A Systematic Study of Data Modalities and Strategies for Co-training Large Behavior Models for Robot Manipulation

Fanqi Lin[†‡], Kushal Arora[†], Jean Mercat[†], Haruki Nishimura[†], Paarth Shah[†], Chen Xu[†], Mengchao Zhang[†], Mark Zolotas[†], Maya Angeles[†], Owen Pfannenstiehl[†], Andrew Beaulieu[†], Jose Barreiros[†]

[†]Toyota Research Institute, Cambridge MA and Los Altos CA, USA
[‡]Tsinghua University, Beijing, China

***Abstract*—Large behavior models (LBMs) have shown strong dexterous manipulation capabilities by extending imitation learning to large-scale training on extensive multi-task robot data, yet their generalization remains limited by the insufficient coverage of available robot data. To expand this coverage without costly additional data collection, recent work increasingly relies on *co-training*: jointly learning from target robot data and heterogeneous data modalities. However, how different co-training data modalities and training strategies affect policy performance remains poorly understood. We present a large-scale empirical study examining five co-training data modalities—standard vision-language data, dense language annotations for robot trajectories, cross-embodiment robot data, human videos, and discrete robot action tokens—across single- and multi-phase training strategies. Our study leverages 4,000 hours of robot and human manipulation data and 50M vision–language samples to train vision-language-action (VLA) policies. We evaluate 89 policies over 58,000 simulation rollouts and 2,835 real-world rollouts. Our results show that co-training with various forms of vision-language and cross-embodiment robot data substantially improves generalization to distribution shifts, unseen tasks, and language following, while discrete action token variants yield no statistically significant benefits. Furthermore, combining effective modalities produces cumulative gains and enables rapid adaptation to unseen long-horizon dexterous tasks via fine-tuning. Together, these results provide a systematic understanding of co-training and practical guidance for building scalable generalist robot policies.**

Project page:https://co-training-lbm.github.io/

I. Introduction

Robot learning is increasingly moving toward generalist models that can perceive, understand, and act in the physical world. Recent efforts have focused on training large behavior models (LBMs) [1]—embodied foundation models trained on large-scale multi-task robot datasets—to produce dexterous manipulation policies. Within this family, vision-language-action models (VLAs) [37, 100, 21, 6, 33, 67, 5] are a representative subclass of LBMs that integrate visual and linguistic inputs for action generation. Despite progress, LBMs still lag behind non-embodied foundation models, such as vision-language models (VLMs) [32, 16, 65, 73], in semantic and spatial understanding and in open-world generalization. This limitation can be attributed to the significant disparity in data scale [25]: robot data is orders of magnitude smaller than the internet-scale text and image corpora used to train VLMs.

To bridge this data gap, many recent works [33, 44, 5, 39, 13, 61, 84] use *co-training*—jointly learning target robot (i.e., the deployment embodiment) data alongside heterogeneous data modalities: standard vision-language data [33, 99, 98, 39], dense language annotations for robot trajectories [92, 44, 93, 84, 61, 11], cross-embodiment robot data [1, 50, 48, 69, 37, 38], human videos [10, 13, 88, 52, 90, 36], and discrete robot action tokens [93, 22, 35, 33]. However, current studies typically evaluate only a subset of these modalities using inconsistent experimental setups, leaving the empirical effectiveness of co-training largely unexplored.

In this work, we systematically investigate how different data modalities and co-training strategies affect policy performance through large-scale experiments toward a generalist LBM. We adopt a VLA architecture consisting of a pretrained VLM backbone and an action head. Our model is trained with flow matching [46, 49] to predict continuous robot actions and next-token objectives for discrete tokens. An overview of our study is illustrated in Fig. 1.

We evaluate five major co-training data modalities: **(1) Standard vision-language data**, vision-language (VL) supervision for commonsense and spatial reasoning; **(2) Dense language annotations for robot trajectories**, scripted captions derived from heuristic action templates and VLM-generated descriptions; **(3) Cross-embodiment robot data**, demonstrations from robot morphologies other than the target embodiment; **(4) Egocentric human videos**, where video frames are paired with either learned latent action tokens or VLM-generated annotations; and **(5) Discrete robot action tokens**, where continuous robot actions are compressed through frequency-based methods (e.g., FAST [58]) or vector quantization-based methods (e.g., VQ-VAE [72]).

We also study strategies for incorporating these modalities at different training phases (i.e., rounds of training with distinct data compositions) and examine whether combining useful co-training modalities yields cumulative performance gains. Additionally, we probe whether co-training improves the quality of the learned policy representations—i.e., whether it provides better inductive priors that enable more efficient fine-tuning on unseen long-horizon, dexterous tasks.

Our policies are evaluated in both simulation and real-world settings. In total, we train and compare 89 VLA policies using

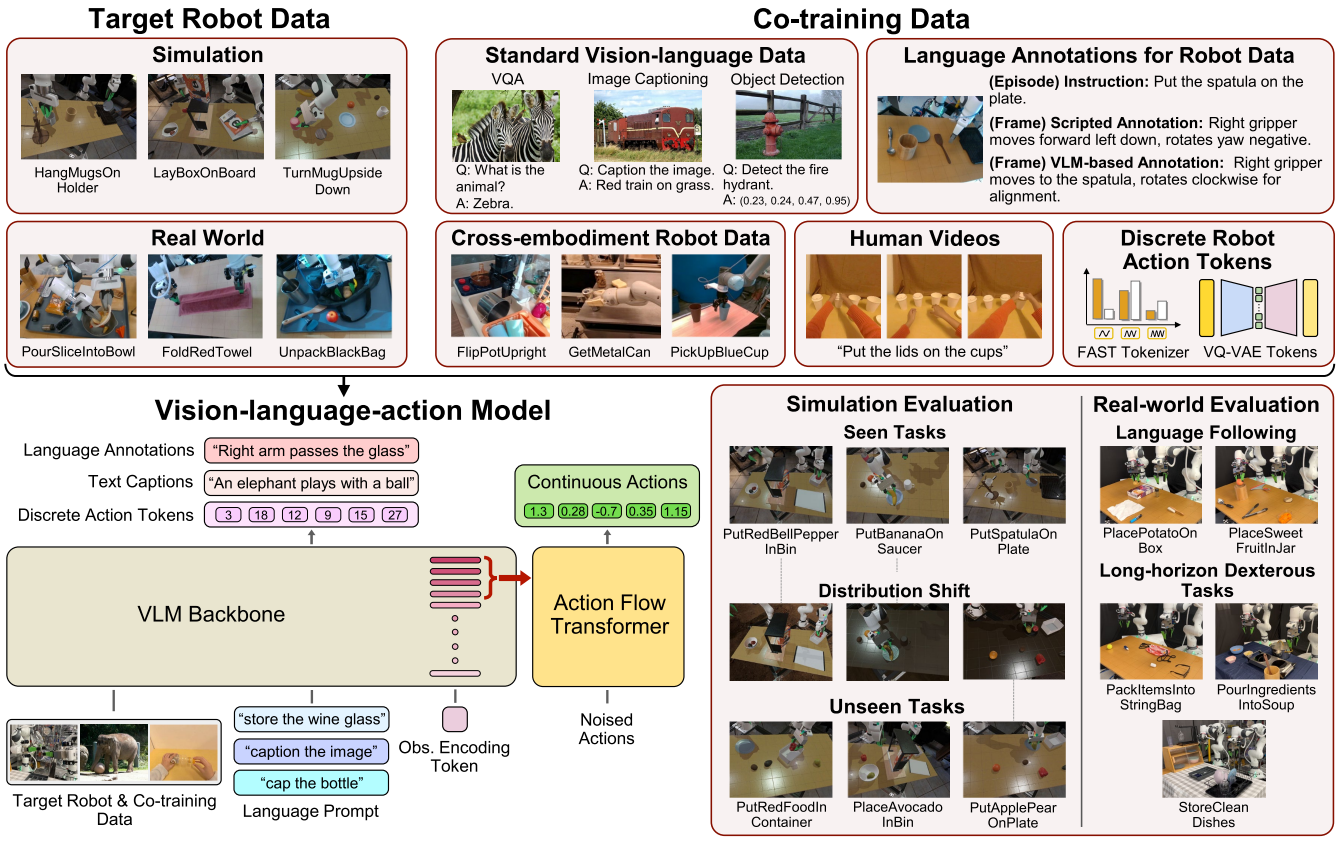

Fig. 1. **Overview of the data, model architecture, and evaluation setup.** Our policy is built on a pretrained vision-language model backbone combined with an Action Flow Transformer. It is trained on target robot data alongside heterogeneous co-training modalities, including standard vision-language data, dense language annotations for robot data, cross-embodiment robot data, human videos, and discrete robot action tokens. We evaluate policies in simulation on seen and unseen tasks, under nominal conditions and distribution shifts, and in the real-world for language following, and long-horizon dexterous manipulation.

about 4,000 hours of robot and human manipulation data plus 50M vision-language samples. The policies are assessed over 58,000 simulation rollouts across seen and unseen tasks, in nominal and distribution shift (DS) settings, and over 2,835 real-world rollouts covering language following and long-horizon dexterous tasks.

