# EmbodiedMidtrain: Bridging the Gap between Vision-Language Models and Vision-Language-Action Models via Mid-training

Anonymous CVPR submission

Paper ID *****

## Abstract

*Vision-Language-Action Models (VLAs) inherit their visual and linguistic capabilities from Vision-Language Models (VLMs), yet most VLAs are built from off-the-shelf VLMs that are not adapted to the embodied domain, limiting their downstream performance. In this work, we propose EmbodiedMidtrain to bridge the gap between VLMs and VLAs. We first characterize the data distribution gap between them, showing that VLA data occupy compact regions that are largely separated from the broader VLM distribution, while the degree of alignment varies substantially both across and within VLM data sources. Then, we build a mid-training data engine that leverages a lightweight learnable proximity estimator to select the most VLA-aligned candidates from a large VLM pool, and mid-trains the VLM on this curated mixture before downstream VLA fine-tuning. Experiments on three robot manipulation benchmarks show that mid-training consistently improves performance across different VLM backbones, achieving results competitive with expert VLAs and off-the-shelf VLMs trained with larger model scale and training budgets. Further analysis reveals that mid-training provides a stronger initialization for VLA fine-tuning, with gains emerging from the earliest steps and widening throughout training. Moreover, the data engine captures both dataset-level and sample-level alignment signals, favoring spatial reasoning over text-centric tasks while preserving the diversity of the VLM data. We will release all code, data and models for future research. This paper is currently under review at COLM 2026.*

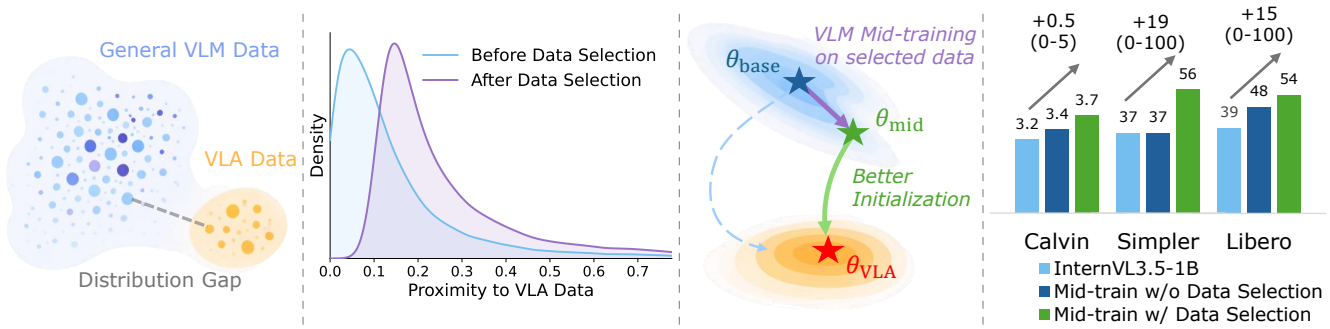
## 1. Introduction

Recent advances in robotics foundation models have given rise to Vision-Language-Action Models (VLAs), which unify vision perception, language understanding, and action generation within a single model to enable generalist robot control across diverse tasks, environments, and embodiments [5, 14, 19]. Most VLAs leverage Vision-Language Models (VLMs) as their backbone, inheriting rich visual and linguistic representations that facilitate rapid adaptation to embodied settings [20, 32, 35].

Despite this progress, a fundamental gap exists: most VLAs initialize from general-purpose *off-the-shelf* VLMs that are not tailored toward embodied action generation. VLM pretraining [1, 23, 29] covers broad vision-language tasks such as captioning, visual question answering, and document understanding, whereas VLA training operates on robotic manipulation trajectories grounded in physical interaction. This distributional mismatch means that, even when a VLM backbone provides a strong initialization for language and visual understanding, its internal representations may not be well-suited for the embodied reasoning that effective action generation demands [47]. Bridging this gap remains an open challenge: simply fine-tuning VLM on curated embodied data does not reliably translate into better VLA performance [50], indicating the need for more principled VLM adaptation to the VLA domain.

In this work, we present *EmbodiedMidtrain*, a mid-training framework that aligns VLM data distribution with the VLA domain, enabling VLAs to truly benefit from the abilities of their VLM backbones. We begin with a deep analysis of the VLM-VLA gap and identify a core issue (Figure 1a): the two domains exhibit a significant data distribution discrepancy. VLA data form compact clusters that are largely separated from the broad, diverse VLM distributions [45]; yet this gap is not uniform, as some VLM samples are inherently more aligned with the VLA domain than others. This reveals that the bulk of VLM data concentrates on capabilities distant from VLA data, leaving downstream VLA fine-tuning to bridge a wide representational gap from a misaligned starting point.

Leveraging this understanding, we develop a mid-training data engine that **reshapes the VLM training distribution toward the VLA domain** before fine-tuning begins (Figure 1b) to bridge the gap. The key idea is that, rather than training on all available VLM data indiscriminately, we can *select* the subset that is most aligned with the target VLA distribution and use it to mid-train the VLM, producing a




(a) VLM-VLA Data Distribution Gap (b) VLM Data Distribution Shift (c) VLM Mid-training (d) Downstream VLA Gains

Figure 1. Overview of *EmbodiedMidtrain*. We analyze the data distribution gap between VLMs and VLAs, and select VLM samples with higher proximity to the VLA domain for mid-training, yielding a stronger initialization for downstream VLA fine-tuning.

stronger initialization for downstream VLA learning. Concretely, we propose a lightweight proximity estimator: a learnable classifier on frozen VLM features that learns to distinguish VLA data from VLM data, and whose predicted scores serve as a continuous measure of each VLM sample's closeness to the VLA domain. Top-scoring samples are then assembled into a curated mid-training mixture, and the VLM is mid-trained on this distribution-aligned dataset to serve as a better initialization for VLA fine-tuning (Figure 1c). This pipeline is lightweight, scalable, and requires no architectural changes to the underlying VLM or VLA.

Experiments on Calvin ABC-D [31], SimplerEnv Bridge [38], and Libero-10 [27] show that our mid-trained 1.1B model achieves a significant performance boost (Figure 1d). Beyond raw performance gain, our analysis reveals several insights into what makes mid-training effective. First, the same curated data transfers across architectures: applying the mixture selected with InternVL3.5-1B to Qwen3VL-2B yields consistent gains, suggesting that proximity-based selection captures transferable properties across VLMs. Second, the learned proximity estimator substantially outperforms hand-crafted alternatives like feature-space distance and perplexity-based metrics. Third, training dynamics analysis shows that the mid-trained model outperforms the baseline from the earliest fine-tuning checkpoints and the gap widens over time, indicating mid-training provides a fundamentally better initialization rather than a transient head start.

