MUSEVLA: ADAPTIVE MULTIMODAL SENSING FOR ROBOTIC MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans use diverse senses to interact with the world, while most Vision-Language-Action (VLA) models for robotics rely solely on RGB observations. This limits their ability to perceive physical properties that are difficult or impossible to infer from RGB cameras, such as temperature, sound, or radar response. We present MuseVLA, an adaptive multimodal sensing VLA model that integrates novel sensors as on-demand tools for robotic manipulation. Given a task instruction and visual context, MuseVLA first generates a sensor token and target description that select the sensing modality to invoke and what to attend to, analogous to a tool call with arguments. It then converts the selected sensor measurement into a *grounded sensor image*, a unified intermediate representation that encodes heterogeneous readings for multimodal fusion and action generation. This design decouples sensor-specific processing from the VLA backbone, enabling efficient integration of diverse modalities. To reduce costly multisensory data collection, our data synthesis pipeline augments existing RGB videos with grounded sensor images for generalization to unseen sensor-guided tasks. We evaluate MuseVLA on a real-world robot across challenging dexterous hand manipulation tasks that require multimodal sensing inputs, including temperature-guided pick-and-place, audio-driven object search, and radar-assisted hidden object retrieval. MuseVLA achieves 80.6% success rate on average, outperforming RGB-only and multisensory VLA baselines significantly, and exhibits strong zero-shot capabilities on unseen tasks.

1 INTRODUCTION

Humans use vision, sound, and touch selectively, treating sensory modalities as *task-oriented tools* rather than combining every sense at all times. They also extend their perception by developing and adopting external sensors and learning when and how to use them. Such task-dependent sensing supports efficient action in complex environments.

Vision-Language-Action (VLA) models have emerged as a powerful paradigm for robotic manipulation by leveraging web-scale vision-language pretraining for strong generalization and instruction following (Black et al., 2024; Kim et al., 2024b; Mees et al., 2024; Black et al., 2025; Shukor et al., 2025; Li et al., 2025b; Zhang et al., 2026). Most of these VLA models rely solely on RGB images as visual input, while robots can access heterogeneous non-visual sensory observations such as thermal, acoustic, or radar signals, capturing complementary physical information beyond vision. Recent efforts incorporate additional modalities such as depth (Li et al., 2025a; Bhat et al., 2025; Qu et al., 2025; Liu et al., 2025), tactile (Bi et al., 2025; Huang et al., 2025), and audio (Zhao et al., 2025; Jones et al., 2025). While these approaches show promising gains, they exhibit several key limitations. First, they typically rely on specialized model architectures or encoders for each sensor modality, which is not applicable to diverse and evolving set of sensors. Second, they depend on large-scale sensory datasets for training, which are often expensive and time-consuming to collect. Lastly, existing models typically assume static sensor usage, lacking task-conditioned, adaptive selection of sensory inputs for efficiency and performance.

In this work, we introduce **MuseVLA**, an adaptive multimodal-sensing VLA model that unifies diverse sensing modalities for robotic manipulation (Figure 1). MuseVLA scales VLA capability by treating sensors as on-demand tools, enabling conditional and scalable sensor perception within a single model. MuseVLA dynamically selects a sensor tool based on task context, among thermal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

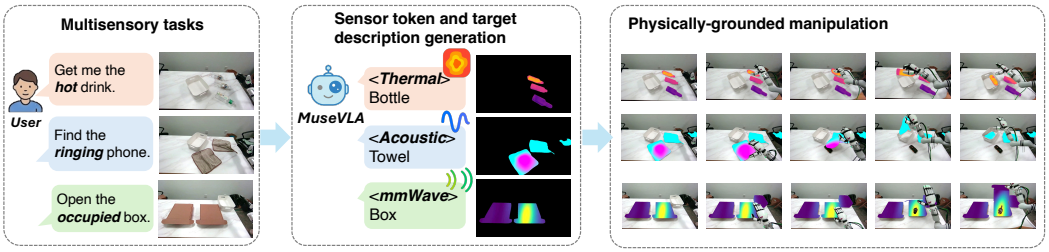


Figure 1: **Adaptive multisensory robotic manipulation.** MuseVLA targets manipulation tasks requiring multimodal sensing beyond RGB. It adaptively selects the suited sensor and generates a target description to construct a grounded sensor image that guides manipulation.

camera, mmWave radar, microphone, etc. This allows the model to acquire task-oriented physical observations, which boosts performance and efficiency. We demonstrate this on dexterous hand, a challenging setting where multisensory manipulation remains largely unexplored.

To achieve this, we introduce learnable *sensor tokens* to enable adaptive sensor selection. Given task instruction and RGB observation, MuseVLA generates the selected sensor token and a target description, analogous to invoking a tool with structured arguments. Sensor observations are then processed and appended as input for subsequent action generation. To support heterogeneous sensors without modality-specific encoders, we use *grounded sensor images* that spatially ground non-visual sensor responses onto the camera image plane over target objects, enabling unified perception across diverse modalities using a pretrained vision encoder. Building on this representation, we develop a data synthesis pipeline that generates multisensory training data from existing RGB datasets, enabling data-efficient training and strong zero-shot generalization.

We evaluate MuseVLA with a real-world multi-sensor robot setup, using challenging robotic manipulation tasks that require sensing modalities, including thermal-guided object pick and place, audio-driven object search, and mmWave radar-assisted hidden object retrieval, and their cascaded multi-stage tasks, each involving manipulating diverse objects with a dexterous hand. Results show 80.6% task success rate on average, significantly outperforming RGB-only baseline and raw sensor heatmaps based VLA baselines by 58% and 47% respectively. MuseVLA also demonstrates effective adaptive sensor selection with high accuracy of generated sensor token and target description. For unseen tasks, MuseVLA trained on synthesized dataset shows strong zero-shot generalization, achieving 66.7% success rate on average.