Taken together, our results provide practical guidance for co-training LBMs. In summary: diverse vision-language data and cross-embodiment robot data consistently improve generalization to distribution shifts, unseen tasks, and language following, while discrete action token variants give no benefit; combining effective modalities yields cumulative performance gains and enables more efficient adaptation to unseen long-horizon dexterous tasks via fine-tuning; and training only on robot data erodes the VLM backbone's visiolinguistic representations while effective co-training maintains them. Through controlled large-scale experiments, our findings establish which co-training signals and strategies are most useful for building scalable generalist robot policies.

II. Method

In this section, we first introduce our co-training framework in Section II-A, including the problem formulation, model architecture, and our co-training and inference strategies. We then describe in Section II-B how the target robot data and diverse co-training datasets are curated for this study.

A. Co-training Framework

1) Problem Formulation: Our objective is to learn a policy π_θ that can leverage diverse co-training data modalities. The policy π_θ takes as input a sequence of n images $I^{1:n}$ and a text prompt ℓ. For continuous robot actions, the model is trained with flow matching (FM) [46, 49] as the learning objective. Specifically, given an action chunk A_t, a FM timestep $\tau \in [0, 1]$, and sampled noise $\epsilon \sim N(0, \mathbf{I})$, we construct a noised action chunk as $A_t^\tau = \tau A_t + (1 - \tau)\epsilon$. The model is then trained by minimizing the loss:

$$\mathcal{L}_{\text{FM}} = \left\| \pi_\theta^a\left(I_t^{1:n}, \ell, A_t^\tau, \tau\right) - (A_t - \epsilon) \right\|^2, \tag{1}$$

where $(A_t - \epsilon)$ and $\pi_\theta^a(\cdot)$ are the ground-truth and predicted flow vector, respectively. For text tokens or discrete action tokens, the model is optimized to minimize the cross-entropy (CE) loss between the ground-truth token sequence $x_{1:M}$ and the predicted logits $\pi_\theta^\ell(\cdot)$, such that:

$$\mathcal{L}_{\text{CE}} = \mathcal{H}\left(x_{1:M}, \pi_\theta^\ell\left(I_t^{1:n}, \ell\right)\right) \tag{2}$$

When jointly optimizing for both continuous and discrete modalities, we combine the objectives into a weighted sum:

$$\mathcal{L} = M_{FM}\mathcal{L}_{FM} + w * M_{CE}\mathcal{L}_{CE}, \tag{3}$$

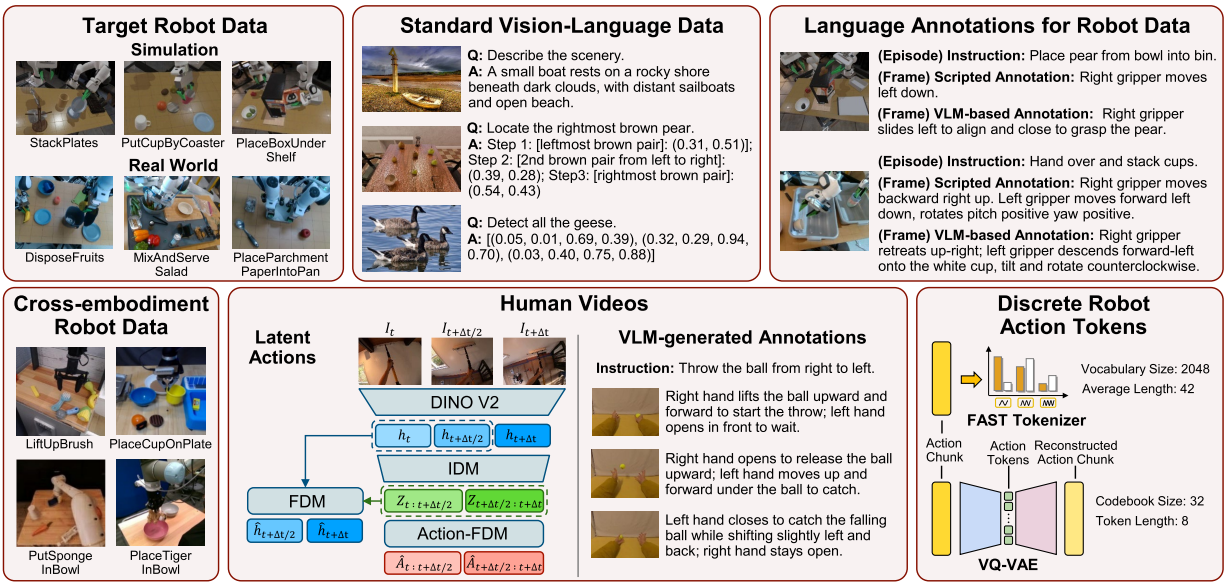

Fig. 2. **Overview of the training data.** Our dataset comprises target robot data collected in both simulation and real-world, and five heterogeneous co-training data modalities: standard vision-language data for commonsense understanding, spatial reasoning, and object grounding; dense language annotations for robot data, generated via heuristic scripting and VLM-based captioning; cross-embodiment robot data capturing diverse robot morphologies and manipulation tasks; human videos, from which we derive either latent action tokens using a latent action model (LAM) or VLM-generated annotations; and discrete robot action tokens, including near-lossless FAST tokens and compact VQ-VAE tokens. These co-training modalities, together with the target robot data, constitute a unified dataset of ∼4,000 hours of manipulation data and 50M vision-language samples.

Here, w is the weight applied to the CE loss, M_{FM} is the FM loss mask indicating whether the model should predict continuous actions for a given sample, and M_{CE} is the mask specifying token positions used to compute the CE loss.

2) Model Architecture: We adopt a VLA architecture (Fig. 1) composed of a pretrained VLM backbone and a flow transformer action head (ActionFT). The VLM is initialized from PaliGemma2-PT (google/paligemma2-3b-pt-224 [65]). It encodes both the observed images and the language prompt describing the task, and can optionally be trained to generate text or discrete action tokens in addition to providing visiolinguistic representations. To obtain a compact representation for action generation, similarly to [45], we introduce a special observation encoding token into the backbone's vocabulary and append it to the end of the text prompt. We extract the hidden state vectors corresponding to this token from the last four layers of the VLM to form a single global conditioning embedding and feed this into the ActionFT. The ActionFT follows the diffusion transformer design introduced in [1]. Our ablation studies (see Appendix 2A) indicate that this compact representation enhances the model's generalization ability to unseen tasks and distribution shift.

3) Co-training and Inference Strategies: Co-training data can be used at different phases of model training. We explore three strategies (Table I): (1) *Single-phase co-training*: train jointly on target robot data and co-training data modalities in a single phase; (2) *Two-phase 1st-phase-only co-training*: train only on co-training modalities during the 1st-phase, then on target robot continuous actions in the 2nd-phase; (3) *Two-phase full co-training*: the same 1st-phase as strategy (2), but train on both co-training and target robot data in the 2nd-phase.