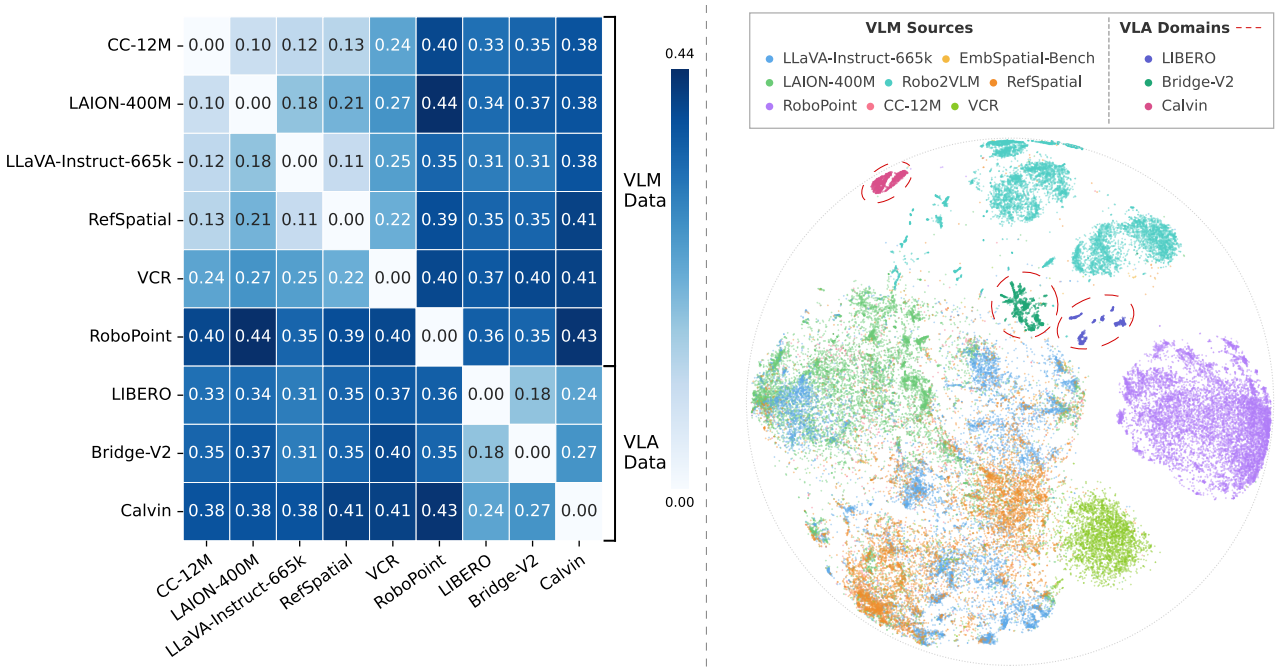
Our main contributions are as follows:

- **A proximity-based mid-training pipeline to bridge the VLM-VLA gap.** We propose *EmbodiedMidtrain*, a mid-training data engine that learns a proximity estimator to score VLM samples by their closeness to the VLA domain, and selects the top-ranked candidates to curate a distribution-aligned mid-training mixture.
- **Consistent VLA performance gains across benchmarks and backbones.** Extensive experiments on Calvin ABC-D, SimplerEnv Bridge, and Libero-10 demonstrate that our mid-trained VLM initialization yields large and consistent improvements over the original backbone, achieving results competitive with substantially larger models.
- **Insights on better VLM initialization for VLAs.** We show that mid-training produces an initialization whose advantages emerge from the earliest fine-tuning steps and amplify over time. We further demonstrate that the learned proximity estimation outperforms hand-crafted alternatives by capturing fine-grained alignment with the VLA domain.

## 2. Related work

**Vision-Language-Action Models.** VLAs extend foundation models like VLMs to generate robot actions, leveraging the broad visual and language understanding inherited from VLMs [5, 19, 32]. Existing designs differ in backbone choice and action generation mechanism. Token-based approaches such as OpenVLA [19] discretize robot actions into tokens for autoregressive generation; OpenVLA-OFT [20] further introduces parallel decoding with learnable action embeddings for efficient continuous action prediction. Flow-matching and diffusion-based approaches such as $\pi_0$ [5] and $\pi_{0.5}$ [35] instead adopt a VLM backbone (PaliGemma [4]) paired with a continuous action generation head. More recent models such as GR00T N1 [32] employ a VLM backbone (Eagle-2 [26]) with a dedicated cross-attention action decoder. While these designs differ architecturally, a common thread is that the VLM backbone is taken *off-the-shelf* from general-purpose training without specific preparation for the embodied domain, which brings a gap that our work identifies and addresses.

**VLM Mid-training.** Many VLMs adopt multi-stage training pipelines, where the model is further trained on curated vision-language data after initial multimodal pretraining or alignment and before final instruction tuning [1, 10, 40]. In this paper, we refer to this intermediate stage as *VLM*




(a) Pairwise MMD distance matrix between VLM and VLA datasets

(b) t-SNE visualization of visual feature distributions across VLM and VLA datasets

Figure 2. **Distribution analysis of VLM and VLA datasets.** (a) Pairwise MMD distances quantify this distribution gap, with cross-group distances larger than within-group distances. (b) VLA datasets form more compact and concentrated clusters that are separated from the broader and more dispersed VLM distributions.

*mid-training*. Similar to mid-training in LLMs, its goal is to adapt a foundation model toward desired domains or capabilities prior to final post-training or task-specific fine-tuning [15, 17, 33, 43]. Our work studies this paradigm in the embodied setting, where mid-training is used to bridge the data distribution gap between general VLM pretraining data and VLA fine-tuning data.

**Embodied-oriented VLMs.** A broad line of work aims to improve VLMs' embodied capabilities, through both embodied-oriented dataset construction [8, 11, 48, 52] and model-level adaptation [6, 9, 18]. Despite these efforts, Zhang et al. [50] show that these gains on finetuning VLMs on embodied tasks do not reliably transfer to downstream VLA task performance, suggesting that current embodied VLM fine-tuning captures a different signal from what VLA execution requires. Alternatively, Vlaser [47] converts in-domain robot trajectories into VQA pairs for VLM fine-tuning. Collectively, these efforts highlight that bridging VLMs and VLAs remains an open problem: existing approaches either improve VLM-side embodied benchmarks without reliably translating into better VLA performance, or require large amounts of in-domain robot data [21, 47, 50]. Our work takes a complementary perspective by operating on the diverse and abundant VLM data rather than crafting task-specific corpora.

## 3. Data distribution gap between VLMs and VLAs

Since most VLAs inherit their visual and linguistic representations from VLM pretraining, the quality of this initialization is fundamentally shaped by the data the VLM was trained on. This motivates us to examine the gap between VLMs and VLAs from the perspective of their training data distributions, where we analyze VLM and VLA data in a shared representation space and characterize the gap with both qualitative and quantitative measurements.

