xu, L. Wright, H. Shojanazeri, M. Ott, S. Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

[96] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020.

[97] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[98] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[99] H. Yang, L. Duan, Y. Chen, and H. Li. Bsq: Exploring bit-level sparsity for mixed-precision neural network quantization. *arXiv preprint arXiv:2102.10462*, 2021.

[100] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, and Z. Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[101] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[102] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[103] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023.

[104] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

A Related Works

A.1 VLA Paradigms for Robot Manipulation

Vision-Language-Action Models (VLAs) refer to multi-modal comprehensive models that can handle visual and language inputs and generate robot actions for control. The word “VLA" was first proposed and studied in RT-2 [55], where they train a VLM to output actions as text tokens for robot control. After that, more VLA works are emerging. According to how they incorporate the task planning process, we divide VLAs into four paradigms and introduce each of them as follows.

PlanningOnly-VLA These works leverage pretrained LLMs or VLMs to reason and perform task planning without generating the low-level action. They break up the given task into simpler sub-tasks that can be performed by either using a set of pre-trained sub-skills [56, 7, 57, 58, 59], or outputting the parameters of pre-defined motions or cost functions for optimization [60, 61, 62, 63, 64, 65, 66, 67]. The problem is that their VLMs and low-level skills usually cannot be trained with further datasets, which frequently places them at a disadvantage compared to other VLA paradigms capable of training on given datasets [68, 69]. Consequently, we do not include PlanningOnly-VLA in this study.

ActionOnly-VLA These works employ an end-to-end fashion to directly map visual and language inputs to robot actions with a multi-modal network. Pioneering works mainly focus on verifying the effectiveness of large-scale robot learning [70, 55, 71, 72], while later works start to explore different model architectures, training objectives, and extra multi-modal representations and information fusion designs to make this paradigm more effective and efficient [3, 6, 5, 73, 74, 75, 51, 76, 77, 36, 78, 79]. In this work, we design VLA-OS-A for this paradigm by synthesizing several advanced model designs that have been verified to be superior in recent works [37, 3, 36].

Integrated-VLA These works use a single model to perform task planning and policy learning simultaneously. According to whether the action generation process is conditioned on the planning embeddings or results, they can be further divided into explicit planning and implicit planning. For explicit planning, EmbodiedCoT [18] and CotVLA [17] generate either language-based or goal-image-based embodied chain-of-thought [47] reasoning before generating actions, and the action generation process is conditioned on the embeddings of CoT. For implicit planning, MDT [80] and PIDM [23] use goal image foresight generation loss as an auxiliary objective for planning, while RoboBrain [26] and ChatVLA [25] train VLA with auxiliary task reasoning loss in language representations. Some recent works also seek to use latent action tokens [81, 82, 52, 83, 84] that serve as forward dynamics representations to generate future images as image foresight planning, and decode these latent actions to real actions with another action head. The inputs to the action head are from the VLM encoder, and they do not need the planning heads (decoder) during inference [81, 82, 52, 83] or they only need one planning forward pass [84], so we also see these methods as implicit planning. In this work, we design VLA-OS-I for this paradigm with various plug-and-play planning heads upon VLA-OS-A for different planning representations, and design corresponding variants for both explicit and implicit planning paradigms as VLA-OS-I-I and VLA-OS-I-E.

Hierarchical-VLA These works use two separate models for task planning and policy learning, with no connection or gradient between them. The idea of hierarchical models has always existed in robotics research [66, 67, 27, 85, 86]. RT-H [22] is the first work of this paradigm, where they use two identical VLMs to generate languages and actions respectively. Later works [20, 87, 4] also follow this idea but use different model architectures for task planning and action generation. Other works seek to generate multi-modal planning results for policy learning, such as image flows or trajectories [88, 19, 21], future videos [89, 54], affordance [90, 91], keypose [92], and keypoints [53]. In this work, we design VLA-OS-H for this paradigm.

A.2 VLA Benchmarks and Evaluations

With the rapid advancement of VLA models, benchmarks and evaluation studies for VLA have also experienced significant growth. Given the complexity and multi-dimentionality of robot manipulation tasks and VLA models, different works usually focus on evaluating one or several specific dimensions of VLA. Some works focus on the VLA model designs and training algorithms, such as different

model architectures and input and output spaces [37, 93]. Other works aim to build benchmark environments and tasks to evaluate different capacities of current VLA models, such as spatial and visual generalization ability [68], long-horizon task reasoning ability [69], and different training data modalities [25]. In this work, we focus on task planning paradigms for VLA and keep the model architectures the same with systematically designed controllable experiments.