TABLE I
DATA COMPOSITION FOR DIFFERENT CO-TRAINING STRATEGIES.

Co-training strategy	**Phase**	**Data composition**	
		Target-robot continuous-action data	**Co-training data**
Single-phase co-training	1st-phase	✓	✓
Two-phase 1st-phase-only co-training	1st-phase	–	✓
	2nd-phase	✓	–
Two-phase full co-training	1st-phase	–	✓
	2nd-phase	✓	✓

We fix the training loss weights and batch data ratios based on ablation experiment results (Appendix 2B). Full training details are provided in Appendix 3. While our primary approach treats co-training data solely as auxiliary supervision, we explore chain-of-thought (CoT) conditioning in action generation to take advantage of co-training data, as shown in Appendix 4.

B. Data Curation

We curate target-robot demonstrations and five co-training modalities, illustrated in Fig. 2.

1) Target Robot Data: We adopt TRI-Ramen [1] as our in-distribution corpus: 523 hours across 403 tasks and 53,411 demonstrations, including TRI-Ramen-Real (478 hours) and TRI-Ramen-Sim (45 hours).

2) Standard Vision-Language Data: To enhance multimodal understanding, we use RoboPoint [91] and RefSpatial [96], covering object grounding and spatial reasoning.

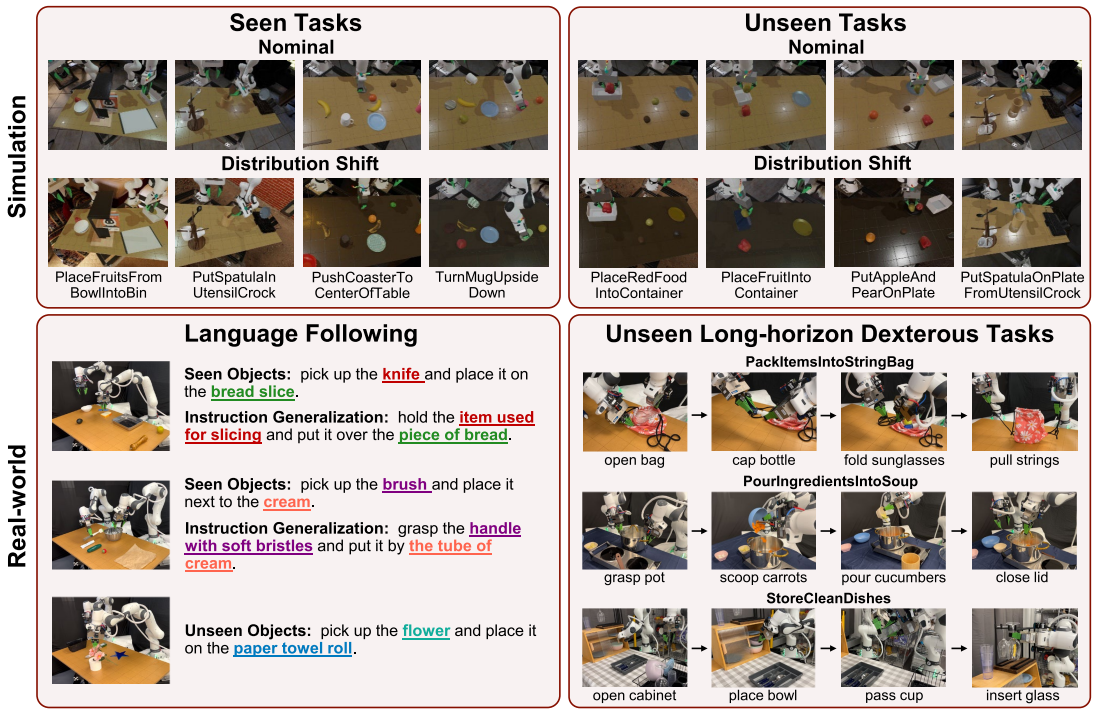

Fig. 3. **Simulation and real-world evaluation.** Policies are evaluated in simulation on 13 seen and 8 unseen tasks under nominal and distribution shift (DS) conditions, where DS introduces appearance changes (e.g., lighting, textures, distractors, camera parameters). Real-world evaluations include language-following experiments with seen objects, instruction generalization through paraphrasing, and unseen objects, as well as adaptation to unseen long-horizon dexterous tasks via fine-tuning.

3) Dense Language Annotations for Robot Trajectories: We augment TRI-Ramen with dense action-grounded text via (i) scripted low-level primitives from relative differences in end-effector poses over the 16-step horizon, and (ii) VLM-generated captions using GPT-5 [53] from downsampled frames, task instruction, scripted action hints, and a world-frame reference image. See Appendix 5B for primitive heuristics, templates, and prompts.

4) Cross-embodiment Robot Data: We use OXE-Ramen [1], a curated Open X-Embodiment subset [55]: 1,150 hours over 12 robot setups, 924 tasks, and 466,415 demonstrations. Observation and action spaces follow the same as in the target robot data.

5) Human Videos: We use egocentric human video datasets (Ego4D [28], EgoDex [31], Something-Something V2 [27], Epic Kitchen [18], HoloAssist [75]), totaling 2,271 hours after filtering (Table S3). We derive (i) discrete latent-action tokens using a latent action model (LAM) trained jointly with robot videos, and (ii) VLM-generated motion descriptions to utilize during training as additional vision-language data. Full curation and latent model details are in Appendix 5C.

6) Discrete Robot Action Tokens: We study two discrete representations for robot action chunks: (i) FAST tokens [58] using an off-the-shelf tokenizer (average length 42.1 with a vocabulary of 2,048 tokens on TRI-Ramen), and (ii) compact VQ-VAE tokens [72] producing 8 tokens per chunk with codebook size 32 (matched to our video latent-action dimensionality). Additional tokenizer and architecture details are in Appendix 5D.

III. Experiments

To systematically investigate the effectiveness of co-training data modalities and strategies, we conduct large-scale experiments to address the following research questions:

1) How do different co-training data modalities, incorporated at different training phases, influence policy performance on various dimensions (in-distribution, generalization to distribution shift, unseen tasks, and language following)?
2) Does combining effective co-training modalities yield cumulative performance gains?
3) Can co-training enhance the quality of learned representations, thereby enabling rapid adaptation via fine-tuning to unseen long-horizon, dexterous tasks?

A. Evaluation

We evaluate policy performance across multiple important dimensions: in-distribution performance, robustness to DS, generalization to unseen tasks, and language following. We additionally assess adaptation to unseen long-horizon dexterous manipulation tasks via fine-tuning. To this end, we conduct large-scale simulation and real-world experiments (Fig. 1, 3).

1) Simulation Benchmark: We adopt the simulation benchmark from [1], built on Drake [70]. The benchmark includes 13 seen tasks and 8 unseen tasks (3 from [1] and 5 newly introduced to further probe generalization; see Appendix 6). Each policy is evaluated with 50 rollouts per task under both nominal and DS conditions, and performance is measured by success rate.

Seen tasks lie within the training distribution, while unseen tasks probe generalization beyond it. Unseen tasks target: (i) semantic understanding (e.g., identifying "red food" or distinguishing fruits from vegetables); (ii) multi-step manipulation (e.g., sequential placement of multiple objects); and (iii) compositional generalization (e.g., recombining previously seen skills or actions in novel task configurations). Here, "unseen tasks" correspond to novel skills absent from the training data, while the underlying objects and environments remain shared.

To evaluate robustness to appearance changes, DS conditions introduce variations in lighting, background, camera parameters, and object/table textures and colors, while nominal conditions match the training distribution.

2) Real-world Evaluation: We evaluate policies on a dual-arm Franka robot platform in two real-world settings. Deployment details are provided in Appendix 7.