To represent VLM and VLA data in a unified distribution space, we extract the last hidden states of a VLM [42] as the feature representation $h(\cdot)$ for each data sample. We first quantify the distance between each pair of datasets with **Maximum Mean Discrepancy** (MMD). Formally, the squared MMD between two datasets $P$ and $Q$ is:

$$\mathrm{MMD}^2(P,Q) = \mathbb{E}_{x,x'\sim P}[k(x,x')] + \mathbb{E}_{y,y'\sim Q}[k(y,y')] - 2\,\mathbb{E}_{x\sim P,\, y\sim Q}[k(x,y)]$$

where $k(x,y) = \exp\!\big(-\|h(x)-h(y)\|_2^2 \,/\, 2\sigma^2\big)$ is a Gaussian RBF kernel applied over the extracted features, with bandwidth $\sigma$ set via the median heuristic [16]. Figure 2a reports the globally normalized pairwise MMD scores across all dataset pairs. We further visualize these features using

t-SNE in Figure 2b.

**The characteristics of data distribution gap between VLM and VLA.** The distribution space of VLM data and VLA data is largely separated, with only a few near-neighbors. Figure 2a shows that MMD distances are generally smaller within the VLM group and within the VLA group than across the two groups, quantitatively confirming a clear distributional mismatch. Figure 2b further illustrates this pattern: VLA datasets form compact clusters that are mostly detached from the main regions occupied by VLM datasets, while only a small subset of the VLM data lies nearby. These results reveal a clear distribution mismatch between VLM data and VLA data that could affect the performance of downstream embodied tasks.

**The data distribution gap exhibits substantial internal heterogeneity.** Although VLM and VLA data are separated overall, their mismatch is not uniform within each dataset. Figure 2b shows that some VLM sources lie noticeably closer to VLA domains than others, with a few local regions exhibiting clear cross-domain proximity despite the global separation. This suggests that the gap between VLM and VLA is better characterized as a spectrum of alignment rather than a binary distinction, revealing substantial heterogeneity in how different parts of the VLM distribution relate to VLA data.

These insights suggest that **enhancing VLMs for embodied tasks requires reshaping the training data distribution toward the VLA domain**. Realizing this goal requires more than coarse dataset-level mixture adjustment; it calls for sample-wise selection within each dataset. This motivates a data-centric mid-training strategy that explicitly prioritizes VLM samples that are most compatible with the target embodied domain.

# 4. Data engine for EmbodiedMidtrain

As shown in Section 3, the full VLM candidate pool contains a large proportion of samples far from the VLA distribution, so mid-training on it indiscriminately may dilute the embodied signal most useful for downstream VLA adaptation. To address this, we propose a mid-training data engine with two considerations: (1) the data mixture should span both general and embodied-oriented VLM sources to preserve diversity, and (2) selection should operate at the *sample* level rather than the dataset level, since even within the same source individual samples vary substantially in their relevance to VLA tasks.

**Proximity-Based Data Selection.** Inspired by target-guided data selection for LLM pretraining [44], we select the VLM subset whose distribution best aligns with the VLA domain. Our core idea is to frame selection as a domain-membership problem: we train a lightweight binary classifier on frozen VLM features to score each candidate sample by how closely it resembles VLA data. After scoring, we rank all VLM samples and retain only the highest-scoring subset for mid-training, yielding a curated corpus that preserves the diversity of the original pool while concentrating on the most VLA-compatible samples.

Let $\mathcal{D}_{\text{VLM}}$ and $\mathcal{D}_{\text{VLA}}$ denote the candidate VLM pool and target VLA corpus, with densities $p_{\text{VLM}}$ and $p_{\text{VLA}}$ over a shared representation space. Our goal is to select a size-$K$ subset whose distribution best aligns with VLA data:

$$\mathcal{D}^*_{\text{VLM}} = \underset{\substack{\mathcal{D}' \subseteq \mathcal{D}_{\text{VLM}} \\ |\mathcal{D}'|=K}}{\operatorname{argmin}} \; d(P_{\mathcal{D}'},\, P_{\text{VLA}}) \tag{1}$$

where $d$ is a distributional divergence. Solving this exactly is intractable, so we relax it to per-sample scoring and top-$K$ selection:

$$\mathcal{D}^*_{\text{VLM}} = \text{top-K}_{x_i \in \mathcal{D}_{\text{VLM}}} \; s(x_i) \tag{2}$$

The key question is how to define the scoring function $s$. A natural choice is the density ratio $p_{\text{VLA}}(x)/p_{\text{VLM}}(x)$, which measures how likely a sample is under the VLA distribution relative to the VLM distribution. However, estimating this ratio directly in high-dimensional feature spaces is difficult. We instead leverage a classical result from density ratio estimation [13]: a binary classifier trained to distinguish two distributions recovers this ratio at optimality:

$$s^*(x) = \frac{p_{\text{VLA}}(x)}{p_{\text{VLA}}(x) + p_{\text{VLM}}(x)} \tag{3}$$

Since $s^*$ is monotonically increasing in the density ratio, ranking by the classifier output is equivalent to ranking by the density ratio.

We instantiate this as a lightweight **proximity estimator** on frozen VLM features, so that scoring is efficient and decoupled from mid-training itself. The estimator applies a learnable scoring function $f$ on top of the frozen VLM's last hidden state $\phi(\mathbf{x})$, followed by a sigmoid:

$$s(\mathbf{x}) = \sigma\big(f(\phi(\mathbf{x}))\big) \tag{4}$$

We train with VLA samples as positives and VLM samples as negatives, using binary cross-entropy loss as the training objective:

$$\begin{aligned}\mathcal{L}_{\text{cls}} = & -\mathbb{E}_{y \sim \mathcal{D}_{\text{VLA}}}\left[\log s(y)\right] \\ & -\mathbb{E}_{x \sim \mathcal{D}_{\text{VLM}}}\left[\log\big(1 - s(x)\big)\right]\end{aligned} \tag{5}$$

After training, we rank all candidate VLM samples by $s(x)$ and retain the top-$K$ to form $\mathcal{D}^*_{\text{VLM}}$ for mid-training. This procedure turns a broad candidate pool into a more targeted corpus for embodied adaptation, preserving the useful diversity of VLM data while shifting the training distribution toward the VLA domain. We use this curated subset in the subsequent mid-training stage of *EmbodiedMidtrain*.

# 5. Experiments

We conduct experiments to demonstrate that the proposed mid-training data engine bridges the VLM-VLA gap. In particular, we show that mid-training with proximity-based data selection consistently improves multiple VLM backbones on downstream VLA tasks, and that the curated data mixture exhibits cross-backbone transferability.