B More Experiments

B.1 Separate Investigation of Task Planning and Policy Learning Parts

It is imperative to discern whether task failures arise from the planning component or policy learning. In this part, we use LIBERO-LONG [28] for Integrated-VLA (only for task planning) and Hierarchical-VLA to separately evaluate the task planning part and policy learning part of the model for all three planning representations. For evaluation, we manually divide each long-horizon task into several sub-tasks, and forcibly reset the environment to the initial state of each subtask. Then we compute the average planning correctness $\mathbb{I}$ (0 or 1) of the planning outcomes and execution success rate $\mathbb{S}$ (0 or 1) from the action head across all subtask start points. Thus, for a given task trajectory, we can get **Decomposition Score (DCS)**$= \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}(p_t)$ and **Instruction Following Score (IFS)**$= \frac{1}{T}\sum_{t=1}^{T}\mathbb{S}(a_t^{seq})$, where T is the total sub-task number and p_t and a_t^{seq} are the planning outcomes and actions generated at subtask t. Note for Hierarchical-VLA, we give the ground truth planning results when testing IFS. Results are shown in Table 3.

Table 3: Separate evaluation of task planning and policy learning modules for different paradigms and representations. Results are averaged from 20 episodes for each task in LIBERO-LONG.

	L		V		IF	
	DCS	IFS	DCS	IFS	DCS	IFS
VLA-OS-I-I	0.79	–	0.83	–	0.92	–
VLA-OS-H	0.81	0.84	0.86	0.93	0.94	0.90

Finding 8: *Hierarchical-VLA performs better than Integrated-VLA in task planning.*

Finding 9: *Visually grounded planning representations are easier for low-level policy to follow.*

B.2 Training Cost and Inference Speed

We report the inference speed and training cost for different paradigms and planning representations. The training cost is calculated by multiplying the total training steps by the per-step time on LIBERO-LONG with 8× A100 NVIDIA GPUs. Results are shown in Figure 1 and 6b.

Finding 10: *The autoregressive property of the language-planning representation head is the principal cause of its higher training cost and slower inference speeds.*

B.3 Data and Model Scalability Experiments

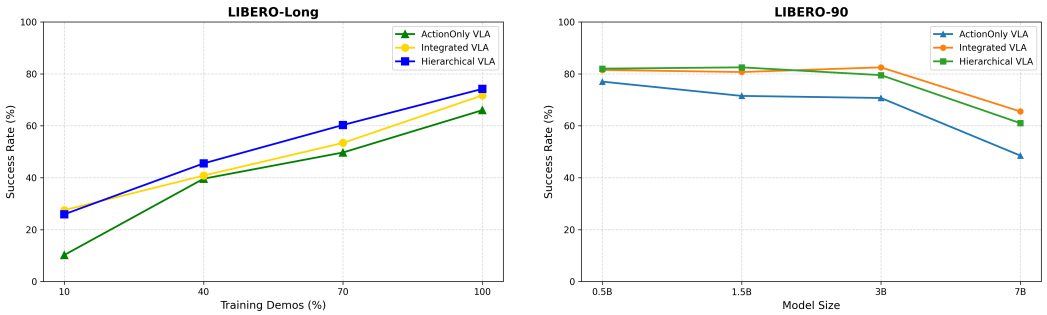
In this part, we perform experiments for data and model scalability of different VLA paradigms. For data scalability, we use LIBERO-LONG [28], a dataset with 10 tasks with a total of 500 demonstrations. We use 10%, 40%, 70%, and 100% of the data to train on three VLA paradigms with the model size S. For model scalability, we use LIBERO-90, a dataset with 90 tasks and 4,500 demonstrations, for the experiment with all training data. We choose Qwen-2.5 LLM backbone with parameters of 0.5B, 1.5B, 3B, and 7B for experiments. Results are shown in Figure 7.

Finding 11: *The performance of all VLA paradigms improves as the amount of action-labeled demonstration data increases, i.e., all VLA paradigms have the data scalability.*

Finding 12: *For tasks trained from scratch with roughly 5,000 demonstrations, the LLM backbone should be limited to 0.5B parameters, or keeping the total model size under 1B parameters.*

B.4 Continual Learning Experiments of Different VLA Paradigms

Continual learning for robot imitation learning [28, 94] measures the degree to which the VLA model forgets old tasks when continuously learning new tasks. In this part, we test the continual learning capacities of three paradigms and three representations on 10 tasks of LIBERO-LONG sequentially. We only use Sequential Finetuning (SEQL) as our lifelong learning algorithm. We use the original metrics from LIBERO [28], including forward transfer (FWT), negative backward transfer (NBT), and area under the success rate curve (AUC). Denote $c_{i,j,e}$ as the agent's success rate on task j when it learned over $i-1$ previous tasks and has just learned e epochs ($e \in 0, 2, \cdots, 20$) on task i. Let $c_{i,i}$ be the best success rate over all evaluated epochs e for the current task i (i.e., $c_{i,i} = \max_e c_{i,i,e}$).