Language Following. We design language-guided pick-and-place experiments covering three scenarios: **(i) Seen Objects:** Objects appear during training; instructions follow the template "pick up [object A] and place it in/on/next to [object B]". **(ii) Instruction Generalization:** Instructions are rephrased to test semantic understanding, including object categories (e.g., "writing tools" referring to pens), attribute-based references (e.g., "the handle with soft bristles" referring to a brush), and paraphrasing robustness. **(iii) Unseen Objects:** Objects are absent from the target-robot training data and instructions follow the same template as (i).

Across all settings, we evaluate 15 spatial layouts with 6–8 tabletop objects each and three language instructions per layout, yielding 45 rollouts per setting. The instruction generalization setting uses identical layouts and targets as the seen-object setting with varied phrasing. The full suite includes 49 seen and 52 unseen objects. We report average task completion percentage. See Appendix 8 for rubrics, procedures, and quality assurance (QA).

Long-horizon Dexterous Manipulation. To assess adaptation to challenging tasks unseen during pretraining, we evaluate on three long-horizon tasks: *PackItemsIntoStringBag*, *PourIngredientsIntoSoup*, and *StoreCleanDishes*. Each task averages 13 intermediate steps spanning 93 seconds and requires fine-grained manipulation beyond pick-and-place (e.g., capping bottles, scooping with a spatula, or inserting a wine glass into a rack). We collect 200 demonstrations per task for fine-tuning. Each checkpoint is evaluated with 30 rollouts per task, reporting average task completion percentage (see Appendix 8).

3) Statistical Analysis Framework: We perform rigorous statistical analysis similar to recent work [1], with pairwise hypothesis tests [78, 64] and Compact Letter Display (CLD) [59] for comparison. Co-training strategies not sharing any CLD alphabet are significantly different in average performance at 5% family-wise error rate (FWER). Bayesian uncertainty estimates are reported for individual strategies, with posterior uncertainty visualized as violin plots overlaid on bar charts. Dots and horizontal lines indicate empirical and posterior means, respectively. Raw empirical distributions are reported in Appendix 9 for the task-completion results, together with more details of our statistical framework.

B. How do different co-training modalities, incorporated at different training phases, influence policy performance?

We evaluate each co-training data modality using the three strategies described in Section II-A3: *single-phase co-training*, *two-phase 1st-phase-only co-training*, and *two-phase full co-training*. We compare the policies obtained from these strategies against a baseline policy trained exclusively on continuous robot actions from TRI-Ramen (namely, *no-co-training baseline*) using the flow-matching objective (M_{CE}=0). All policies are first evaluated in simulation, with results summarized in Fig. 4. Modalities found to be effective are then evaluated in real-world language-following experiments (Fig. 5).

Standard Vision-language Data Co-training. (1) As shown in Fig. 4A and 5A, co-training with standard VL data substantially improves robustness to DS, generalization to unseen tasks, and language following, with no statistically significant change in in-distribution performance (seen tasks). (2) Fine-tuning the pretrained VLM on our curated VL data enhances its representation for robot manipulation tasks, as evidenced by gains from *two-phase 1st-phase-only co-training* over the baseline (Fig 4A, 5A). (3) Continuing to co-train with VL data during the 2nd-phase further enhances performance on unseen tasks and language following (particularly with unseen objects). We hypothesize that this continued exposure allows the model to retain the rich, generalizable knowledge from the VL corpus, which is absent in the robot data, thereby preventing catastrophic forgetting.

Dense Language Annotations for Robot Data Co-training. (1) Co-training with scripted (Fig. 4B, 5B) and VLM-based (Fig. 4C, 5C) annotations for robot data improves the model's robustness to DS, generalization to unseen tasks, and language-following, with no statistically significant change in in-distribution performance. (2) Owing to greater linguistic diversity and richer descriptions of object-environment interactions, VLM-based annotations yield more substantial improvements on unseen tasks and language following compared to scripted annotations (Fig. 4B-C, 5B-C). (3) In two-phase co-training, incorporating these annotations during the 2nd-phase yields no additional benefit, suggesting that they are most effective when used exclusively during the 1st-phase (Fig. 4B-C, 5B-C). We posit that, because these annotations describe the same physical trajectories as the robot action data, their utility primarily lies in bootstrapping language-action alignment during the 1st-phase training, rather than introducing new information during the 2nd-phase.

Cross-embodiment Robot Data Co-training. (1) As shown in Fig. 4D and 5D, co-training with cross-embodiment robot data improves robustness to DS, generalization to unseen tasks, and language following, with no statistically significant change in in-distribution performance. (2) Cross-embodiment robot data is most effective, providing the largest gains, when confined to the first phase of two-phase co-training, particularly for unseen-task generalization and robustness under DS

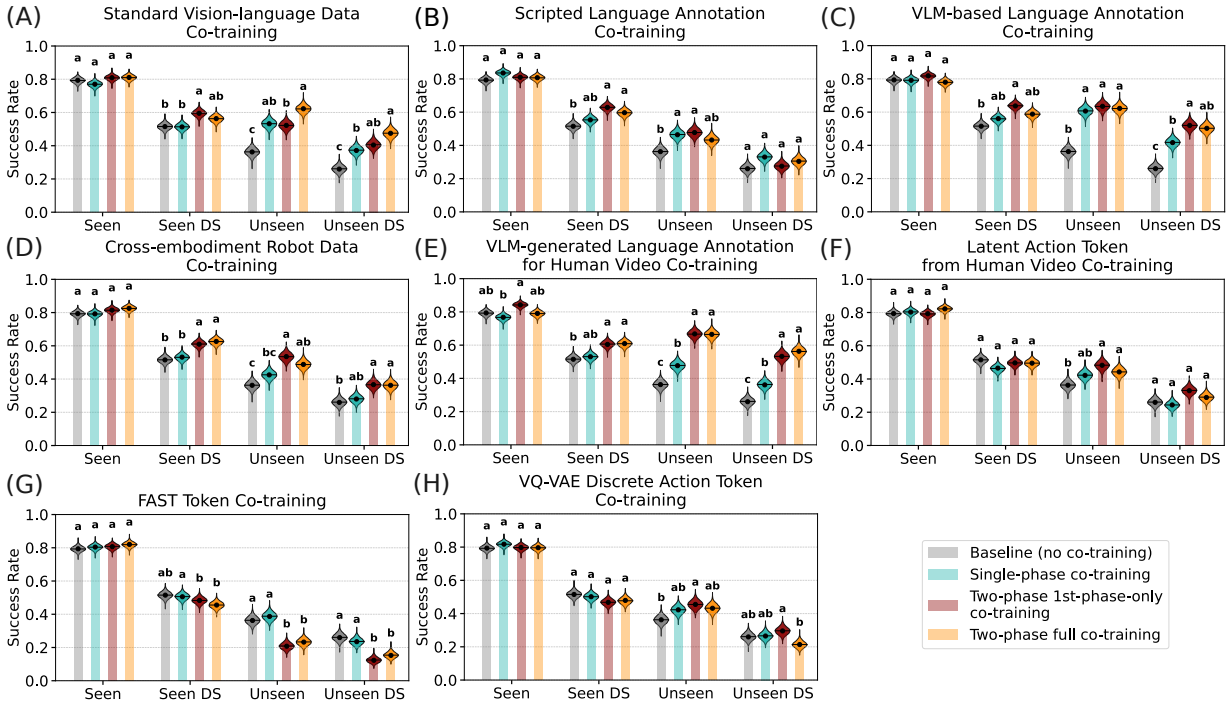

Fig. 4. **Simulation ablation of co-training data and strategies.** Comparison of the no-co-training baseline with policies co-trained on a single data modality across sequential training phases. Policies are evaluated on seen and unseen tasks under nominal and distribution shift conditions (A–H denote data modalities).