## 5.1. Setup

**VLM data source.** To construct the candidate pool for data selection, we collect a diverse set of VLM datasets spanning both general-purpose and embodied-oriented sources. For **general VLM data**, following Qwen-VL [1], we include image-captioning data from subsets of LAION-400M [36], CC-12M [7] with captions relabeled by BLIP [22], instruction-following data from LLaVA-Instruct-665k [28], and Visual Commonsense Reasoning (VCR) [49]. For **embodied-oriented VLM data**, we additionally include RefSpatial [52] for spatial referring and reasoning, EmbSpatial-Bench [11] for embodied spatial understanding, Robo2VLM [8] for robotic visual question answering, and RoboPoint [48] for spatial affordance prediction.

**VLM mid-training.** We first construct the mid-training dataset using the data engine described in Section 4. We leverage a balanced mixture of training data in VLA fine-tuning as the target VLA data for proximity estimator learning. To prevent overfitting, we apply early stopping at 90% validation accuracy. The resulting proximity estimator scores and selects candidate VLM samples to conduct mid-training on InternVL3.5-1B [42] and Qwen3VL-2B [2] using the LLaMA-Factory [51] framework. We train all model parameters with a global batch size of 256 for 5,000 steps. See Appendix A for more implementation details.

**VLA fine-tuning.** After VLM mid-training, we initialize a VLA from the resulting VLMs and fine-tune it following the VLA training pipeline of VLM4VLA [50], where the VLM backbone is cascaded with a two-branch MLP action decoder that predicts continuous arm actions and binary gripper actions. This architecture is designed to be generic across different VLM backbones and capable of fully leveraging each VLM's intrinsic knowledge. The architecture and training details are provided in Appendix B.

**Evaluation benchmarks.** We evaluate on three simulated manipulation benchmarks that cover multiple aspects of embodied control, following VLM4VLA's evaluation protocol for fair and reproducible comparison. **Calvin ABC-D** [31] trains on the ABC splits and evaluates on the unseen split D over 1,000 five-subtask sequences, testing generalization to novel scene configurations. **SimplerEnv Bridge** [24, 38] is a real-to-sim benchmark with four tabletop manipulation tasks and 24 randomized trials each, reporting mean success rate. **LIBERO-10** [27], the most challenging suite in the LIBERO benchmark, comprises 10 long-horizon tasks evaluated on 50 trials per task.

## 5.2. Baselines

We compare our method against two categories of baselines: expert VLA models and VLAs trained from off-the-shelf VLMs. These baselines are either experimented or reproduced by VLM4VLA. For fair comparison, we also report model size and training budgets, measured by the total number of samples seen. This allows us to assess not only downstream performance, but also the efficiency of our approach, showing that VLM mid-training with proximity-based data selection can achieve competitive results with substantially fewer training resources. Additional implementation details are provided in the Appendix B.

**Expert VLA Baselines.** We first compare with representative expert VLA models, including OpenVLA [19] and $\pi_0$ [5]. OpenVLA is built on Llama-2-7B with DINOv2 and SigLIP visual encoders, and models robot control by autoregressively predicting discretized action tokens. $\pi_0$ is based on Paligemma-1 VLM and models continuous robot actions using flow matching.

**Off-the-shelf VLM Baselines.** To evaluate how pretrained VLM abilities transfer to downstream action learning, we further consider VLAs directly finetuned from a diverse set of off-the-shelf VLMs spanning different architectures and scales. These VLMs include the Qwen2.5VL and Qwen3VL family [2, 3], which represents a strong line of general-purpose open-source VLMs; the Paligemma family (Paligemma-1 [4] and Paligemma-2 [37]), which is widely adopted for downstream adaptation; and KosMos-2 [34], which provides a grounding-oriented alternative.

## 5.3. Main results

Table 1 presents the main results. With proximity-based mid-training, our models achieve consistent and substantial improvements across all three benchmarks, demonstrating that a well-curated mid-training mixture can significantly strengthen a VLM's readiness for downstream action learning. More details are provided in Appendix C.

**Competitive performance at a fraction of the scale.** Despite being the smallest model in the comparison, the mid-trained InternVL3.5-1B surpasses both expert VLA baselines on Calvin ABC-D and outperforms several off-the-shelf VLMs that are 3–8$\times$ larger, including Paligemma-1, Paligemma-2, and KosMos-2. On SimplerEnv-Bridge and Libero-10, it reaches performance on par with much larger VLMs such as the Qwen family, while using only a fraction of their training budgets. These results suggest that what matters most is not the volume of pretraining data a VLM has seen, but how well the mid-training data aligns with the downstream embodied distribution.

**Cross-backbone transferability.** To examine the generality of our data engine beyond a single backbone, we apply the

| Model | Size | # Samples Seen (Calvin / Simpler / Libero) | Calvin (Tasks Completed in a Row) 1 | 2 | 3 | 4 | 5 | Avg. Len.↑ | Simpler↑ | Libero↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| **Expert VLA Baselines*** | | | | | | | | | | |
| OpenVLA (Llama-2) | 7.7B | 7.7M / 25.6M / 25.6M | 0.792 | 0.644 | 0.499 | 0.368 | 0.245 | 2.548 | 4.2 | 53.7 |
| $\pi_0$ (Paligemma-1) | 3.1B | 7.7M / 25.6M / 25.6M | 0.896 | 0.785 | 0.786 | 0.610 | 0.532 | 3.509 | 60.4 | 46.0 |
| **Off-the-shelf VLM Baselines*** | | | | | | | | | | |
| Qwen2.5VL-3B | 3.8B | 7.7M / 25.6M / 25.6M | 0.922 | 0.842 | 0.766 | 0.700 | 0.626 | 3.856 | 48.0 | 43.0 |
| Qwen2.5VL-7B | 8.3B | 7.7M / 25.6M / 25.6M | 0.935 | 0.864 | 0.807 | 0.758 | 0.693 | 4.057 | 46.8 | 45.0 |
| Qwen3VL-2B | 2.1B | 7.7M / 25.6M / 25.6M | 0.943 | 0.882 | 0.831 | 0.776 | 0.710 | 4.142 | 49.0 | 55.8 |
| Qwen3VL-4B | 4.4B | 7.7M / 25.6M / 25.6M | 0.933 | 0.857 | 0.790 | 0.719 | 0.644 | 3.943 | 56.3 | 44.4 |
| Qwen3VL-8B | 8.8B | 7.7M / 25.6M / 25.6M | 0.940 | 0.868 | 0.797 | 0.746 | 0.684 | 4.035 | 58.3 | 46.2 |
| Qwen3VL-30B-A3B | 30B-A3B | 7.7M / 25.6M / 25.6M | 0.939 | 0.877 | 0.820 | 0.757 | 0.682 | 4.075 | 44.8 | 46.8 |
| Paligemma-1 | 2.9B | 7.7M / 25.6M / 25.6M | 0.914 | 0.813 | 0.692 | 0.599 | 0.488 | 3.506 | 55.3 | 44.2 |
| Paligemma-2 | 3.0B | 7.7M / 25.6M / 25.6M | 0.901 | 0.775 | 0.669 | 0.575 | 0.486 | 3.406 | 57.3 | 46.2 |
| KosMos-2 | 1.7B | 7.7M / 25.6M / 25.6M | 0.878 | 0.721 | 0.591 | 0.498 | 0.408 | 3.096 | 60.4 | 55.0 |
| **VLM with *EmbodiedMidtrain* (Ours)** | | | | | | | | | | |
| InternVL3.5-1B | 1.1B | 1.0M / 4.1M / 4.1M | 0.909 | 0.754 | 0.606 | 0.498 | 0.406 | 3.173 | 36.5 | 39.0 |
| + *EmbodiedMidtrain* | 1.1B | 1.0M / 4.1M / 4.1M | 0.935 | 0.838 | 0.737 | 0.653 | 0.551 | **3.714** | **56.3** | **54.2** |
| Qwen3VL-2B | 2.1B | 1.0M / 4.1M / 4.1M | 0.887 | 0.747 | 0.612 | 0.527 | 0.432 | 3.205 | 38.5 | 33.8 |
| + *EmbodiedMidtrain* | 2.1B | 1.0M / 4.1M / 4.1M | 0.922 | 0.808 | 0.700 | 0.623 | 0.533 | 3.584 | 45.8 | 40.2 |