(a) Data scalability experiments on LIBERO-LONG with different planning paradigms with 0.5B LLM backbone. Success rates are calculated among 20 evaluation episodes among the 3 best checkpoints.

(b) Model scalability experiments on LIBERO-90 of all VLA paradigms. Success rates are calculated among 20 evaluation episodes among the 3 best checkpoints. We select the best checkpoint before 100k steps.

Figure 7: Data and model scalability experiments across different VLA paradigms.

Then, we find the earliest epoch e_i^* in which the agent achieves the best performance on task i (i.e., $e_i^* = \arg\min_e c_{i,i,e_i} = ci, i$), and assume for all $e \geq e_i^*, c_{i,i,e} = ci, i$. Given a different task $j \neq i$, we define $c_{i,j} = ci, j, e_i^*$. Then the three metrics are defined as follows:

$$\begin{aligned}
\text{FWT} &= \sum_{k\in[K]} \frac{\text{FWT}_k}{K}, \quad \text{FWT}_k = \frac{1}{11} \sum_{e\in\{0\ldots50\}} c_{k,k,e}, \\
\text{NBT} &= \sum_{k\in[K]} \frac{\text{NBT}_k}{K}, \quad \text{NBT}_k = \frac{1}{K-k} \sum_{\tau=k+1}^{K} \left(c_{k,k} - c_{\tau,k}\right), \\
\text{AUC} &= \sum_{k\in[K]} \frac{\text{AUC}_k}{K}, \quad \text{AUC}_k = \frac{1}{K-k+1} \left(\text{FWT}_k + \sum_{\tau=k+1}^{K} c_{\tau,k}\right).
\end{aligned} \tag{1}$$

Results are shown in Table 4 and Table 5.

Table 4: Continual learning results on LIBERO-LONG of three different VLA paradigms. The VLA-OS-I and VLA-OS-H are trained with three planning representations together.

	FWT(↑)	NBT(↓)	AUC(↑)
VLA-OS-A	0.71	**0.32**	0.25
VLA-OS-I	0.75	0.43	0.29
VLA-OS-H	**0.80**	0.45	**0.32**

Table 5: Continual learning results on LIBERO-LONG of three different planning representations (Language (L), Visual (V), and Image Foresight (IF)) on VLA-OS-I.

	FWT(↑)	NBT(↓)	AUC(↑)
L	0.72	0.47	0.26
V	0.74	0.40	**0.28**
IF	**0.75**	**0.39**	0.27

Finding 13: VLA paradigms with task planning (Integrated-VLA and Hierarchical-VLA), compared to the non-planning paradigm (ActionOnly-VLA), achieve higher forward transfer but incur faster forgetting.

Finding 14: Visually grounded planning representations deliver superior forward transfer and exhibit slower forgetting relative to language-based planning representations.

C Benchmarks and Dataset Details

C.1 VLM Pretraining Dataset

The LLM backbones we choose are Qwen-2.5 [16] series. Since they are not VLM, we first pretrain it to VLM with LLaVa v1.5 [41] data mixture, which consists of two subsets used for a multi-stage training pipeline. The first subset consists of a 558K sample mixture of examples sourced from various captioning datasets, while the second consists of 665K multimodal instruct tuning examples comprised of synthetic data generated in [41], as well as examples from existing vision-language training sets. According to the conclusion from Prismatic-VLMs [39], we only use the first subset to train the VLM in a single-stage optimization procedure, that is, directly fine-tuning all parameters. We implement the training code with PyTorch using Fully Sharded Data Parallel (FSDP [95]) and BF16 mixed precision and train the VLM with 2 epochs for all Qwen2.5 model types (0.5B, 1.5B, 3B, and 7B). The training hyperparameters are shown in Table 6.

Table 6: Training hyperparameters of VLM for Qwen2.5 LLM.

Hyperparameter	Value
Batch Size	64
Max Gradient Norm	1.0
Weight Decay	0.1
Learning Rate	2e-5
Optimizer	AdamW
Scheduler	Warmup & Cosine Decay
Warmup Ratio	0.03

C.2 LIBERO Dataset

The LIBERO Dataset [28] contains five subsets: LIBERO-Spatial, LIBERO-Object, LIBERO-GOAL, LIBERO-LONG, and LIBERO-90. The first four subsets contain 10 tasks for each of them, with 50 demonstrations for each task. The last subset contains 90 tasks with also 50 demonstrations for each task. All tasks have a language instruction that describes the task. We use the two camera views for all subsets (the agentview and eye-in-hand view). We use a resolution of 224×224 for each view of the image. The action space is 7-dim, containing 6-dim $\delta x, \delta y, \delta z, \delta roll, \delta pitch, \delta yaw$ and 1-dim gripper open/close. We use a history length of 2 and a future action length of 8.