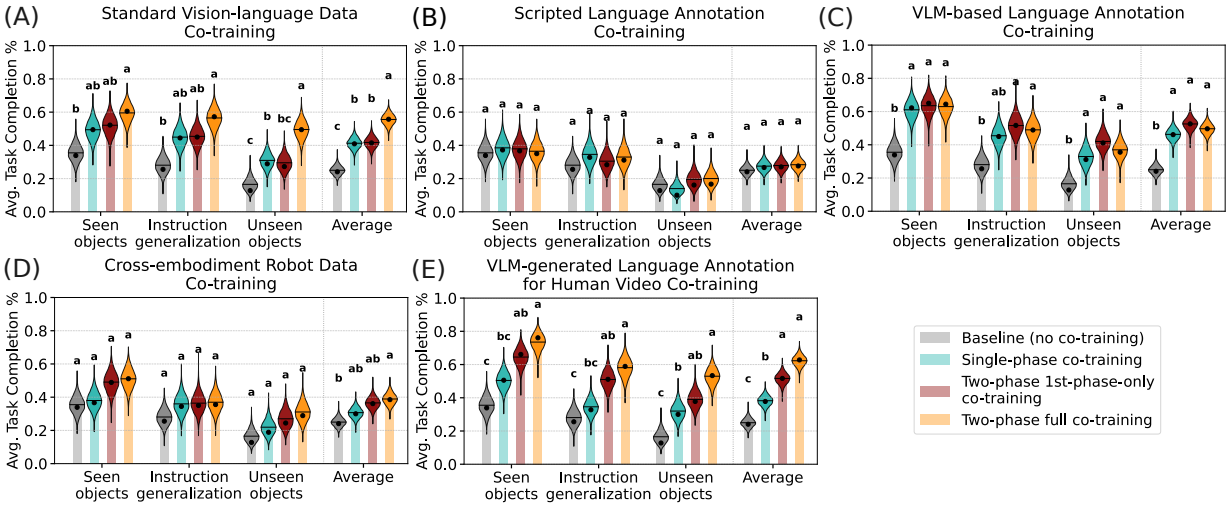

Fig. 5. **Real-world ablation of co-training data and strategies.** Performance of the no-co-training baseline and policies co-trained with a single data modality across training phases, evaluated at language-following with seen objects, instruction generalization, and unseen objects (A–E denote data modalities).

(Fig. 4D). When included during the second phase of training in the *two-phase full co-training*, it yields negligible additional benefit for language-following. We hypothesize that the diverse morphologies and manipulation strategies from cross-embodiment data are most valuable for learning generalizable visual and behavioral representations during the 1st-phase training. In the 2nd-phase, the model benefits from specializing to the target embodiment, at which point continued exposure to other embodiments provides limited value.

Human Video Co-training. *A) Latent Actions:* In *single-phase co-training*, the model jointly learns continuous actions for TRI-Ramen and discrete latent action tokens extracted from all video data (TRI-Ramen, OXE-Ramen, and human videos). For two-phase methods, the model learns latent actions from all video data in the 1st-phase. In *two-phase full co-training*, it learns both TRI-Ramen continuous actions and latent actions from all video data during the 2nd-phase.

Single-phase co-training with latent actions yields no improvement over the baseline (Fig. 4F), whereas *single-phase co-training* with other effective modalities (e.g., standard vision language data) consistently improves performance. On the other hand, for two-phase co-training: (1) latent action

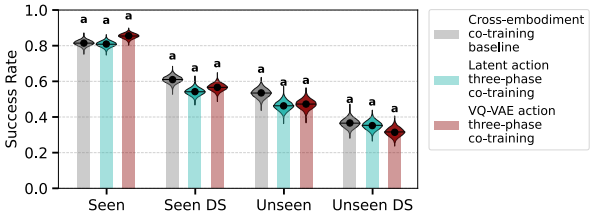

Fig. 6. **Ablation of latent action and VQ-VAE action co-training under data and compute-rich settings.** Simulation results indicate that these modalities show no performance gains over the cross-embodiment baseline.

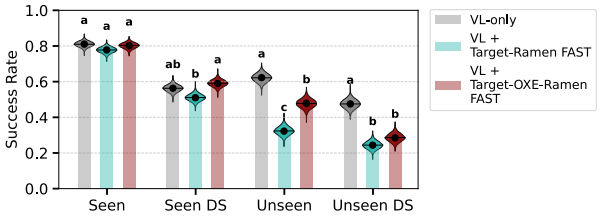

Fig. 7. **Ablation of FAST token co-training on a broad mix of robot and standard vision-language (VL) data.** FAST co-training along with VL data fails to improve overall performance and degrades generalization on unseen tasks compared to the VL-only co-training baseline.

1st-phase training improves performance on unseen tasks, and (2) incorporating latent actions during the 2nd-phase provides no additional benefit. These results suggest that the benefits of latent action 1st-phase training out of two phases may stem from increased computation rather than genuine knowledge transfer. To investigate this, we design a data- and compute-rich setting, comparing two policies: (a) Cross-embodiment co-training baseline: 1st-phase trained on all robot data (TRI-Ramen and OXE-Ramen), and 2nd-phase on TRI-Ramen; this baseline is equivalent to the *two-phase 1st-phase-only co-training* with cross-embodiment data without latent action co-training. (b) Latent action three-phase co-training: (i) train on latent actions from all video data, (ii) train on all robot data for continuous actions, and (iii) train on TRI-Ramen. Notably, the only difference between these two methods is that the latter includes an additional initial training phase on all video data. As shown in Fig. 6, the added initial latent action training phase provides no benefit in this data- and compute-rich setting.

Given that prior works [10, 5, 13] highlight the utility of latent action pretraining in a low target-robot-data regime, we further explore its effectiveness across varying scales of robot data, ranging from single-task to the full robot dataset. Fig. 8 shows that latent action 1st-phase training improves performance in the low target-robot-data regime, but these benefits diminish as the quantity of fine-tuning robot data increases.

B) VLM-generated Annotations: (1) As shown in Fig. 4E and 5E, co-training with VLM-generated annotations for human videos improves robustness to DS, generalization to unseen tasks, and language following, with no statistically significant change in in-distribution performance. (2) In two-phase co-training (Fig. 5E), continuing to incorporate these annotations during the 2nd-phase enhances language-following performance, particularly for unseen objects. We attribute this benefit to the rich diversity of motions, objects, and environments in human videos, which is absent from TRI-Ramen. Joint training during the 2nd-phase allows the model to maintain this broader world knowledge, rather than narrowing to the distribution of target robot data.

Discrete Robot Action Tokens Co-training. *A) FAST Tokens:* Our results suggest that FAST token co-training fails to improve performance across all dimensions and degrades generalization on unseen tasks (Fig. 4G). Prior works [33, 22] show that FAST token co-training could improve performance when pretrained on a broad mix of robot and standard VL data. To examine this claim, we compare three approaches: (a) VL + Ramen-OXE-Ramen FAST: 1st-phase training on all robot data using FAST tokens, alongside VL data. (b) VL + TRI-Ramen FAST: 1st-phase training only on TRI-Ramen using FAST tokens, alongside VL data. (c) VL-only: 1st-phase training with VL data only. All methods employ an identical 2nd-phase: continuous action learning on TRI-Ramen through flow matching, with standard VL data co-training.

As illustrated in Fig. 7, FAST token co-training fails to improve overall performance and degrades generalization on unseen tasks. However, including OXE-Ramen during the 1st-phase significantly outperforms training on TRI-Ramen alone, indicating that FAST co-training might prove beneficial when scaled to substantially larger robot datasets, though it remains ineffective at our current data scale. We attribute this to the nature of FAST tokens as near-lossless action representations: co-training with FAST tokens may bias the VLM backbone toward learning precise action mappings rather than generalizable features.

B) VQ-VAE Discrete Action Tokens: (1) VQ-VAE discrete action token co-training yields marginal improvements on unseen tasks but slightly degrades performance under DS (Fig. 4H). (2) Incorporating VQ-VAE tokens during the 2nd-phase of two-phase co-training provides no additional benefit.

Similar to latent action co-training, we examine its effectiveness in a data- and compute-rich setting by comparing a cross-embodiment co-training baseline against VQ-VAE discrete action three-phase co-training policy that begins with learning VQ-VAE discrete action tokens on all robot data during the 1st-phase. As illustrated in Fig. 6, VQ-VAE co-training yields no improvement in this regime.