To summarize, we make the following contributions:

- We present MuseVLA, an adaptive multimodal sensing VLA model that enables scalable and efficient sensor integration as on-demand tools for dexterous-hand manipulation.
- We introduce sensor tokens and target description generation for adaptive sensor selection. We further propose a data synthesis pipeline based on grounded sensor image representation, which synthesizes multisensory datasets from existing RGB robotic datasets for generalization to unseen tasks.
- We evaluate MuseVLA on a suite of challenging robotic manipulation tasks and demonstrate its superior performance in leveraging sensing inputs. We include a video as supplementary material and will open-source our data and code to facilitate further research.

2 Related Work

VLA models for robotic manipulation. Vision-language-action models process language instructions and visual observations to generate robot actions end-to-end (Mees et al., 2024; Kim et al., 2024b; Wen et al., 2025; Ze et al., 2024; Black et al., 2024; Shukor et al., 2025). Unlike conventional manipulation policies trained on specific tasks (Chebotar et al., 2023; Luo et al., 2023), VLA models leverage web-scale pretraining to achieve strong generalization and instruction following. Recent efforts have scaled VLA models across diverse embodiments and large-scale data (Intelligence et al., 2026; Wu et al., 2026; Bjorck et al., 2025; Team, 2026). Complementary to these scaling

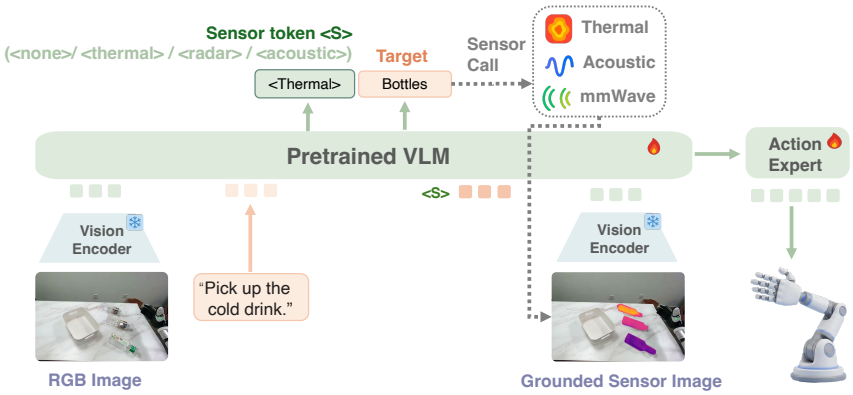


Figure 2: **Overview of MuseVLA model.** Given an RGB image and task instruction, MuseVLA generates a sensor token and target description. The selected sensor is invoked to construct a grounded sensor image, which is appended for manipulation action generation. We co-train VLM backbone and action expert end-to-end on real-world and synthesized multisensory datasets.

efforts, MuseVLA integrates sensing modalities as on-demand tools to empower VLA models for multisensory manipulation. We explore challenging dexterous hand manipulation tasks (Li et al., 2025b; Gao et al., 2026) that require fine-grained spatial understanding and precise control.

Multi-sensor fusion and Multi-modal LLMs. Fusing complementary sensors is critical for domains including autonomous driving (Liu et al., 2022; Wolters et al., 2024; Lin et al., 2024; Zheng et al., 2025; Palladin et al., 2024; Li et al., 2025c; Kim et al., 2024a; Xiong et al., 2025), and 3D scene understanding (Wang et al., 2024; Han et al., 2025). Sensing modalities provide richer physical information and more robust representations than a single camera. Many fusion methods in autonomous driving and scene understanding are designed for perception with sensor-specific architectures, rather than for general action generation and manipulation. MuseVLA presents a scalable and adaptive multi-sensor fusion approach within a unified VLA framework for manipulation. Recent large language models also unify diverse sensor modalities. OneLLM (Han et al., 2024) aligns eight modalities to language, X-Fi (Chen & Yang, 2024) supports variable modality combinations for human sensing, and HoloLLM (Zhou & Yang, 2025) incorporates uncommon sensors into an MLLM. They focus on multimodal understanding or human sensing, whereas MuseVLA learns which physical sensor to invoke adaptively and generates dexterous robot actions.

Multisensory VLA models. Recent works extend VLA models with additional sensing modalities, including depth for spatial understanding (Li et al., 2025a; Bhat et al., 2025; Qu et al., 2025; Patratskiy et al., 2025; Zhen et al., 2024), tactile feedback for force-aware control (Bi et al., 2025; Huang et al., 2025; Yu et al., 2025), speech for human-robot interaction (Zhao et al., 2025), and multisensory fusion (Liu et al., 2025; Guo et al., 2025). OmniVLA (Guo et al., 2025) leverages multiple sensor measurements but ingests all available modalities at all times and does not show zero-shot generalization to unseen tasks. These approaches establish complementary ways to condition robot policies on additional sensors, but generally process a predetermined input modality set rather than learning which physical sensor to invoke from the task context. MuseVLA builds on and goes beyond these works by introducing learnable sensor tokens for task-conditioned adaptive sensor selection that boosts performance and efficiency, data synthesis for data-efficient training and generalization to unseen tasks, and validating these designs on dexterous hand manipulation tasks.

3 Adaptive Multimodal Sensing

Task definition. Given a manipulation task with language instruction $\boldsymbol{l}$, the robot is equipped with sensors $\mathcal{S} = \{S_1, \ldots, S_N\}$, each providing observation $\boldsymbol{s}_i$. The goal is to learn a VLA model $\boldsymbol{\pi}$ that maps RGB observations and sensory inputs to an action chunk $A = (\boldsymbol{a}_t, \ldots, \boldsymbol{a}_{t+H})$:

$$\boldsymbol{\pi} : (\boldsymbol{l},\ \boldsymbol{o}_t,\ \boldsymbol{s}_{1,t}, \ldots, \boldsymbol{s}_{N,t}) \rightarrow A \tag{1}$$

where $\boldsymbol{o}_t$ is the RGB observation and $\boldsymbol{s}_{i,t}$ is the observation from sensor S_i at time t. However, naively incorporating all sensor input is inefficient, as not all are relevant to every task. It also

increases training and inference cost and constrains the model to a fixed sensor set, limiting flexible model deployment across robotic platforms.