Following OpenVLA [5], we further clean up the original LIBERO datasets by:

- We filter out all "no-op" actions from the dataset, i.e., actions that have near-zero magnitude in the translation and rotation components and do not change the state of the robot's gripper.
- We replay all demonstrations in the corresponding simulation environments and filter out the demonstrations that fail to complete the task (as determined by the environments' success criteria).

C.3 The COLOSSEUM Dataset

For 3D manipulation tasks and generalization experiments, we use The Colosseum [29] as our task benchmark. This benchmark contains 20 single-arm manipulation tasks in simulation. Each task has various variants such as lighting, distractors, physical properties perturbations, and camera pose. The cameras in this benchmark are depth cameras, so we can get the depth map and then get the point cloud observations by fusing all cameras. We follow 3D-DA [42] to preprocess the 3d observations to point cloud tokens. Then we send the point cloud tokens to the action head (or the low-level action head) as additional inputs, together with the original multi-view images. This makes the action heads have the 3D-aware property. For each task, we have 100 demonstrations. The action space is 8-dim,

containing 3-dim $\delta x, \delta y, \delta z$ and 4-dim $\delta w, \delta q_x, \delta q_y, \delta q_z$ as the delta quaternion for rotation, and 1-dim gripper open/close. We use a history length of 2 and action length of 8.

C.4 The Real-World Deformable Manipulation Dataset

For deformable object manipulation tasks, we design three real-world deformable object manipulation tasks: unfold the jeans, fold the handkerchief, and straighten the rope, as shown in Figure 4. We use two camera views for these tasks, where a third-view camera is mounted on another X-Arm, and an eye-in-hand-view camera is mounted on the main X-Arm, as shown in Figure 8. We collect 100 demonstrations for each task with human teleoperation. The cameras we use are RealSense D435i. We freeze the rotation of the X-Arm, so the action space is 4-dim: 3-dim $\delta x, \delta y, \delta z$ and 1-dim gripper open/close. The average horizon of these tasks is 214 steps. We also use an observation history length of 2 and a future action length of 8.

(a) The jeans unfold task. (b) The handkerchief folding task. (c) The rope straightening task.

Figure 8: The real-world deformable object manipulation tasks.

C.5 The DexArt Dataset

We use the DexArt [31] benchmark for dexterous manipulation tasks. This benchmark contains four dexterous manipulation tasks built on the Sapien [96] simulator, including *turn on the faucet*, *open the laptop*, *lift the bucket*, and *open the toilet*. The original benchmark is a reinforcement learning benchmark, and they provide the official trained policy checkpoint. We load these checkpoints and collect 100 demonstrations for each task. We use one camera view for each task.

C.6 The FurnitureBench Dataset

For long-horizon complex manipulation tasks, we choose FurnitureBench [30] as our task benchmark. This benchmark provides corresponding simulation environments called FurnitureSim, and it provides demonstrations for four tasks: *cabinet*, *lamp*, *one-leg*, and *round-table*. Each task has 100 demonstrations. The action space is 8-dim, containing 3-dim $\delta x, \delta y, \delta z$ and 4-dim $\delta w, \delta q_x, \delta q_y, \delta q_z$ as the delta quaternion for rotation, and 1-dim gripper open/close. We use three camera views as input.

C.7 The PerAct2 Dataset

For dual-arm manipulation tasks, we choose PerAct2 [32] as our task benchmark. We use five tasks in this benchmark: *handover item*, *lift ball*, *put bottle in fridge*, *straighten rope*, and *sweep to dustpan*. As in The Colosseum, we make this benchmark a 3D task benchmark. Each task has 100 demonstrations. The action space is 22-dim, where 16-dim is for the dexterous hand joint values and 6-dim is for the end-effector. For this dataset, we do not use the image foresight planning.

C.8 The Real-World Rigid-Body Manipulation Dataset

To further verify our conclusions in the real-world setting, we design 5 manipulation tasks in a single-arm manipulation setting, as shown in Figure 9. We collect 50 demonstrations for each task. The action space is 7-dim.