Summary. (1) Co-training with diverse VL data and cross-embodiment robot data substantially enhances the model's generalization to DS, unseen tasks, and language-following capabilities. Notably, owing to their information richness, co-training with standard VL data and language annotations for human videos benefits both 1st-phase and 2nd-phase co-training, whereas language annotations for robot trajectories and cross-embodiment data are primarily effective during 1st-phase in two-phase co-training. (2) Across all the effective co-training data modalities, standard VL data, VLM-based language annotations for robot data, and language annotations for human videos are the most beneficial. These three specific modalities are all in the form of diverse VL data, suggesting

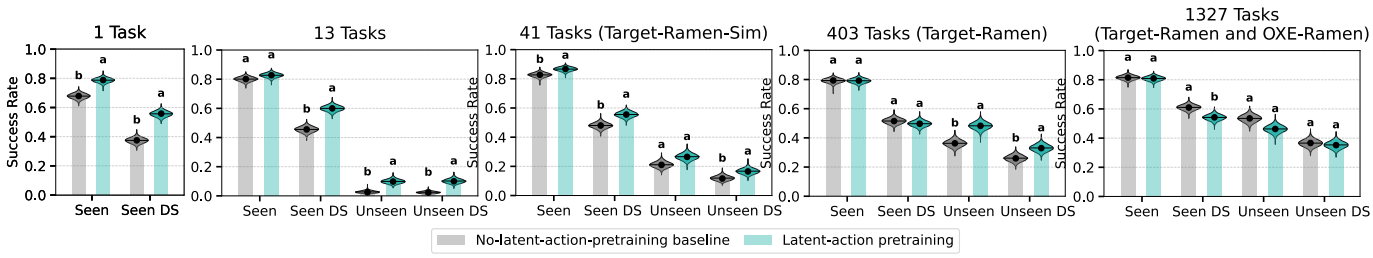

Fig. 8. **Ablation of latent action co-training on increasing data quantity.** Latent action 1st-phase in a multi-phase training yields performance gains in the low target-robot data regime, with diminishing returns as the quantity of fine-tuning robot data increases.

that strengthening VL understanding of the VLM backbone translates into better robot policies. (3) Discrete action tokens (including latent actions extracted from videos, FAST tokens, and action tokens learned from VQ-VAE) co-training yield no statistically significant performance improvements in our experiments. Specifically, co-training with FAST tokens decreases generalization, while latent actions from videos only provide benefits in the low target-robot-data regime, with benefits diminishing as the proportion of robot data increases. (4) Across all co-training modalities examined, we observe no statistically significant impact on in-distribution performance.

Fig. 9 compares the best training strategies for each useful co-training modality. Specifically, for standard VL and VLM-generated language annotations for human videos, the best models correspond to *two-phase full co-training*. For scripted and VLM-based language annotation and cross-embodiment robot data, the best models correspond to *two-phase 1st-phase-only co-training*. Co-training with standard VL data, VLM-based annotations for robot data, and annotations for human videos most substantially improves performance on unseen tasks and language following. Furthermore, benefiting from the rich information absent in robot demonstrations, co-training with standard VL data and annotations for human videos more effectively enables the model to recognize unseen objects compared to VLM-based annotations for robot data.

C. Does combining effective co-training modalities yield cumulative performance gains?

Having identified effective co-training data modalities along with their optimal training strategies, we further investigate whether combining these modalities yields cumulative benefits. We conduct an ablation study where we incrementally add each effective data source (training details are provided in Table S1): (1) Baseline without co-training, (2) +Vision-Language-Data: Standard VL data co-training only, (3) +Robot-Annotation-Data: Adding dense language annotations (both scripted and VLM-based) for robot data, (4) +Human-Video-Annotation-Data: Adding VLM-generated annotations for human videos, (5) +Cross-Embodiment-Robot-Data (namely, Final Model): Adding cross-embodiment robot data.

As shown in Fig. 10, combining effective co-training modalities yields consistent cumulative performance gains across all evaluation dimensions. Our Final Model achieves strong performance across all settings, attaining a 72.6% empirical

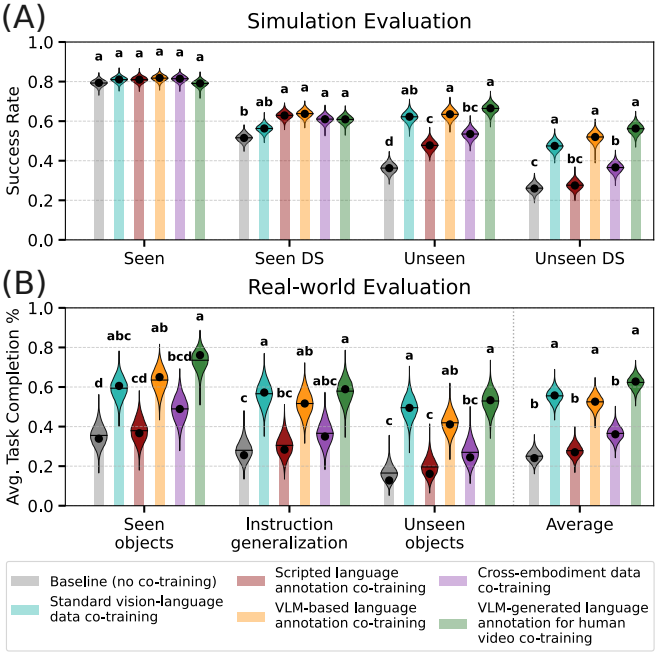

Fig. 9. **Performance of policies trained with the best co-training strategies for effective co-training data modalities.** A) Simulation results on seen and unseen tasks under nominal and distribution-shift conditions. B) Real-world language-following performance on seen objects, instruction generalization, and unseen objects. Diverse vision-language and cross-embodiment robot data substantially enhance the model's generalization to distribution shifts, unseen tasks, and language-following without affecting in-distribution performance.

success rate on simulation unseen tasks (36.4% improvement over baseline) and 69.4% empirical average task completion on real-world language following (45.3% improvement over baseline[1]). In Appendix 10, we benchmark how the combined effective modalities shape the VLM backbone of our policies.

D. Can co-training enhance the quality of learned representations, thereby enabling rapid adaptation to unseen long-horizon, dexterous tasks?

Prior works [1, 6, 33] have demonstrated that high-quality pretraining enables policies to rapidly adapt to downstream unseen tasks through fine-tuning. We investigate whether co-training enhances learned representation quality by fine-tuning (FT) on our suite of unseen long-horizon, dexterous manipulation tasks. We compare three approaches: (1) FT Final Model: fine-tune our Final Model from Section III-C (trained with all

[1]Improvement rates are calculated using the empirical means.

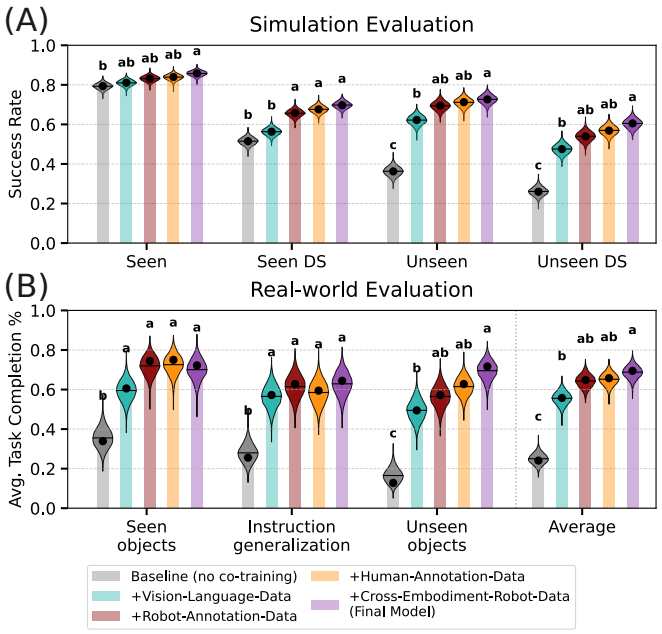

Fig. 10. **Performance of policies co-trained with the effective data modalities additively combined.** A) Simulation results on seen and unseen tasks under nominal and distribution shift conditions. B) Real-world evaluations for language-following performance across seen objects, instruction generalization, and unseen objects settings. Combining the effective co-training data modalities yields cumulative gains in policy performance.

effective co-training modalities), (2) FT Baseline: fine-tune the no-co-training baseline model (trained only on TRI-Ramen), (3) Single Task: a single-task policy without pretraining.