Decoupling sensors with VLA model. We introduce an adaptive sensor selection mechanism that decomposes the task into three stages: (1) the VLM selects a task-relevant sensor $\boldsymbol{l_s}$ and a sensing target description $\boldsymbol{l_d}$; (2) the selected sensory observation is processed into an intermediate representation $\boldsymbol{m}$; and (3) this representation is autoregressively appended for action generation.

$$\boldsymbol{\pi} : (\boldsymbol{l}, \boldsymbol{o}_t) \to (\boldsymbol{l_s}, \boldsymbol{l_d}), \quad \mathcal{G} : (\boldsymbol{o}_t, \boldsymbol{s}_{i,t}, \boldsymbol{l_d}) \to \boldsymbol{m}_{i,t}, \quad \boldsymbol{\pi} : (\boldsymbol{l}, \boldsymbol{o}_t, \boldsymbol{l_s}, \boldsymbol{l_d}, \boldsymbol{m}_{i,t}) \to A \tag{2}$$

Here, $\mathcal{G}$ is the sensor grounding function that constructs a grounded sensor image $\boldsymbol{m}_{i,t}$ from the RGB observation, the selected sensor's observation $\boldsymbol{s}_{i,t}$ (with i specified by $\boldsymbol{l_s}$), and the target description. The same model $\boldsymbol{\pi}$ is reused for both sensor selection and action generation, maintaining a unified end-to-end architecture. By decoupling sensors from the VLA backbone via an intermediate representation, sensor and backbone capabilities can be extended independently without repeating RGB–action pretraining from scratch. It also brings efficient inference and training since only the relevant sensor is processed.

Sensor data representation. We use grounded sensor images as the intermediate representation $\boldsymbol{m}$, which encodes diverse sensor data into RGB space by overlaying sensor heatmaps onto task-relevant regions (Figure 3). This unifies sensing modalities without specialized encoders, inspired by how humans ground sensory physical information onto specific objects or regions in the RGB space. This unified format also enables sensory dataset synthesis from existing RGB robotic datasets without dedicated sensor data generators.

Given the RGB observation $\boldsymbol{o}_{\mathrm{RGB}}$ and spatially aligned 2D sensor observation $\boldsymbol{s}$, a segmentation module f_{seg} produces a binary mask $\boldsymbol{M}$ highlighting the target region specified by $\boldsymbol{l_d}$, and the grounded sensor image is constructed as:

$$\boldsymbol{M} = f_{seg}(\boldsymbol{o}_{\mathrm{RGB}}, \boldsymbol{l_d}), \quad \boldsymbol{m} = \boldsymbol{M} \odot \boldsymbol{s} + (1 - \boldsymbol{M}) \odot \boldsymbol{o}_{\mathrm{RGB}} \tag{3}$$

For radar and acoustic arrays, this projection provides directional grounding to a visible region or proxy. Each sensor uses a lightweight, one-time spatial calibration analogous to standard RGB-camera calibration; integrating a new device does not require retraining the VLA from scratch. We leave sensor data processing and alignment details in Section A.

4 Methodology

4.1 MuseVLA Model

Model architecture. As shown in Figure 2, MuseVLA uses pretrained VLM, PaliGemma-2 (Steiner et al., 2024), as backbone for its transferable vision-language understanding and generation capability. The VLM consists of a Gemma-2 (Team et al., 2024) language model and a SigLIP (Zhai et al., 2023) vision encoder with linear projection for alignment. It takes in one or a sequence of images and a language instruction. For the action expert, we use a diffusion transformer (DiT) model (Peebles & Xie, 2023). We use the pretrained VLM weights from VITRA (Li et al., 2025b) to utilize its large-scale human-hand data based pretraining.

Sensor tokens. We introduce learnable *sensor tokens* for adaptive sensor selection. Given the task instruction and RGB image, MuseVLA generates a sensor token as it would for any other discrete language token. The **<None>** token selects RGB-only operation when additional sensing is unnecessary. The **<Thermal>**, **<Acoustic>**, and **<mmWave>** tokens select corresponding sensing modality input, enabling temperature sensing, sound-based localization or monitoring, and radar-based detection of hidden objects, respectively. These serve as representative examples, and additional sensor modalities can be added with corresponding tokens and data.

Target description and segmentation. As shown in Figure 3, we construct grounded sensor images by overlaying sensor heatmaps on masked RGB regions. The VLM backbone generates a target description $\boldsymbol{l_d}$ specifying the objects of interest for robotic manipulation, together with the sensor token $\boldsymbol{l_s}$, treating this as a language grounding (image captioning) task based on the task instruction and RGB observation (Equation (2)). For example, given "Pick up the hot mug", the model generates

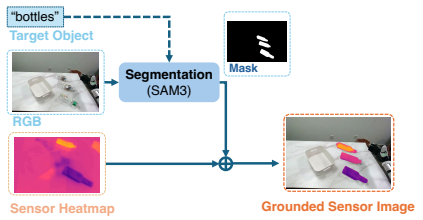


Figure 3: **Grounded sensor image processing.** We perform semantic segmentation with target description and overlay sensor heatmap at masked RGB regions.

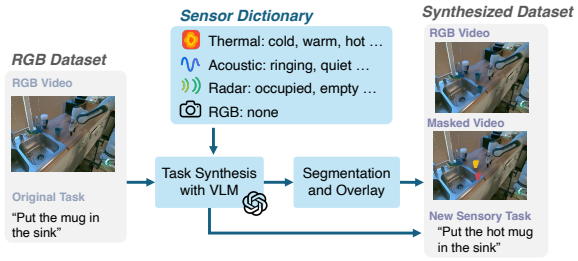


Figure 4: **Data synthesis pipeline.** We inject physical keywords into task instructions of existing RGB datasets and overlay color-coded masks on segmented target objects to synthesize multisensory training data.