Table 1. Main results across Calvin ABC-D, SimplerEnv-Bridge, and Libero-10. **# Samples Seen** is reported as the training budgets on *Calvin / SimplerEnv / Libero*. * Results for Expert VLA Baselines and Off-the-shelf VLM Baselines are reported by VLM4VLA.

same mid-training data, selected using InternVL3.5-1B's feature space, to Qwen3VL-2B. Despite being curated with representations from a different model, the selected data also yield clear gains for Qwen3VL-2B across all three benchmarks after mid-training. This suggests that proximity-based selection captures distributional properties that are not specific to a single backbone, but instead reflect a more general alignment with embodied VLA tasks.

## 6. Analysis

Having established the effectiveness of our approach, we now examine its design choices and behavior in detail. We first ablate two core components of our data engine: the effect of data selection versus random sampling, and the choice of proximity measurement (Section 6.1). We then analyze the training dynamics of VLA fine-tuning to understand *when* the benefits of mid-training emerge (Section 6.2), and finally inspect the selected data to understand *what* the proximity estimator learns to prefer (Section 6.3).

### 6.1. Ablations

We ablate two central design choices in our data engine: the advantage of proximity-based selection over random sampling, and the effectiveness of different proximity measurements. Table 2 summarizes the results.

**Random selection.** Randomly sampling from the candidate pool consistently underperforms our learned estimator on all three benchmarks, showing that naive mid-training on unfiltered data is insufficient to bridge the distribution gap. This suggests that the gains do not come merely from additional mid-training, but from identifying and retaining the subset of VLM data that is better aligned with the VLA domain. Proximity-based selection is therefore critical to unlocking the benefit of mid-training.

| Setting | Calvin↑ | Simpler↑ | Libero↑ |
|---|---|---|---|
| Random Selection | 3.398 | 43.8 | 48.4 |
| *Proximity Measurements* | | | |
| Feat.-space Avg. Dist. | 3.126 | 53.1 | 51.2 |
| VLA-cond. Perplexity | 3.159 | 55.2 | 48.0 |
| Delta Perplexity | 1.527 | 39.6 | 54.2 |
| Learned Estimator (*Ours*) | **3.714** | **56.3** | **54.2** |

Table 2. Ablation results for random selection and different proximity measurements on mid-training InternVL3.5-1B backbone.

**Proximity measurements.** Beyond our learned proximity estimator, we evaluate three alternatives that capture VLM-VLA proximity through different hand-crafted signals.

*Feature-space average distance.* For each candidate VLM sample $x$, we compute its average $\ell_2$ distance to all VLA samples in the frozen VLM representation space:

$$d_{\text{avg}}(x) = \frac{1}{|\mathcal{D}_{\text{VLA}}|} \sum_{y \in \mathcal{D}_{\text{VLA}}} \|\phi(x) - \phi(y)\|_2 , \qquad (6)$$

where $\phi(\cdot)$ denotes the last hidden state of the frozen VLM, following the same feature space used in Section 3. Samples with smaller $d_{\text{avg}}(x)$ are ranked as closer to the VLA distribution and selected first.

*VLA-conditioned perplexity.* We fine-tune the VLM on converted in-domain VLA data in which robot actions are represented as text tokens, yielding a VLA-conditioned model

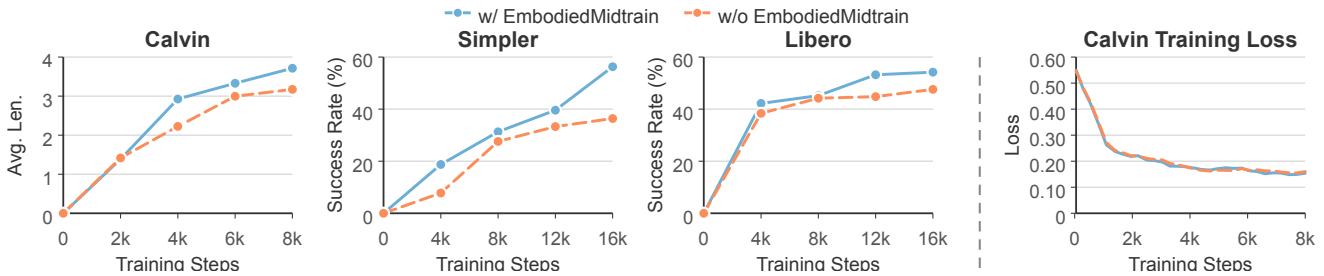



Figure 3. The training dynamics across VLA tasks for VLMs with and without *EmbodiedMidtrain*, including downstream VLA task performances (left) and training loss (right).

with parameters $\theta_{\text{VLA}}$. The perplexity of a candidate sample $x = (x_1, \ldots, x_T)$ under this model is

$$\text{PPL}_{\text{VLA}}(x) = \exp\left( -\frac{1}{T} \sum_{t=1}^{T} \log p_{\theta_{\text{VLA}}}(x_t \mid x_{<t}) \right). \quad (7)$$

Samples with lower $\text{PPL}_{\text{VLA}}(x)$ are considered more compatible with the VLA domain and selected first.

*Delta perplexity.* Delta perplexity measures the perplexity change induced by VLA fine-tuning,

$$\Delta\text{PPL}(x) = \text{PPL}_{\text{VLA}}(x) - \text{PPL}_{\text{VLM}}(x), \quad (8)$$

where $\text{PPL}_{\text{VLM}}(x)$ is the perplexity under the original pre-trained VLM. A more negative $\Delta\text{PPL}(x)$ indicates that the sample becomes more predictable after VLA adaptation, suggesting stronger alignment with the VLA domain. Samples are ranked by $\Delta\text{PPL}(x)$ in ascending order for selection.