As shown in Fig. 11, benefiting from our curated co-training data and co-training strategies, our Final Model rapidly acquires new skills through fine-tuning, achieving 90.2% average task completion with only 200 demonstrations—a 22.8% improvement over the FT Baseline and a 42.9% improvement over the Single Task policy. This demonstrates that effective co-training substantially enhances representation quality, thereby enabling more fine-grained action learning in downstream tasks. We observe that failures in the FT Baseline and the Single Task policy predominantly stem from insufficient precision in manipulation: they frequently fail to align and secure the cap onto the bottle in *PackItemsIntoStringBag*, misalign the spatula for grasping in *PourIngredientsIntoSoup*, and struggle to grasp the transparent cup in *StoreCleanDishes*. In contrast, the FT Final Model consistently executes these fine-grained manipulations with high precision.

IV. Discussion, Limitations, And Future Work

We present a large-scale empirical study that systematically dissects the impact of diverse co-training data and strategies on the performance of LBMs. Our findings reveal that co-training with vision-language data and cross-embodiment robot data substantially enhances generalization to DS, unseen tasks, and language following capabilities, while discrete action token variants yield no statistically significant benefits. Furthermore, we show that combining effective modalities produces cumulative performance gains and enables rapid adaptation to

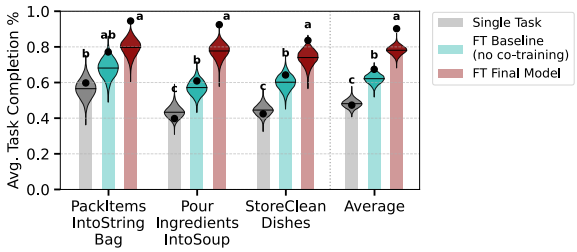

Fig. 11. **Adaptation to unseen long-horizon, dexterous tasks via fine-tuning.** Real-world evaluation of a Single Task policy trained from scratch, a fine-tuned (FT) Baseline pretrained only on robot data, and our FT Final Model pretrained with all effective co-training modalities. The FT Final Model achieves higher average task completion, highlighting the role of co-training in learning transferable, high-quality representations.

dexterous, long-horizon tasks via fine-tuning.

Notably, among all useful co-training modalities, diverse vision-language data—including standard datasets and rich annotations for robot and human videos—demonstrate the most substantial improvements. This observation resonates with the Good Regulator Theorem [17], which states that a system must incorporate an internal model (implicit or explicit) of its operating world to effectively regulate it. In our setting, strong foundation models (such as VLMs) provide precisely such internal models that have rich semantic and spatial understanding of the physical world. Our results suggest that progress towards truly generalist robot policies is intrinsically linked to advances in these foundation models.

While our study provides promising insights, several limitations should be acknowledged. First, while we examine various sources of vision-language data, we do not systematically analyze their impact by task taxonomy (e.g., visual question answering, image captioning, object detection, spatial reasoning). Understanding how different vision-language task categories affect specific policy capabilities would enable more targeted and sample-efficient data curation. Second, we explore only coarse-grained representations for human videos through latent actions and language annotations. As hand pose estimation techniques advance and dexterous robotic hands continue to evolve, explicitly extracting fine-grained dexterous motions from human video may become a valuable co-training signal. Finally, our study focuses exclusively on imitation learning; exploring co-training within alternative learning paradigms, such as world modeling or reinforcement learning, remains an open frontier for developing scalable generalist policies.

Acknowledgments

We thank Muhammad Zubair Irshad, Vitor Campagnolo Guizilini, Sedrick Keh, and Swati Gupta for the manuscript reviews; Kosei Tanada, Takuto Ito, Phoebe Horgan, Gordon Richardson, and Aykut Onol for evaluation quality assurance; Sam Shereda, Mariah Smith-Jones, and Matthew Tran for evaluation and data collection; Alejandro Castro and Joseph Masterjohn for simulation support, and to the TRI's PROPS team for assistance with experimental setup.

References

[1] Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, et al. A careful examination of large behavior models for multitask dexterous manipulation. *arXiv preprint arXiv:2507.05331*, 2025.

[2] Erik Bauer, Elvis Nava, and Robert K Katzschmann. Latent action diffusion for cross-embodiment manipulation. *arXiv preprint arXiv:2506.14608*, 2025.

[3] Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language. *arXiv preprint arXiv:2403.01823*, 2024.

[4] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[5] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[6] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

[7] Nico Bohlinger, Grzegorz Czechmanowski, Maciej Krupka, Piotr Kicki, Krzysztof Walas, Jan Peters, and Davide Tateo. One policy to run them all: an end-to-end learning approach to multi-embodiment locomotion. *arXiv preprint arXiv:2409.06366*, 2024.

[8] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[9] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2502.14420*, 2025.

[10] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

[11] William Chen, Suneel Belkhale, Suvir Mirchandani, Oier Mees, Danny Driess, Karl Pertsch, and Sergey Levine. Training strategies for efficient embodied reasoning. *arXiv preprint arXiv:2505.08243*, 2025.

[12] Xiaoyu Chen, Junliang Guo, Tianyu He, Chuheng Zhang, Pushi Zhang, Derek Cathera Yang, Li Zhao, and Jiang Bian. Igor: Image-goal representations are the atomic control units for foundation models in embodied ai. *arXiv preprint arXiv:2411.00785*, 2024.

[13] Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, et al. Villa-x: enhancing latent action modeling in vision-language-action models. *arXiv preprint arXiv:2507.23682*, 2025.

[14] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[15] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

[16] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[17] Roger C Conant and W Ross Ashby. Every good regulator of a system must be a model of that system. *International journal of systems science*, 1(2):89–97, 1970.

[18] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11): 4125–4141, 2020.

[19] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv e-prints*, pages arXiv–2409, 2024.

[20] Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. *arXiv preprint arXiv:2408.11812*, 2024.

[21] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. *ICML23: Proceedings of the 40th International*

Conference on Machine Learning, 2023.

[22] Danny Driess, Jost Tobias Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Z Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, et al. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *arXiv preprint arXiv:2505.23705*, 2025.

[23] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024.

[24] Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot visual imitation learning via meta-learning. In *Conference on robot learning*, pages 357–368. PMLR, 2017.

[25] Ken Goldberg. Good old-fashioned engineering can close the 100,000-year "data gap" in robotics, 2025.

[26] Ankit Goyal, Hugo Hadfield, Xuning Yang, Valts Blukis, and Fabio Ramos. Vla-0: Building state-of-the-art vlas with zero modification. *arXiv preprint arXiv:2510.13054*, 2025.

[27] 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.

[28] 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.

[29] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[30] Asher J Hancock, Xindi Wu, Lihan Zha, Olga Russakovsky, and Anirudha Majumdar. Actions as language: Fine-tuning vlms into vlas without catastrophic forgetting. *arXiv preprint arXiv:2509.22195*, 2025.

[31] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

[32] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[33] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054.

[34] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[35] Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea open-world dataset and g0 dual-system vla model. *arXiv preprint arXiv:2509.00576*, 2025.

[36] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13226–13233. IEEE, 2025.

[37] 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.

[38] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success, 2025. *URL https://arxiv.org/abs/2502.19645*, 2025.

[39] Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.

[40] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Masquerade: Learning from in-the-wild human videos using data-editing. *arXiv preprint arXiv:2508.09976*, 2025.

[41] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos. *arXiv preprint arXiv:2503.00779*, 2025.

[42] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39): 1–40, 2016.