"the mugs" as the target description. We then use a pretrained segmentation model, SAM3 (Carion et al., 2025), to obtain a binary mask of the described objects and overlay the sensor heatmap, producing grounded sensor images that encode both visual and physical sensing information in RGB space. The sensor token, modality-specific colormap, and fixed normalization range jointly define the representation's physical semantics, so rendered colors correspond to calibrated temperature or signal-power values rather than arbitrary labels. When the description identifies multiple candidates, SAM3 produces their instance masks and we retain sensor values over the mask union. We choose SAM3 for its strong generalization to unseen objects. Keeping segmentation separate from the VLA backbone also potentially enables asynchronous mask updates during action generation to reduce latency overhead.

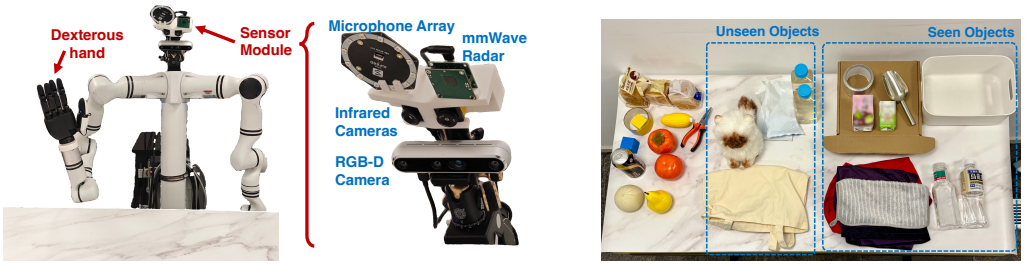
Sensor-guided action generation. After obtaining the grounded sensor image $\boldsymbol{m}$, we append it to the model input after encoding with the VLM vision encoder. With the unified RGB representation, we reuse the same vision encoder for all sensing modalities without requiring specialized encoders. We append a learnable cognition token as extra input as in VITRA. The model then generates a conditioning feature for action expert based on the task instruction, RGB observation, generated sensor token, target description, and grounded sensor image, as in Figure 2. The action expert takes the visual-language feature, robot states, and noisy actions to predict the denoised action sequence for robotic manipulation.

4.2 Data Synthesis Pipeline

Multisensory robot datasets are scarce compared with RGB datasets, and hardware-dependent signal formats and distributions make raw sensor data difficult to synthesize. Our grounded sensor images instead encode physical properties as colors over target regions in RGB images. This enables multisensory data synthesis from existing RGB demonstrations (Figure 4).

Data synthesis. We construct a sensor dictionary mapping each modality to a diverse set of physical property keywords (e.g., thermal $\rightarrow$ "hot", "cold", "warm"). For each RGB video episode, we randomly sample a modality and keyword, inject the keyword into the task instruction (e.g., "Pick up the *hot* mug"), and prompt a VLM (GPT-5.2) to generate a target description. SAM3 then segments the described object, and we overlay a color corresponding to the keyword on the masked region to simulate sensor observations (Figure 4). For example, for the "hot" keyword, we overlay a yellow color mask on the segmented object region. The color mapping is randomized around a base color per keyword for diversity. Lastly, we combine the color-masked video with the original RGB video as synthesized data. We also apply a data augmentation that clones the masked region to a shifted location with a wrong sensor color, encouraging the model to rely on sensor information rather than spatial priors. The synthetic stage supplies diverse sensor-conditioned tasks, objects, and actions rather than reproducing raw sensor physics; the mixed real-world trajectories expose the model to hardware-specific noise and measurement distributions.

Dataset details. We apply the data synthesis pipeline on three existing RGB video robotic manipulation datasets, including MolmoAct (Lee et al., 2025), AgiBotWorld-Alpha (contributors, 2024), and VITRA (Li et al., 2025b) to build a multisensory dataset that covers diverse scenarios, tasks,



(a) Robot hardware setup.

(b) Objects involved in the tasks.

Figure 5: Evaluation setup. We set up a robot arm with a 12DoF dexterous hand and a multi-sensor module and evaluate MuseVLA on a variety of task objects spanning different sensory modalities.

and manipulation actions. In total, we synthesize 9.6K episodes with 1.05 M frames, covering over 1000 objects. We randomly synthesize different sensing modalities on each episode to increase the synthesized dataset diversity.

4.3 Training with Multisensory Data

MuseVLA training includes two parts (Figure 2): VLM training for adaptive sensor selection and target description generation, and VLA training for sensor-guided action generation. We mix real-world multisensory dataset on our downstream robotic tasks with the synthesized dataset, which enables large-scale pretraining for zero-shot generalization on unseen tasks in Section 5.4.

VLM training loss. Using the task instruction and RGB video as input, we train the VLM backbone to generate the appropriate sensor token and target description. We optimize the following VLM loss:

$$\mathcal{L}_{\text{VLM}} = \mathcal{L}_{\text{sensor}} + \mathcal{L}_{\text{target}} \tag{4}$$

where $\mathcal{L}_{\text{sensor}}$ and $\mathcal{L}_{\text{target}}$ are the cross-entropy losses for sensor token and target description generation.

VLM and VLA co-training. For VLA training, we incorporate the grounded sensor images as additional input and jointly train the VLM backbone with the diffusion action expert end-to-end. The overall training loss combines the VLM loss with the diffusion MSE loss (Peebles & Xie, 2023):

$$\mathcal{L} = \mathcal{L}_{\text{VLM}} + \lambda\, \mathbb{E}_{\tau,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{a}^\tau, \tau, \mathbf{c})\|_2^2\right], \tag{5}$$

where $\mathbf{a}^\tau$ is the noisy action at diffusion step τ, $\mathbf{c}$ denotes the conditioning inputs (visual-language feature from VLM and robot states), ϵ and ϵ_θ are the ground-truth and predicted noise. λ is a hyperparameter balancing the two objectives and we set $\lambda = 1\text{e}{-2}$ in our experiments. Joint optimization preserves both accurate VLM outputs and effective action generation. Our experiments show that omitting the VLM loss degrades VLM outputs.