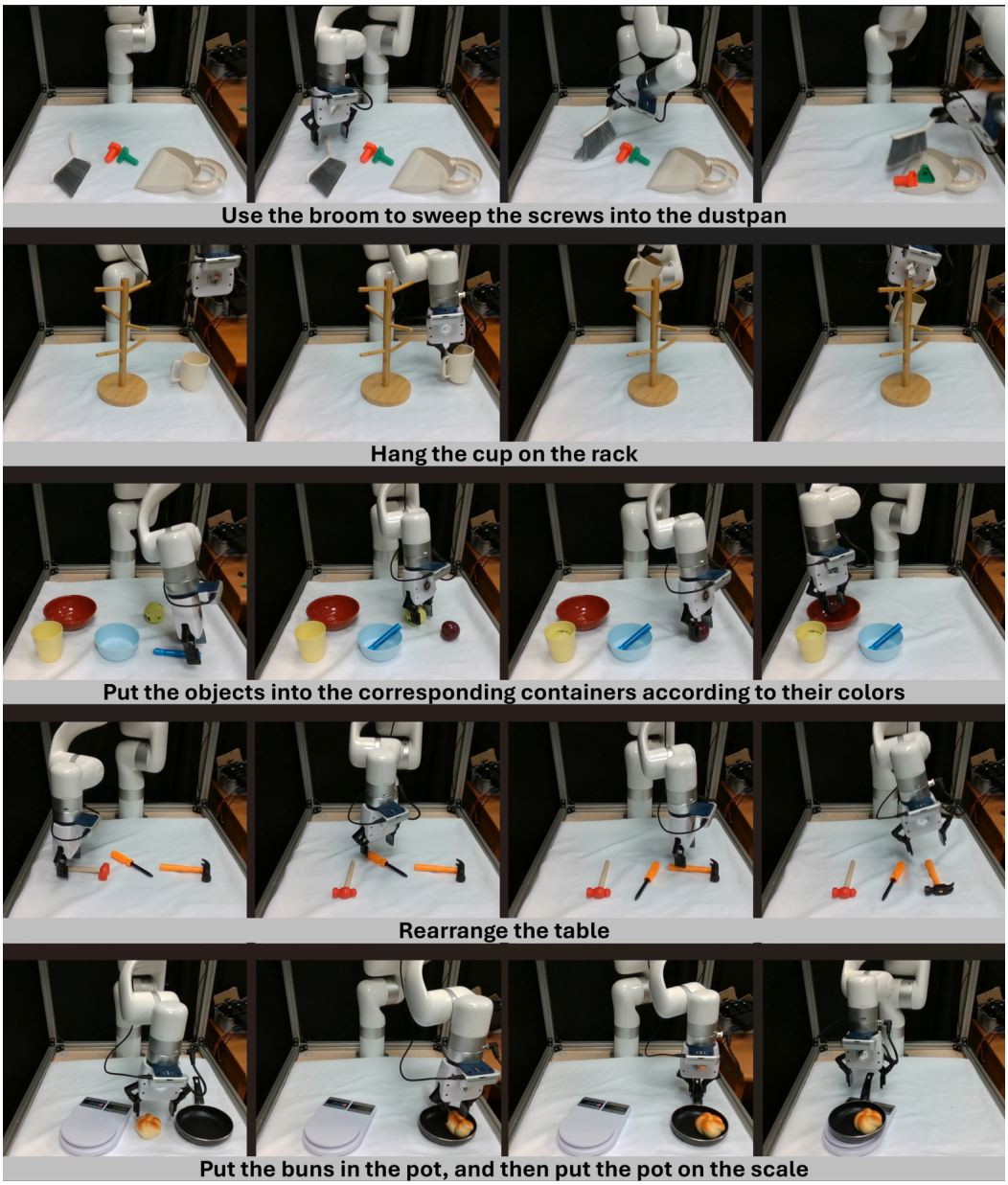


Figure 9: Real world manipulation tasks.

D Reasoning Dataset Annotation

D.1 Language Reasoning Dataset

This dataset contains language-based planning results for the task that understands the scene and decomposes the task, as used in [20, 69, 25, 22]. We design a unified language planning format and structure applicable to all manipulation tasks with 8 different keys, including `Task`, `Plan`, `Subtask`, `Subtask Reason`, `Move`, `Move Reason`, `Gripper Position`, and `Object Bounding Boxes`. For example, for the task *open the top drawer of the cabinet*, the reasoning data should be:

```
TASK: Open the top drawer of the cabinet. PLAN: 1. Approach the
cabinet. 2. Locate the top drawer. 3. Locate and grasp the drawer
handle. 4. Open the drawer. 5. Stop. VISIBLE OBJECTS: akita black bowl
 [100, 129, 133, 155], plate [17, 131, 56, 158], wooden cabinet [164,
75, 224, 175] SUBTASK REASONING: The top drawer has been located; the
robot now needs to position itself to grasp the handle. SUBTASK:
Locate and grasp the drawer handle. MOVE REASONING: Moving left aligns
 the robot's end effector with the drawer handle. MOVE: move left
GRIPPER POSITION: [167, 102, 166, 102, 165, 102, 164, 102, 162, 102,
161, 102, 160, 102, 158, 102, 156, 102, 154, 102, 153, 102, 151, 102,
149, 102, 147, 102, 145, 102, 143, 102]
```

Similarly to EmbodiedCoT [18], we provide an overview of our data labeling pipeline in Figure 3. To obtain a comprehensive understanding of the scene, we first query the Prismatic-7B VLM [39], which outputs a detailed scene description. Next, we derive low-level motion primitives by analyzing the robot's proprioceptive state across a 10-step prediction horizon, assuming a static camera viewpoint, and translating these movement traces into a set of pre-defined action templates (e.g., "move left", "move up"). To construct the full reasoning trace, we use Gemini1.5 [97] to synthesize higher-level plans. Given the task instruction, scene description, and step-wise movements, Gemini1.5 generates a structured plan that includes a sequence of sub-tasks, as well as the specific sub-task relevant to each step. Additionally, it provides concise justifications for both the movement taken and the associated sub-task.