All three alternatives are less consistent than our learned estimator, which confirms that hand-crafted metrics only partially capture VLA alignment, whereas the learned estimator provides a more robust and transferable signal.

## 6.2. Training dynamics

Figure 3 compares VLA fine-tuning trajectories from the original InternVL3.5-1B backbone and our mid-trained variant, evaluated at intermediate checkpoints throughout training. The mid-trained model already achieves higher performance in the early stages of fine-tuning, providing direct evidence that **proximity-based mid-training yields a better initialization for VLA learning**. Moreover, this advantage is not a transient head start, since the mid-trained model consistently outperforms the baseline throughout the entire training trajectory, and the gap widens rather than narrowing over time. Notably, this difference is not clearly reflected in training loss, which remains highly similar across the two initializations. This suggests that training loss alone does not fully capture the quality of the learned initialization. Together with the consistently stronger downstream performance of the mid-trained model, these results indicate that the benefits of mid-training persist throughout VLA fine-tuning and translate into a backbone better suited to VLA tasks.

## 6.3. Analysis of selected VLM data

Figure 4a presents the proximity score distributions across eight candidate VLM datasets, visualized as violin plots. While all datasets concentrate in the low-to-moderate score range, the distribution shapes vary noticeably across datasets. Among them, RefSpatial achieves the highest average scores while VCR receives the lowest, indicating that the estimator assigns clear *dataset-level* preferences. At the same time, the within-dataset score spread shows that the estimator also performs fine-grained *sample-level* selection, retaining only the most VLA-aligned samples even from high-scoring datasets.

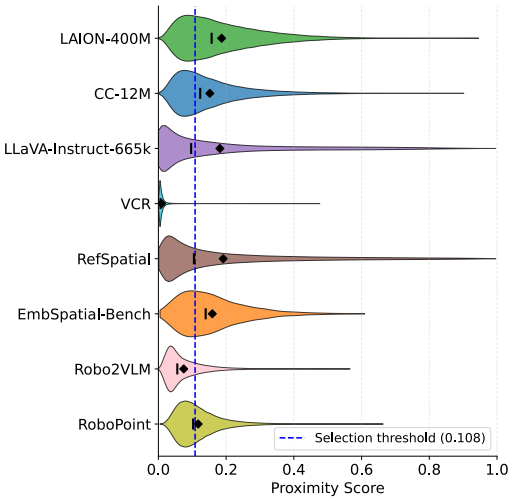
Figures 4b and 4c illustrate this with representative examples. The high-scoring sample requires grounding spatial references and exercising spatial reasoning crucial to embodied manipulation. The low-scoring sample involves recognizing text on a book cover, bearing little relevance to embodied tasks. This suggests that the proximity estimator learns to distinguish visual reasoning patterns that transfer to embodied control from those that do not, without requiring manual domain expertise.

The overall proximity score distribution before and after data selection is presented in Figure 5. Compared with the full candidate pool, the selected subset exhibits a distribution shift toward higher proximity scores, with part of the low-score mass removed, showing that the reshaped data distribution is more aligned with the VLA domain. Further analysis of dataset composition is provided in Appendix D.

## 6.4. Diversity preservation after data selection

We further verify that our proximity-based selection preserves the diversity of the original VLM dataset rather than collapsing the mid-training corpus onto a narrow region near the VLA domain. To quantify the diversity of a data subset, we follow Xing et al. [45] to adopt the *uniformity* metric proposed by Wang and Isola [41]. Formally, given a dataset $D_i$, we define its diversity as

$$S_{\text{diversity}}^{D_i} \triangleq \frac{1}{\mathbb{E}_{u,v \sim D_i}\left[e^{-t\|u-v\|_2^2}\right]}, \quad (9)$$

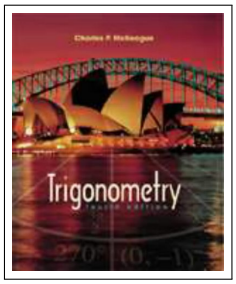



(a) Per-dataset proximity score

**Q:** *You are standing at the point marked by the coordinate point at point (0.878, 0.780). Which object is directly in front of you?*
**A:** *The white matte truck at lower right.*

**Q:** *Locate a point on the yellow metallic crane at upper right.* ⟨format instructions...⟩
**A:** `[(0.976, 0.244)]`

(b) High-scoring sample




**Q:** *Who wrote this book?*
**A:** *Charles P. McKeague.*

**Q:** *What is the title of this book?*
**A:** *Trigonometry.*

(c) Low-scoring sample

Figure 4. Analysis of proximity-based data selection. (a) Distribution of proximity scores across VLM data sources. (b) A high-scoring sample from RefSpatial requiring spatial grounding and reasoning. (c) A low-scoring sample: a book cover with text-only VQA.

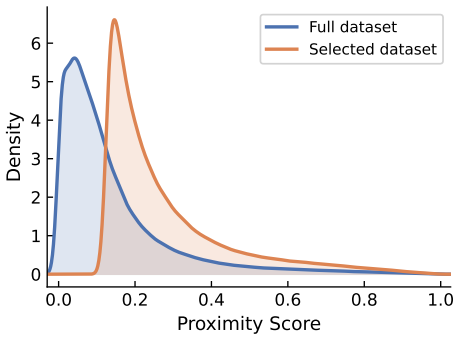



Figure 5. VLM data distribution shift of proximity score.

| Dataset | Diversity |
|---|---|
| VLM Data | 1.85 |
| General VLM Data | 1.96 |
| Embodied-oriented VLM Data | 1.62 |
| VLA Data | 1.26 |
| Selected VLM Data | 1.93 |

Table 3. Diversity of VLM data, VLA data and VLM data selected by EmbodiedMidtrain.

where $u$ and $v$ are feature representations drawn from $D_i$ in the frozen VLM's last hidden state space (following Section 3), and $t$ is a temperature parameter controlling the kernel bandwidth (we apply $t = 2$ in our calculation). A higher $S^{D_i}_{\text{diversity}}$ indicates greater spread among samples, reflecting higher diversity.

Table 3 reports the diversity scores. VLA data is the most concentrated, with a score of 1.26 that is consistent with the compact clusters observed in Section 3, while the general VLM pool is the most diverse at 1.96. Our selected subset reaches 1.93, closely matching the general VLM pool and substantially exceeding both the embodied-oriented subset at 1.62 and the VLA data. Proximity-based selection thus shifts the data distribution toward the VLA domain without collapsing it, retaining the broad visual and linguistic coverage that makes the mixture effective as a mid-training signal.