5 Experiments

Our evaluation covers the following: (1) task success rates on challenging manipulation tasks that require multimodal sensing inputs. (2) adaptive sensor selection performance and improved efficiency. (3) benefits of synthesized dataset pretraining, showing generalization to unseen tasks.

5.1 Experiment Setup

Robot setup. We set up a table-top experiment environment using a robot equipped with 12-DoF Robotera XHand dexterous hands and a sensor suite, as in Figure 5. The sensor suite includes an RGB-D camera (Intel RealSense), two thermal cameras (infiRay T2S), an mmWave radar (Calterah 4T4R 60GHz radar), and a microphone array (Sipeed 6+1Mic Array). Thermal, radar, and acoustic measurements yield 2D heatmaps of temperature, reflection intensity, and sound intensity, respectively, aligned with the RGB camera's field of view (Section A). To collect robotic manipulation

Table 1: Task success rates comparison across three types of sensing modality tasks. The p column reports the two-sided Fisher's exact test on pooled end-to-end success against MuseVLA (Ours).

	Per-modality Success Rate				Per-stage Success Rate			
	Thermal	Acoustic	mmWave	*Average*	Sensing	Manipulation	*Score*	p
π_0-RGB	33.3%	25.0%	4.17%	20.8%	48.6%	43.1%	0.458	<0.001
$\pi_{0.5}$-RGB	16.7%	33.3%	8.33%	19.4%	41.7%	33.3%	0.375	<0.001
π_0-Raw	16.7%	41.7%	25.0%	27.8%	86.1%	27.8%	0.569	<0.001
$\pi_{0.5}$-Raw	16.7%	33.3%	20.8%	23.6%	83.3%	29.2%	0.563	<0.001
MuseVLA-RGB	12.5%	33.3%	20.8%	22.2%	41.7%	43.1%	0.424	<0.001
MuseVLA-Raw	41.7%	25.0%	33.3%	33.3%	91.7%	33.3%	0.625	<0.001
MuseVLA-RawAdapt	70.8%	41.7%	66.7%	59.7%	93.1%	59.7%	0.764	<0.05
MuseVLA-AllInput	50.0%	54.2%	66.7%	56.9%	87.5%	63.9%	0.757	0.021
MuseVLA (Ours)	**83.3%**	**58.3%**	**87.5%**	**76.4%**	**95.8%**	**77.8%**	**0.868**	–

Table 2: Adaptive sensor selection accuracy.

	Training Tasks		Unseen Tasks	
	Sensor	Target	Sensor	Target
PaliGemma-2	0%	13.0%	0%	9.5%
MuseVLA w/o pretrain	100%	**100%**	85%	40.5%
MuseVLA (pretrained)	**100%**	93.5%	**100%**	**82.0%**

demonstration data, we use a teleoperation system with a pair of MANUS5 gloves to control the robot hands and record action sequences and all sensors' measurements.

Task setting. We design three categories of manipulation tasks that demand beyond-RGB sensing modalities, each with multiple sub-task instruction variants and target objects. (1) *Thermal-guided pick-and-place*: The robot picks up a drink with a specified temperature and places it into a basket, e.g., *"Pick up the hot drink and place it into the basket."* Instructions cover three thermal keywords (*hot*, *cold*, *room-temperature*) over two types of drinks with varying positions and orientations. (2) *Acoustic-grounded object search*: The robot localizes a hidden sound source using spatial audio cues and removes the covering, e.g., *"Pick up the clothes/towels covering the ringing phone and place it into the basket"*. Coverings include clothes and towels of varying shapes and colors. (3) *mmWave radar-guided object search*: The robot uses the mmWave radar to see through closed boxes, opens the box that contains an item, e.g., *"Open the occupied box."* Hidden items include drinks, shovels, and tape rolls, with box positions varied across trials. In total, we collect 720 teleoperated demonstration episodes, covering 10 different sub-task instructions, 7 objects with varying placements across 3 sensing modalities, as our real-world training dataset. Figure 5b shows seen objects in our datasets. Figure 1 shows task examples and Figure 7 shows successful action trajectories. We evaluate tasks with distracting objects to test the model's ability to ground on the correct target object.

Training setup. We initialize from VITRA (Li et al., 2025b) weights for both VLM backbone and action expert, and perform VLM and VLA co-training on 64 A100 (40GB) GPUs with a batch size of 512 for 20K steps ($\sim$20 hours) using a learning rate of $1e{-}5$ and AdamW optimizer. We train and evaluate a single MuseVLA model across all tasks, which is more challenging than per-task models.

Evaluation metrics. We report two types of metrics. (i) *Task success rate* is the fraction of trials in which the robot completes the full task end-to-end. (ii) *Task score* decomposes a trial into a *sensing* sub-task and a *manipulation* sub-task, each contributing 0.5: 0.5 for selecting and grounding on the correct target object (e.g., the hot drink, the occupied box), and 0.5 for performing the correct manipulation (picking and placing in the container, opening the box, removing the covering). The per-stage *sensing* and *manipulation* success rates in Table 1 are averaged across trials. Manipulation success measures whether the requested physical operation is completed, even if applied to the wrong target, whereas end-to-end success requires both correct target selection and manipulation. We report two-sided Fisher's exact p-values for pooled success against MuseVLA.

Table 3: Impact of pretraining on synthesized multisensory data. We report success rates on training tasks and zero-shot generalization to unseen tasks.