[43] Yi Li, Yuquan Deng, Jesse Zhang, Joel Jang, Marius Memmel, Raymond Yu, Caelan Reed Garrett, Fabio Ramos, Dieter Fox, Anqi Li, et al. Hamster: Hierarchical action models for open-world robot manipulation. *arXiv preprint arXiv:2502.05485*, 2025.

[44] Fanqi Lin, Ruiqian Nai, Yingdong Hu, Jiacheng You, Junming Zhao, and Yang Gao. Onetwovla: A unified vision-language-action model with adaptive reasoning. *arXiv preprint arXiv:2505.11917*, 2025.

[45] Juyi Lin, Amir Taherin, Arash Akbari, Arman Akbari, Lei Lu, Guangyu Chen, Taskin Padir, Xiaomeng Yang, Weiwei Chen, Yiqian Li, Xue Lin, David Kaeli, Pu Zhao, and Yanzhi Wang. Vote: Vision-language-action optimization with trajectory ensemble voting. *arXiv preprint arXiv:2507.05116*, 2025. URL https://arxiv.org/abs/2507.05116.

[46] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[47] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[48] Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025.

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

[50] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

[51] Yangcen Liu, Woo Chul Shin, Yunhai Han, Zhenyang Chen, Harish Ravichandar, and Danfei Xu. Immimic: Cross-domain imitation from human videos via mapping and interpolation. *arXiv preprint arXiv:2509.10952*, 2025.

[52] Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025.

[53] OpenAI. Gpt-5 is here, 2025. URL https://openai.com/gpt-5/.

[54] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[55] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya 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.

[56] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[57] Baoqi Pei, Yifei Huang, Jilan Xu, Guo Chen, Yuping He, Lijin Yang, Yali Wang, Weidi Xie, Yu Qiao, Fei Wu, et al. Modeling fine-grained hand-object dynamics for egocentric video representation learning. *arXiv preprint arXiv:2503.00986*, 2025.

[58] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

[59] Hans-Peter Piepho. An algorithm for a letter-based representation of all-pairwise comparisons. *Journal of Computational and Graphical Statistics*, 13(2):456–466, 2004.

[60] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J Yoon, Ryan Hoque, Lars Paulsen, et al. Humanoid policy˜ human policy. *arXiv preprint arXiv:2503.13441*, 2025.

[61] Delin Qu, Haoming Song, Qizhi Chen, Zhaoqing Chen, Xianqiang Gao, Xinyi Ye, Qi Lv, Modi Shi, Guanghui Ren, Cheng Ruan, et al. Eo-1: Interleaved vision-text-action pretraining for general robot control. *arXiv preprint arXiv:2508.21112*, 2025.

[62] Lucy Xiaoyang Shi, Brian Ichter, Michael Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025.

[63] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

[64] David Snyder, Asher James Hancock, Apurva Badithela, Emma Dixon, Patrick Miller, Rares Andrei Ambrus, Anirudha Majumdar, Masha Itkina, and Haruki Nishimura. Is your imitation learning policy better than mine? policy comparison with near-optimal stopping. In *Proceedings of the Robotics: Science and Systems Conference (RSS) XXI*, 2025.

[65] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024.

[66] Tony Tao, Mohan Kumar Srirama, Jason Jingzhou Liu, Kenneth Shaw, and Deepak Pathak. Dexwild: Dexterous human interactions for in-the-wild robot policies. *arXiv preprint arXiv:2505.07813*, 2025.

[67] Gemini Robotics Team, Abbas Abdolmaleki, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Ashwin Balakrishna, Nathan Batchelor, Alex Bewley, Jeff Bingham, et al. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342*, 2025.

[68] Gemini Robotics Team, S Abeyruwan, J Ainslie, JB Alayrac, MG Arenas, T Armstrong, A Balakrishna, R Baruch, M Bauza, M Blokzijl, et al. Gemini robotics: Bringing ai into the physical world, 2025. *URL https://arxiv. org/abs/2503.20020*, 2025.

[69] 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.

[70] Russ Tedrake et al. Drake: Model-based design and verification for robotics, 2019.

[71] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[72] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[73] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[74] Xiaofeng Wang, Kang Zhao, Feng Liu, Jiayu Wang, Guosheng Zhao, Xiaoyi Bao, Zheng Zhu, Yingya Zhang, and Xingang Wang. Egovid-5m: A large-scale video-action dataset for egocentric video generation. *arXiv preprint arXiv:2411.08380*, 2024.

[75] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20270–20281, 2023.

[76] Yating Wang, Haoyi Zhu, Mingyu Liu, Jiange Yang, Hao-Shu Fang, and Tong He. Vq-vla: Improving vision-language-action models via scaling vector-quantized action tokenizers. *arXiv preprint arXiv:2507.01016*, 2025.

[77] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[78] Bernard L Welch. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[79] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*, 2025.

[80] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*, 2025.

[81] Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation. *arXiv preprint arXiv:2505.21864*, 2025.

[82] Mingxing Xu, Wenrui Dai, Chunmiao Liu, Xing Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint arXiv:2001.02908*, 2020.

[83] Zhenjia Xu, Zhou Xian, Xingyu Lin, Cheng Chi, Zhiao Huang, Chuang Gan, and Shuran Song. Roboninja: Learning an adaptive cutting policy for multi-material objects. *arXiv preprint arXiv:2302.11553*, 2023.

[84] Ganlin Yang, Tianyi Zhang, Haoran Hao, Weiyun Wang, Yibin Liu, Dehui Wang, Guanzhou Chen, Zijian Cai, Junting Chen, Weijie Su, et al. Vlaser: Vision-language-action model with synergistic embodied reasoning. *arXiv preprint arXiv:2510.11027*, 2025.

[85] Jonathan Yang, Dorsa Sadigh, and Chelsea Finn. Polybot: Training one policy across robots while embracing variability. *arXiv preprint arXiv:2307.03719*, 2023.

[86] Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, Quan Vuong, Chelsea Finn, Dorsa Sadigh, and Sergey Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. *arXiv preprint arXiv:2402.19432*, 2024.

[87] Shuai Yang, Hao Li, Yilun Chen, Bin Wang, Yang Tian, Tai Wang, Hanqing Wang, Feng Zhao, Yiyi Liao, and Jiangmiao Pang. Instructvla: Vision-language-action instruction tuning from understanding to manipulation. *arXiv preprint arXiv:2507.17520*, 2025.

[88] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.

[89] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu.

Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14022–14032, 2024.

[90] Chengbo Yuan, Rui Zhou, Mengzhen Liu, Yingdong Hu, Shengjie Wang, Li Yi, Chuan Wen, Shanghang Zhang, and Yang Gao. Motiontrans: Human vr data enable motion-level learning for robotic manipulation policies. *arXiv preprint arXiv:2509.17759*, 2025.

[91] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

[92] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

[93] Andy Zhai, Brae Liu, Bruno Fang, Chalse Cai, Ellie Ma, Ethan Yin, Hao Wang, Hugo Zhou, James Wang, Lights Shi, et al. Igniting vlms toward the embodied space. *arXiv preprint arXiv:2509.11766*, 2025.

[94] Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11142–11152, 2025.

[95] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

[96] Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025.

[97] Xueyang Zhou, Yangming Xu, Guiyao Tie, Yongchao Chen, Guowen Zhang, Duanfeng Chu, Pan Zhou, and Lichao Sun. Libero-pro: Towards robust and fair evaluation of vision-language-action models beyond memorization. *arXiv preprint arXiv:2510.03827*, 2025.

[98] Zhongyi Zhou, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. Chatvla-2: Vision-language-action model with open-world embodied reasoning from pretrained knowledge. *arXiv preprint arXiv:2505.21906*, 3, 2025.

[99] Zhongyi Zhou, Yichen Zhu, Minjie Zhu, Junjie Wen, Ning Liu, Zhiyuan Xu, Weibin Meng, Yaxin Peng, Chaomin Shen, Feifei Feng, et al. Chatvla: Unified multimodal understanding and robot control with vision-language-action model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5377–5395, 2025.

[100] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.