However, during experiments, we observed that the quality of the generated reasoning, referred to as *initial reasoning* in Figure 3, was often suboptimal, exhibiting two major issues. First, there was inconsistency in the planning outputs: even for the same task, the language descriptions of sub-tasks varied significantly. This stems primarily from the inherent randomness in responses from large language models such as Gemini1.5. Second, we found a mismatch between the generated plans and the actual trajectories. This issue was particularly pronounced in complex, long-horizon tasks (e.g., FurnitureBench [30]), where the provided inputs—task instruction, scene description, and step-wise movement primitives—were insufficient for the model to infer coherent and accurate planning steps. As a result, the low quality of the *initial reasoning* posed challenges for training the planning head, as the model struggled to learn meaningful mappings from observations to such plannings.

To address these issues, we applied a filtering and refinement process to the *initial reasoning*. Specifically, for each task, Gemini or human experts selected and edited the task descriptions and high-level plans produced by Gemini to ensure consistency across episodes of the same task. Once a canonical task and plan were established, we prompted Gemini again to regenerate the step-wise reasoning under this fixed structure. This process yielded the *final reasoning* in Figure 3, which aligns better with the trajectories and provides more coherent supervision for training the planning head.

In addition to the reasoning generated by Gemini, we also incorporate object bounding boxes and gripper positions into the final annotations. For real-world data, we adopt a labeling strategy similar to EmbodiedCoT [18], leveraging vision-language models to annotate object locations from visual inputs. For simulation data, we exploit the availability of camera intrinsics and extrinsics to project 3D gripper positions into 2D image coordinates. Object bounding boxes can also be directly extracted using simulator-provided segmentation masks, enabling efficient and accurate annotation of the visual scene.

Finally, we represent language planning in the following format:

- Task: A concise natural language description of the goal the robot needs to achieve.
- Plan: A high-level sequence of steps to accomplish the task, typically numbered and described in imperative language.
- Subtask: A mid-level action derived from the plan, typically one step at a time, to be executed next.
- Subtask Reason: A rationale explaining why the current subtask is necessary or meaningful in context.
- Move: A specific low-level movement command to guide the robot toward completing the subtask.
- Move Reason: A justification of the chosen movement, often grounded in spatial alignment or task constraints.
- Gripper Position: A list of 2D coordinates that define the intended trajectory or position of the robot's gripper in image space. This often reflects the gripper's pixel-level alignment with the target object.
- Object Bounding Boxes: A list of objects currently detected in the scene, each annotated with a bounding box in pixel coordinates $[x_1, y_1, x_2, y_2]$

D.2 Visual Reasoning Dataset

This dataset will generate visual representations in the language format for task planning. Compared to pure language-based representations, these visual representations have better spatial semantic information and are more grounded in the input images, which are used in recent multi-modal learning works [43, 44]. In this work, we use three keys, including `object bounding boxes`, `end-effector flow`, and `target object affordance` as the visual planning representations.

As shown in Figure 3, we use discrete location tokens on the input image to represent visual planning results. For an image with width W and height H, we evenly divide both the width and height into P segments each, thus we use $P \times P$ discrete bins to represent the visual pictures and each bin consists of $(W/P) \times (H/P)$ pixels. We use a new location token `<loc i>` to represent the i-th bin token from top-left to bottom-right, and increase the tokenizer's word vocabulary to add these bin tokens. For bounding boxes, we use the top-left and bottom-right bins to represent them. For end-effector flows, we use a sequence of bins to represent them. For affordances, we use a region of bins to represent the target regions. In this work, $W = H = 224$, and $P = 32$, i.e., each bin consists of 7×7 pixels. For example, for the task *Put the cream cheese box and the butter in the basket*, the visual reasoning data should be:

```
VISUAL OBJECT BBOXES: alphabet soup <loc_500, loc_632>, cream cheese <
loc_353, loc_452], tomato sauce <loc_461, loc_624>, ketchup <loc_341,
loc_503>, orange juice <loc_538, loc_767>, milk <loc_563, loc_791>,
butter <loc_684, loc_783>, basket <loc_448, loc_775>. VISUAL EE FLOW:
loc_387, loc_387, loc_387, loc_419, loc_419, loc_419, loc_419, loc_419
, loc_419, loc_419, loc_419, loc_451, loc_451, loc_451, loc_451,
loc_451>. VISUAL AFFORDANCE: loc_354, loc_355, loc_356, loc_386,
loc_387, loc_388, loc_418, loc_419, loc_420>
```