	Seen tasks				Unseen tasks				
	Thermal	Acoustic	mmWave	*Average*	Thermal	Acoustic	mmWave	*Average*	p
MuseVLA-Raw	41.7%	25.0%	33.3%	33.3%	31.3%	25.0%	18.8%	25.0%	<0.001
MuseVLA w/o pretrain	83.3%	58.3%	**87.5%**	76.4%	25.0%	31.3%	25.0%	27.1%	<0.001
MuseVLA w/o synth. overlay	79.2%	58.3%	83.3%	73.6%	31.3%	43.8%	43.8%	39.6%	0.014
MuseVLA (pretrained)	**87.5%**	**70.8%**	83.3%	**80.6%**	**75.0%**	**56.3%**	**68.8%**	**66.7%**	–

Baselines. We compare against two baseline types by finetuning frontier VLA models, π_0 and $\pi_{0.5}$, on our dataset: (i) *VLA-RGB*, standard RGB-only VLA models that receive no sensor input; (ii) *VLA-Raw*, VLA models that receive spatially aligned sensor heatmaps from all available sensors concatenated as additional image inputs, but without the grounded sensor image processing or adaptive sensor selection.

5.2 Multisensory Manipulation Task Performance

Task success rates. Table 1 reports success rates and scores over 24 independent trials per modality. Here, MuseVLA (Ours) uses only real-world training data; Table 3 evaluates synthesized-data pretraining. MuseVLA achieves 76.4% average success, exceeding RGB-only baselines (22.2% for MuseVLA-RGB and 20.8% for π_0-RGB) and raw-heatmap baselines (33.3% for MuseVLA-Raw and 27.8% for π_0-Raw). Acoustic tasks are harder because they require grasping soft fabric. Raw-heatmap baselines often select the correct target but fail at manipulation, indicating that unprocessed sensor inputs limit the benefit of multisensory perception. Compared with the matched MuseVLA-AllInput variant, adaptive selection improves pooled end-to-end success from 56.9% to 76.4% (two-sided Fisher's exact $p = 0.021$) while processing only the selected sensor image.

Ablation study. We evaluate three ablations: (1) removing grounded sensor images (MuseVLA-RawAdapt), (2) removing adaptive sensor selection (MuseVLA-AllInput), and (3) removing both (MuseVLA-Raw). All three reduce task success (Table 1). MuseVLA-RawAdapt outperforms MuseVLA-Raw, indicating that selecting the relevant heatmap reduces interference from irrelevant sensors even without grounding. MuseVLA-AllInput retains the same backbone and grounded sensor representation as MuseVLA, but jointly processes all three sensing modalities instead of only the selected one. This joint use of sensor inputs is similar to OmniVLA (Guo et al., 2025). The comparison with MuseVLA therefore isolates the benefit of adaptive sensor selection within a common architecture.

Multi-sensor tasks. We test two multi-stage tasks (Figure 8): (i) *Radar → RGB*, opening an occupied box using radar, then RGB-only pick-and-place into a basket; and (ii) *Radar → Thermal*, radar-guided box opening followed by selecting a hot, cold, or room-temperature drink among multiple drinks. Over 12 trials each, Radar → RGB achieves 66.7% end-to-end success (100% localization, 83.3% opening, 66.7% pick-and-place), and Radar → Thermal achieves 75.0% (100% localization, thermal selection, and opening; 75.0% pick-and-place), demonstrating sensor switching within one task.

5.3 Adaptive Sensor Selection

VLM generation accuracy. We evaluate MuseVLA's adaptive sensor selection in Table 2 across 200 samples each from training and unseen tasks. We measure two accuracy metrics: (1) *sensor selection accuracy*, the fraction of correct sensor token predictions, and (2) *target description accuracy*, the fraction of semantically correct target descriptions (evaluated via GPT-5.2). A human verifier independently audits 50 randomly sampled target descriptions and finds no mismatch with the GPT-based judgments. MuseVLA substantially outperforms the PaliGemma-2 base VLM on both metrics, demonstrating the effectiveness of our VLM training. Pretraining on synthesized data further boosts target description accuracy from 40.5% to 82.0% on unseen tasks, which is critical for zero-shot generalization.

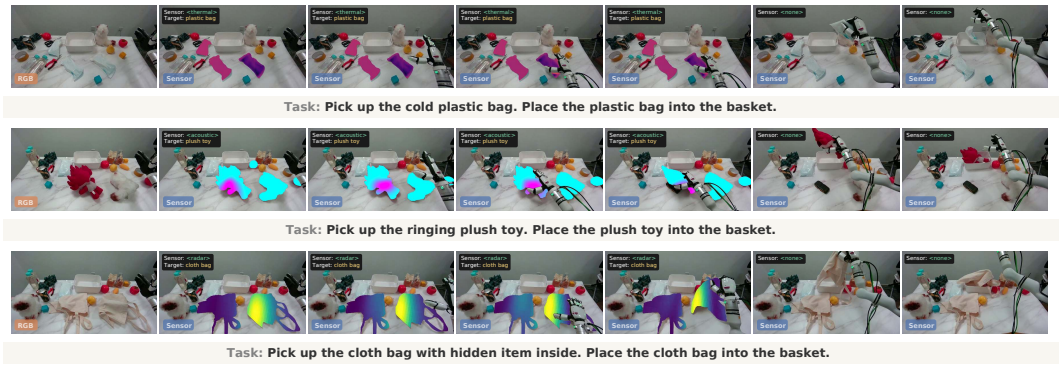


Figure 6: Execution trajectories of unseen task evaluation.

Inference efficiency. We measure GPU memory on an A100 80 GB GPU over 100 samples. With three modalities, adaptive selection reduces inference memory from 13.15 ± 0.01 GB for AllInput to 6.43 ± 0.01 GB, excluding the 11.9 GB shared model weights, and reduces training peak memory from 42.50 ± 0.10 GB to 33.68 ± 0.07 GB at batch size 8. Because MuseVLA processes RGB and only one selected sensor view, its memory remains constant as sensor count grows, whereas AllInput memory increases with additional views (Table 4).

5.4 Effectiveness of Pretraining

Unseen task setting. We evaluate *unseen tasks* over 16 independent trials per modality, recombining sensors with new targets. These tasks are unseen in the training datasets and are evaluated without additional task-specific training. For example, *"Pick up the cloth bag with the ringing item inside and place it into the basket"* requires acoustic localization and grasping an unseen object. Figures 6 and 9 show execution trajectories.