## 7. Conclusion

We presented *EmbodiedMidtrain*, a mid-training pipeline that bridges the data distribution gap between VLMs and VLAs through proximity-based data selection. Built on frozen VLM features, our lightweight estimator identifies the VLM samples most aligned with the VLA domain and uses them to construct a more effective mid-training set. Across three manipulation benchmarks, this strategy consistently improves performance competitive with models trained at larger scales and training budgets. The selected data also transfer to Qwen3VL-2B, boosting its performance despite being curated with InternVL3.5-1B features.

Our analysis further shows that mid-training provides a stronger initialization for downstream VLA learning: the advantage emerges from the earliest stages of fine-tuning and continues to grow throughout training. We also find that the selected data exhibit meaningful preferences at both the dataset and sample levels, while preserving the diversity of the original VLM dataset. We hope these findings can inform future efforts to better prepare vision-language backbones for embodied intelligence.

# References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2, 5

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 5

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5

[4] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 2, 5

[5] 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. $\pi_0$: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 1, 2, 5

[6] Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024. 3

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 5

[8] Kaiyuan Chen, Shuangyu Xie, Zehan Ma, and Ken Goldberg. Robo2vlm: Visual question answering from large-scale in-the-wild robot manipulation datasets, 2025. 3, 5

[9] Xinyi Chen, Yilun Chen, Yanwei Fu, Ning Gao, Jiaya Jia, Weiyang Jin, Hao Li, Yao Mu, Jiangmiao Pang, Yu Qiao, Yang Tian, Bin Wang, Bolun Wang, Fangjing Wang, Hanqing Wang, Tai Wang, Ziqin Wang, Xueyuan Wei, Chao Wu, Shuai Yang, Jinhui Ye, Junqiu Yu, Jia Zeng, Jingjing Zhang, Jinyu Zhang, Shi Zhang, Feng Zheng, Bowen Zhou, and Yangkun Zhu. Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy, 2025. 3

[10] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. 2

[11] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–355, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3, 5

[12] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024. 14

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3, 2014. 4

[14] GR Team. Gemini robotics 1.5: Pushing the frontier of generalist embodied agents. *arXiv preprint arXiv:2510.03342*, 2025. 1

[15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina

Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil,

Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. 3

[16] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. 3

[17] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024. 3

[18] Yuheng Ji et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:2502.21257*, 2025. 3

[19] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 1, 2, 5

[20] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, 2025. 1, 2

[21] Taeyoung Kim, Jimin Lee, Myungkyu Koo, Dongyoung Kim, Kyungmin Lee, Changyeon Kim, Younggyo Seo, and Jinwoo Shin. Contrastive representation regularization for vision-language-action models, 2025. 3

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 5

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 1

[24] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In *Proceedings of The 8th Conference on Robot Learning*, pages 3193–3215, 2025. 5

[25] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 14

[26] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, Nadine Chang, Karan Sapra, Amala Sanjay Deshmukh, Tuomas Rintamaki, Matthieu Le, Ilia Karmanov, Lukas Voegtle, Philipp Fischer, De-An Huang, Timo Roman, Tong Lu, Jose M. Alvarez, Bryan Catanzaro, Jan Kautz, Andrew Tao, Guilin Liu, and Zhiding Yu. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv:2501.14818*, 2025. 2

[27] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023. 2, 5

[28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 5

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 1

[30] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6924–6934, 2025. 14

[31] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7327–7334, 2022. 2, 5

[32] NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots, 2025. 1, 2

[33] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William

Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. 3

[34] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 5

[35] 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. *arXiv preprint arXiv:2504.16054*, 2025. 1, 2

[36] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114, 2021. 5

[37] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024. 5

[38] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023. 2, 5

[39] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024. 14

[40] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[41] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*. JMLR.org, 2020. 7

[42] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 3, 5

[43] Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*, 2025. 3

[44] Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 4

[45] Youguang Xing, Xu Luo, Junlin Xie, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Shortcut learning in generalist robot policies: The role of dataset diversity and fragmentation. In *Conference on Robot Learning*, pages 3239–3266, 2025. 1, 7

[46] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025. 14

[47] 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. 1, 3

[48] 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 in robotics. In *Proceedings of The 8th Conference on Robot Learning*, pages 4005–4020. PMLR, 2025. 3, 5

[49] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[50] Jianke Zhang, Xiaoyu Chen, Qiuyue Wang, Mingsheng Li, Yanjiang Guo, Yucheng Hu, Jiajun Zhang, Shuai Bai, Junyang Lin, and Jianyu Chen. Vlm4vla: Revisiting vision-language-models in vision-language-action models. *arXiv preprint arXiv:2601.03309*, 2026. 1, 3, 5, 13

[51] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. 5, 13

[52] 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. 3, 5

# A. Implementation Details for VLM Mid-training

We apply a simple-yet-effective learnable scoring function $f(\cdot)$ described in Section 4 as a linear layer on top of frozen VLM representations. For proximity estimator training, we apply a batch size of 128. The training data are sampled in a balanced manner from the VLM candidate pool and the target VLA data, as described in Section 5.1. Due to early-stopping, our binary classifier training usually stops at 75 to 100 steps. We then run inference with this proximity estimator on all VLM data to select a subset of top 1280k samples by proximity score for further VLM training.

In VLM mid-training, we leverage the LLaMA-Factory [51] as the training framework. We perform full-parameter supervised fine-tuning on InternVL3.5-1B, unfreezing the vision encoder, multi-modal projector, and language model. Table 4 summarizes the key hyperparameters.

| Configuration | Mid-training |
|---|---|
| VLM backbone | InternVL3.5-1B |
| Trainable parameters | All (full fine-tuning) |
| Sequence length | 1024 |
| Optimizer | AdamW |
| Peak learning rate | $5 \times 10^{-5}$ |
| Learning rate schedule | Cosine decay |
| Warmup ratio | 0.1 |
| Numerical precision | bfloat16 |
| Training steps | 5,000 |
| Per-device batch size | 32 |
| Gradient accumulation | 2 |
| Global batch size | 256 |

Table 4. Key hyperparameters for VLM mid-training.

# B. Implementation Details for VLA Fine-tuning

**Training Setup.** We follow the VLM4VLA adaptation design and evaluation protocol [50] for downstream VLA adaptation. Table 5 summarizes the key training settings for our InternVL3.5-1B-based VLA models. Across all experiments, models take a single-view image as visual input and do not use robot state information, and are fine-tuned end-to-end. Input images are first standardized to $224 \times 224$ and then resized to match the input resolution required by each VLM backbone. For InternVL3.5-1B, the final input resolution is $448 \times 448$.

For the baseline models reported in Table 1 including expert VLAs and off-the-shelf VLMs, we directly use the results from VLM4VLA and do not rerun them in this work. We refer readers to VLM4VLA for the corresponding model-specific training details and hyperparameter settings.