Specifically, given a manipulation task $\mathcal{T}$ consisting of N steps $(i.e. 1, 2, ..., N)$, we take the following steps to generate visual-based planning representations $\{\mathcal{V}_i^{box}, \mathcal{V}_i^{flow}, \mathcal{V}_i^{afford}\}_{i=1}^N$:

1. **Object Bounding Boxes:** We first get instance semantic maps $S = \{S_i\}_{i=1}^N \in \mathcal{R}^{N \times H \times W}$ from the simulation engine to compute binary masks for each object in each frame. Next, we sequentially apply $cv2.morphologyEx()$ to reduce noise and reconnect fragmented regions, $cv2.findContours()$ to detect object contours, and $cv2.boundingRect()$ to compute the rectangular bounding box for each detected object. Finally, we annotate the location token of

the top-left and bottom-right bins for each bounding box. The final bounding box visual annotation for task $\mathcal{T}$ can be formulated as $\left\{\mathcal{V}_i^{box}\right\}_{i=1}^{N}$, where $\mathcal{V}_i^{box} = \left\{(loc_j^{tl}, loc_j^{br})\right\}_{j=1}^{m_i}$.

2. **End-effector Flow:** The end-effector flow visual annotation is obtained by directly labeling the location tokens corresponding to the gripper positions in the language-based planning representation. Formally, the end-effector flow annotation for task $\mathcal{T}$ can be formulated as $\left\{\mathcal{V}_i^{flow}\right\}_{i=1}^{N}$, where $\mathcal{V}_i^{flow} = loc_i^{gripper}$.
3. **Object Affordance:** The object affordance is represented as a heatmap centered on the target object to be fetched. We first identify the target object by detecting changes in all bounding boxes (e.g. shifts in location or variations in size). Next, we employ the pretrained SAM2 [98] model to infer a precise object mask within the target bounding box. Finally, we compute a Gaussian heatmap centered at the gripper position within the object mask to model the affordance. Location tokens corresponding to regions with affordance values exceeding a predefined threshold are then annotated in a top-left to bottom-right order. The final object affordance annotation for task $\mathcal{T}$ can be formulated as $\left\{\mathcal{V}_i^{afford}\right\}_{i=1}^{N}$, where $\mathcal{V}_i^{afford} = \{loc_j\}_{j=1}^{n_i}$.

D.3 Image Foresight Reasoning

Image Foresight (IF) reasoning dataset will imagine a future goal frame as the most general representation for task planning. There is no special effort here to label the goal image. We just select the future image from the trajectory.

Here we want to introduce more about the image generation head. In this work, we use an image generation head for planning based on [45]. It auto-regressively generates the image in a coarse-to-fine paradigm proposed by [46]. Given an input image, it iteratively quantizes the residual image following a coarse-to-fine resolution schedule $\{(h_k, w_k)\}_{k=1}^{K}$. It also applies a technique called **B**itwise **S**elf-**C**orrection (BSC) to mitigate the performance gap between training and testing caused by teacher-forcing training.

Formally, inside each quantization iteration k, the tokenizer does the following steps:

1. **Calculate and Quantize Residual:** It computes the difference between the original raw feature $\boldsymbol{F}$ and the reconstructed flipped feature from the previous iteration ($\boldsymbol{F}_{k-1}^{flip}$). This residual is then interpolated to the current resolution (h_k, w_k) and quantized following [99] to produce tokens at the current resolution $\boldsymbol{R}_k = \text{quantize}(\text{down}(\boldsymbol{F} - \boldsymbol{F}_{k-1}^{flip}, (h_k, w_k)))$.
2. **Apply Random Flipping For BSC:** A random flipping operation ($\text{Random\_Flip}(\cdot)$) is applied to the quantized residual $\boldsymbol{R}_k$ based on a probability p. This results in the flipped residual $\boldsymbol{R}_k^{flip} = \text{Random\_Flip}(\boldsymbol{R}_k, p)$.
3. **Reconstruct Flipped Feature:** The algorithm reconstructs the cumulative flipped feature $\boldsymbol{F}_k^{flip}$ up to the current iteration. It does this by interpolating *all* previously generated flipped residuals ($\boldsymbol{R}_i^{flip}$ for i from 1 to k) to the original image resolution (h, w) and sums them together: $\boldsymbol{F}_k^{flip} = \sum_{i=1}^{k} \text{up}(\boldsymbol{R}_i^{flip}, (h, w))$.

During inference, generation starts from a global conditioning signal, for example, the text embedding in a T2I generation setting. Notably, it generates *all* tokens of a resolution at once, distinguishing this method from the raster-scanning generation paradigm.