Pretraining effectiveness. We compare MuseVLA with and without pretraining on synthesized data in Table 3, using 24 trials per modality on seen tasks and 16 trials per modality on unseen tasks. Both variants use VITRA initialization; "with/without pretraining" refers to including synthesized multisensory data in MuseVLA training. MuseVLA with pretraining shows comparable performance on seen tasks, while significantly improving average unseen task success by 39.6% (from 27.1% to 66.7%), with gains across three modalities. We observe a 12.5% gain for seen acoustic tasks due to improved dexterous hand manipulation when grabbing soft clothes. The gain for unseen zero-shot performance stems from both more accurate sensor selection and target descriptions (Table 2) and better manipulation actions learned from more diverse manipulation tasks in the synthesized dataset. To separate sensor-grounding supervision from additional RGB manipulation data, MuseVLA w/o synth. overlay retains the same 9.6K additional trajectories and 20K-step budget but removes their sensor overlays. Its average success is 73.6% on seen tasks and 39.6% on unseen tasks, compared with 80.6% and 66.7% for MuseVLA. On unseen tasks, removing the overlays reduces pooled sensing success ($p = 1.29 \times 10^{-4}$) and end-to-end success significantly ($p = 0.014$), showing that the generalization gain is not explained by additional RGB trajectories alone.

6 Conclusion

MuseVLA treats sensors as on-demand tools for dexterous manipulation, predicting a sensor token and target description before constructing a *grounded sensor image* for action generation. Its data synthesis pipeline augments RGB videos to reduce multisensory data requirements. On a real-world robot, it achieves 80.6% average success across thermal-, audio-, and radar-guided tasks, outperforms RGB-only and multisensory baselines, and generalizes zero-shot to unseen tasks. Limitations include the modest real-world dataset and dependence on accurate segmentation; larger datasets and stronger segmentation models may relax these constraints.

AI USE STATEMENT

We used LLMs to assist with language and LaTeX polishing, synthetic dataset construction (Section 4.2), and target-description evaluation (Section 5.3). The authors take full responsibility for the final manuscript, methods, results, and claims.

REPRODUCIBILITY STATEMENT

Section 4 describes the model, data synthesis pipeline, and training objectives. Section 5.1 specifies the robot and sensor hardware, task definitions, training configuration, evaluation metrics, and baselines. Sections 5 and 5.4 reports evaluation trial counts. Section A provides sensor processing and calibration details; Section B reports memory scaling, ablations, and robustness tests; and Section C shows additional execution trajectories.

REFERENCES

Vineet Bhat, Yu-Hsiang Lan, Prashanth Krishnamurthy, Ramesh Karri, and Farshad Khorrami. 3D CAVLA: Leveraging depth and 3D context to generalize vision language action models for unseen tasks. *arXiv preprint arXiv:2505.05800*, 2025.

Jianxin Bi, Kevin Yuchen Ma, Ce Hao, Mike Zheng Shou, and Harold Soh. VLA-Touch: Enhancing vision-language-action models with dual-level tactile feedback. *arXiv preprint arXiv:2507.17294*, 2025.

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.

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.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert 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. In Joseph Lim, Shuran Song, and Hae-Won Park (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 27–30 Sep 2025. URL https://proceedings.mlr.press/v305/black25a.html.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Yevgen Chebotar, Quan Vuong, Karol Hausman, Fei Xia, Yao Lu, Alex Irpan, Aviral Kumar, Tianhe Yu, Alexander Herzog, Karl Pertsch, et al. Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions. In *Conference on Robot Learning*, pp. 3909–3928. PMLR, 2023.

Xinyan Chen and Jianfei Yang. X-Fi: A modality-invariant foundation model for multimodal human sensing, 2024. URL https://arxiv.org/abs/2410.10167.

AgiBot World Colosseum contributors. Agibot world colosseum. https://github.com/OpenDriveLab/AgiBot-World, 2024.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

Heyu Guo, Shanmu Wang, Ruichun Ma, Shiqi Jiang, Yasaman Ghasempour, Omid Abari, Baining Guo, and Lili Qiu. Omnivla: Physically-grounded multimodal vla with unified multi-sensor perception for robotic manipulation. *arXiv preprint arXiv:2511.01210*, 2025.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. OneLLM: One framework to align all modalities with language. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26574–26585. IEEE, 2024.

Xiaofeng Han, Shunpeng Chen, Zenghuang Fu, Zhe Feng, Lue Fan, Dong An, Changwei Wang, Li Guo, Weiliang Meng, Xiaopeng Zhang, et al. Multimodal fusion and vision-language models: A survey for robot vision. *arXiv preprint arXiv:2504.02477*, 2025.

Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-VLA: Unlocking vision-language-action model's physical knowledge for tactile generalization. *arXiv preprint arXiv:2507.09160*, 2025.

Physical Intelligence, Bo Ai, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Greg Balke, Kevin Black, George Bokinsky, Shihao Cao, Thomas Charbonnier, et al. $\pi_{0.7}$: A steerable generalist robotic foundation model with emergent capabilities. *arXiv preprint arXiv:2604.15483*, 2026.

Joshua Jones, Oier Mees, Carmelo Sferrazza, Kyle Stachowicz, Pieter Abbeel, and Sergey Levine. Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5961–5968. IEEE, 2025.

Jisong Kim, Minjae Seong, and Jun Won Choi. Crt-fusion: Camera, radar, temporal fusion using motion information for 3d object detection. *Advances in Neural Information Processing Systems*, 37:108625–108648, 2024a.

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*, 2024b.

Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, Winson Han, Wilbert Pumacay, Angelica Wu, Rose Hendrix, Karen Farley, Eli VanderBilt, Ali Farhadi, Dieter Fox, and Ranjay Krishna. MolmoAct: Action reasoning models that can reason in space, 2025. URL `https://arxiv.org/abs/2508.07917`.