**VLA Architecture.** We follow the VLM4VLA adaptation design to convert a VLM into a VLA. Specifically, the multimodal representation produced by the VLM backbone is mapped to latent control features, which are then decoded by a small MLP-based action head into chunked robot actions. Continuous arm actions are supervised with Huber loss, while binary gripper actions are supervised with binary cross-entropy loss.

| Configuration | VLA Fine-tuning |
|---|---|
| Trainable parameters | All (full fine-tuning) |
| Action chunk size | 10 (Calvin ABC-D)<br>4 (Simpler-Bridge, Libero-10) |
| Optimizer | AdamW |
| Peak learning rate | $5 \times 10^{-5}$ |
| Learning rate schedule | Cosine decay |
| Warmup steps | 500 |
| Numerical precision | bfloat16 |
| Training steps | 16,000 |
| Per-device batch size | 32 |
| Gradient accumulation | 2 |
| Global batch size | 256 |

Table 5. Key hyperparameters for VLA fine-tuning.

# C. Detailed Experimental Results

Table 6 provides full per-scene breakdowns for SimplerEnv-Bridge, corresponding to the aggregate scores reported in Table 1.

# D. Selected Data Mixture Composition

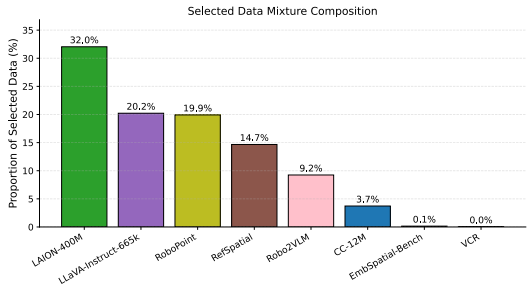



Figure 6. Composition of the selected mid-training data mixture after proximity-based selection, shown as the percentage of each VLM source dataset in the final mixture.

Figure 6 shows the composition of the mid-training mixture after proximity-based data selection. LAION-400M contributes the largest share (32.0%), followed by LLaVA-Instruct-665k (20.2%) and RoboPoint (19.9%). RefSpatial and Robo2VLM account for 14.7% and 9.2% respectively, while CC-12M contributes a small fraction (3.7%).

| Model | Size | # Samples Seen | Carrot | Eggplant | Spoon | Cube | Simpler↑ |
|---|---|---|---|---|---|---|---|
| **Expert VLA Baselines*** | | | | | | | |
| OpenVLA (Llama-2) | 7.7B | 25.6M | 4.2 | 0.0 | 0.0 | 12.5 | 4.2 |
| $\pi_0$ (Paligemma-1) | 3.1B | 25.6M | 62.5 | 100.0 | 54.2 | 25.0 | 60.4 |
| **Off-the-shelf VLM Baselines*** | | | | | | | |
| Qwen2.5VL-3B | 3.8B | 25.6M | 20.8 | 91.7 | 79.2 | 0.0 | 48.0 |
| Qwen2.5VL-7B | 8.3B | 25.6M | 12.5 | 100.0 | 75.0 | 0.0 | 46.8 |
| Qwen3VL-2B | 2.1B | 25.6M | 20.8 | 95.8 | 79.2 | 0.0 | 49.0 |
| Qwen3VL-4B | 4.4B | 25.6M | 54.2 | 95.8 | 75.0 | 0.0 | 56.3 |
| Qwen3VL-8B | 8.8B | 25.6M | 58.3 | 95.8 | 79.2 | 0.0 | 58.3 |
| Qwen3VL-30B-A3B | 30B-A3B | 25.6M | 29.2 | 79.2 | 70.8 | 0.0 | 44.8 |
| Paligemma-1 | 2.9B | 25.6M | 50.0 | 91.7 | 75.0 | 4.2 | 55.3 |
| Paligemma-2 | 3.0B | 25.6M | 75.0 | 75.0 | 79.2 | 0.0 | 57.3 |
| KosMos-2 | 1.7B | 25.6M | 37.5 | 100.0 | 75.0 | 29.2 | 60.4 |
| **VLM with *EmbodiedMidtrain* (Ours)** | | | | | | | |
| InternVL3.5-1B | 1.1B | 4.1M | 0.0 | 100.0 | 45.8 | 0.0 | 36.5 |
| +*EmbodiedMidtrain* | 1.1B | 4.1M | 20.8 | 91.7 | 70.8 | 41.7 | **56.3** |
| Qwen3VL-2B | 2.1B | 4.1M | 37.5 | 75.0 | 41.7 | 0.0 | 38.5 |
| +*EmbodiedMidtrain* | 2.1B | 4.1M | 41.7 | 96.8 | 45.8 | 0.0 | 45.8 |

Table 6. Detailed results on SimplerEnv-Bridge (per-scene success rates: Carrot, Eggplant, Spoon, Cube). **# Samples Seen** reports the SimplerEnv fine-tuning budgets. * Results for Expert VLA Baselines and Off-the-shelf VLM Baselines are reproduced and reported by VLM4VLA.

EmbSpatial-Bench and VCR are nearly absent from the selected mixture (0.1% and 0.0%).

Two patterns are worth noting. First, the dominance of LAION-400M is not because the dataset is uniformly close to the VLA domain, but rather because its sheer scale means that even a small fraction of high-scoring samples translates into a large absolute count. This highlights the value of sample-level selection: the estimator identifies a useful subset from a massive, predominantly out-of-domain source that would otherwise be discarded entirely by dataset-level filtering. Second, the selected mixture reflects a balance between general and embodied-oriented data. Embodied-oriented datasets such as RoboPoint and RefSpatial are retained at much higher rates relative to their original pool sizes, consistent with the score distributions in Figure 4a, yet they do not dominate the mixture. This suggests that the proximity estimator naturally curates a diverse mixture that combines spatial and embodied reasoning from specialized sources with complementary visual knowledge from large-scale general data.

## E. Capability Retention After VLM Mid-training

While our primary goal is to improve VLMs as initializations for downstream VLA adaptation, it is also important to examine how mid-training affects the original capability profile of the VLM itself. To this end, we evaluate the mid-trained model on VLM benchmarks including BLINK [12], POPE [25], VisuLogic [46], 3DSRBench [30] and SpatialEval [39]. Results are reported in Table 7.

| Model | POPE | BLINK | VisuLogic | 3DSRBench | SpatialEval |
|---|---|---|---|---|---|
| InternVL3.5-1B | 86.33 | 43.45 | 21.00 | 47.87 | 49.82 |
| +*EmbodiedMidtrain* | 86.29 | 40.45 | 24.90 | 49.51 | 48.00 |

Table 7. Performance of InternVL3.5-1B on VLM benchmarks before and after mid-training.

Overall, mid-training largely preserves the original VLM capability profile while inducing a selective shift in performance. Performance remains nearly unchanged on POPE, improves on VisuLogic and 3DSRBench, and decreases moderately on BLINK and SpatialEval. This suggests that VLM mid-training does not uniformly maintain all original capabilities, but instead reorients the model toward skills that are more relevant to embodied downstream adaptation.