We select [45] as our image generation head based on three primary advantages. Firstly, it surpasses the state-of-the-art diffusion-based models [100, 101, 102] in performance on academic benchmarks and in human preference evaluations. Secondly, [45] achieves lower inference latency compared to prevalent diffusion models, a critical requirement for embodied planning within the hierarchical VLA framework. Third, our experiments indicate that the training loss of [45] serves as a stronger predictor of the final quality of foresight image generation while necessitating fewer hyperparameter

adjustments, such as the noise scheduling required by the diffusion models. In practice, when the loss drops below 0.1, it indicates that the training is complete.

E VLA-OS Model Details

E.1 Action Head Details

The action head for all VLA-OS models is in the same architecture, with only different numbers of layers. It is a block-wise causal attention transformer with the same number of layers as the LLM backbone, as introduced in Section 2.2. Let $[KV_1, \cdots, KV_n]$ be the KV tokens from the LLM, $[t]$ be the denoising timestep embedding token, $[q]$ be the proprioceptive token, and $[a_t, \cdots, a_{t+H-1}]$ be the action token, sequentially. The tokens in each block can attend to itself and blocks before it, but cannot attend to blocks after it. The hyperparameters of the action head are shown in Table 7.

Table 7: Hyperparameter of the action head transformer.

	Layer Number	Hidden Size	Dropout	Head	Non-Linear Func
Action Head S	24	512	0.1	8	GELU
Action Head S	28	512	0.1	8	GELU
Action Head S	36	512	0.1	8	GELU
Action Head S	28	512	0.1	8	GELU

The low-level action head used for VLA-OS-H is also a transformer. It has a separate convolutional neural network (CNN) for encoding the input images, the visual planning images, and the image foresight image. For other parts, it keeps the same setting as the normal action head.

E.2 Planning Head Details

All three planning head transformers share the same network structure (the VAE encoder and decoder of the image foresight planning head are frozen). The planning head takes as input the keys and values from each layer of the LLM backbone. The hyperparameters of the planning head are shown in Table 8.

Table 8: Hyperparameter of the planning head transformer.

	Layer Number	Hidden Size	Dropout	Head	Non-Linear Func
Action Head S	24	512	0.1	8	GELU
Action Head S	28	512	0.1	8	GELU
Action Head S	36	512	0.1	8	GELU
Action Head S	28	512	0.1	8	GELU

E.3 Training Loss Details

The action heads can be trained with either L1 behavior cloning loss, or the flow matching loss. L1 loss is shown to be better than L2 MSE loss for VLA [5, 35]. The L1 loss is:

$$\mathcal{L}_{L1}(\theta) = \mathbb{E}_{s,a\in\mathcal{D}}\big|\pi_\theta(s) - a\big|. \tag{2}$$

The flow matching loss is:

$$\mathcal{L}_{FM} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I),t\sim U(0,1),s,a\sim\mathcal{D}}||\pi_\theta(x_t, t|s) - u_t||_2^2, \tag{3}$$

where $x_t = (1-t)\epsilon + ta$ and $u_t = \frac{d}{dt}x_t = a - \epsilon$.

The planning head losses for the language planning head and the visual planning head are the standard next-token prediction loss. The loss for the image foresight planning head follows the original paper [45].

E.4 Training Details of Hierarchical-VLAs

The training process of Hierarchical-VLAs can have multiple choices, since they inherently incorporate two models that are not connected by backward gradients. In this work, we aim to reuse the trained model to the greatest extent possible to reduce the cost of repeated training. Thus, for Hierarchical-VLAs, we first borrow the trained VLMs backbone as well as the planning heads from Integrated-VLAs, and finetune them on the planning datasets to get the high-level models for the Hierarchical-VLAs. Later, for the low-level model, we extract the Keys and Values from the high-level LLM backbone and use them as the embedding of the input visual and language signals and send them to each layer of the low-level action head. For the planning outputs from the high-level planning heads, we use different models to encode them: we use a frozen Qwen-2.5 7B [16] model to encode the language planning outputs to get the sentence embeddings, a common Convolutional Neural Network (with 6-channel inputs of the current 3-channel image and a 3-channel visual planning results) to encode the visual planning outputs, and a common Convolutional Neural Network (with 3-channel inputs of the goal image) to encode the image foresight planning outputs. The gradient of the low-level action head will not go backward through the high-level VLM backbone.

F Limitations

The limitations of this paper are: 1) despite the VLA-OS family encompassing a wide array of task planning paradigms for VLA, there remain several designs and variants that we have not yet covered, such as using latent actions [81, 52] for image generation rather than VAR [46, 45]-like generator in VLA-OS, video generation for planning [54, 89], and scene flow for planning [19, 103]; 2) we didn't explore embodiment transfer, sim2real transfer, and 2D to 3D transfer problems for VLA; 3) our training dataset remains limited to fewer than 10,000 trajectories, and we have not yet investigated the research questions that arise from pretraining on larger datasets such as the OXE [104] dataset.