Chengmeng Li, Junjie Wen, Yan Peng, Yaxin Peng, Feifei Feng, and Yichen Zhu. PointVLA: Injecting the 3D world into vision-language-action models. *arXiv preprint arXiv:2503.07511*, 2025a.

Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025b.

Yiheng Li, Yang Yang, and Zhen Lei. Rctrans: Radar-camera transformer via radar densifier and sequential decoder for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5048–5056, 2025c.

Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu. Rcbevdet: Radar-camera fusion in bird's eye view for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14928–14937, June 2024.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022.

Zhuoyang Liu, Jiaming Liu, Jiadong Xu, Nuowei Han, Chenyang Gu, Hao Chen, Kaichen Zhou, Renrui Zhang, Kai Chin Hsieh, Kun Wu, et al. MLA: A multisensory language-action model for multimodal understanding and forecasting in robotic manipulation. *arXiv preprint arXiv:2509.26642*, 2025.

Jianlan Luo, Perry Dong, Jeffrey Wu, Aviral Kumar, Xinyang Geng, and Sergey Levine. Action-quantized offline reinforcement learning for robotic skill learning. In *Conference on Robot Learning*, pp. 1348–1361. PMLR, 2023.

Oier Mees, Dibya Ghosh, Karl Pertsch, Kevin Black, Homer Rich Walke, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

Edoardo Palladin, Roland Dietze, Praveen Narayanan, Mario Bijelic, and Felix Heide. Samfusion: Sensor-adaptive multimodal fusion for 3d object detection in adverse weather. In *European Conference on Computer Vision*, pp. 484–503. Springer, 2024.

Maxim A Patratskiy, Alexey K Kovalev, and Aleksandr I Panov. Spatial traces: Enhancing vla models with spatial-temporal understanding. *arXiv preprint arXiv:2508.09032*, 2025.

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

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. SpatialVLA: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

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.

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.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Generalist AI Team. Gen-1: Scaling embodied foundation models to mastery. *Generalist AI Blog*, 2026. https://generalistai.com/blog/apr-02-2026-GEN-1.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19757–19767, June 2024.

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.

Philipp Wolters, Johannes Gilg, Torben Teepe, Fabian Herzog, Anouar Laouichi, Martin Hofmann, and Gerhard Rigoll. Unleashing hydra: Hybrid fusion, depth consistency and radar for unified 3d perception. *arXiv preprint arXiv:2403.07746*, 2024.

Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Shuailei Ma, He Sun, Yong Wang, Zhenqi Qiu, Houlong Xiong, Ziyu Wang, Shuai Zhou, Yiyu Ren, Kejia Zhang, Hui Yu, Jingmei Zhao, Qian Zhu, Ran Cheng, Yong-Lu Li, Yongtao Huang, Xing Zhu, Yujun Shen, and Kecheng Zheng. A pragmatic vla foundation model. *arXiv preprint arXiv:2601.18692v1*, 2026.

Weiyi Xiong, Zean Zou, Qiuchi Zhao, Fengchun He, and Bing Zhu. Lxlv2: Enhanced lidar excluded lean 3d object detection with fusion of 4d radar and camera. *IEEE Robotics and Automation Letters*, 2025.

Jiawen Yu, Hairuo Liu, Qiaojun Yu, Jieji Ren, Ce Hao, Haitong Ding, Guangyu Huang, Guofan Huang, Yan Song, Panpan Cai, et al. Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation. *arXiv preprint arXiv:2505.22159*, 2025.

Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, Xue Bin Peng, and Jiajun Wu. Generalizable humanoid manipulation with 3d diffusion policies. *arXiv preprint arXiv:2410.10803*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Chubin Zhang, Jianan Wang, Zifeng Gao, Yue Su, Tianru Dai, Cai Zhou, Jiwen Lu, and Yansong Tang. Clap: Contrastive latent action pretraining for learning vision-language-action models from human videos. *arXiv preprint arXiv:2601.04061*, 2026.

Wei Zhao, Pengxiang Ding, Min Zhang, Zhefei Gong, Shuanghao Bai, Han Zhao, and Donglin Wang. VLAS: Vision-language-action model with speech instructions for customized robot manipulation. *arXiv preprint arXiv:2502.13508*, 2025.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3D-VLA: A 3D vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.

Lianqing Zheng, Jianan Liu, Runwei Guan, Long Yang, Shouyi Lu, Yuanzhe Li, Xiaokai Bai, Jie Bai, Zhixiong Ma, Hui-Liang Shen, et al. Doracamom: Joint 3d detection and occupancy prediction with multi-view 4d radars and cameras for omnidirectional perception. *arXiv preprint arXiv:2501.15394*, 2025.

Chuhao Zhou and Jianfei Yang. HoloLLM: Multisensory foundation model for language-grounded human sensing and reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 81772–81797, 2025. doi: 10.52202/085713-2464.

A Implementation Details

A.1 Sensor Heatmap Construction

The thermal camera directly outputs a 2D temperature map over pixel coordinates, requiring no additional spatial processing. For the mmWave radar and microphone array, each array element k records a complex-valued measurement $z_k = a_k e^{j\phi_k}$. We convert these per-element signals into spatially resolved 2D heatmaps using standard digital beamforming, a well-established signal processing technique. We consider the received power from a direction (θ, φ) (azimuth, elevation):

$$P(\theta,\varphi) = 20\log_{10}\left\|\sum_{k=1}^{K} a_k\, e^{j\phi_k}\, e^{-j\Delta_k(\theta,\varphi)}\right\|, \quad \Delta_k(\theta,\varphi) = \frac{2\pi}{\lambda}(d_k^x \cos\varphi\, \sin\theta + d_k^y \sin\varphi)\,, \quad (6)$$

where λ is the carrier wavelength and (d_k^x, d_k^y) is the position of element k in the array plane. The resulting azimuth-elevation heatmap captures spatial distribution of the received power, such as mmWave reflection intensity or received sound intensity. We map each heatmap into RGB space using sensor-specific colormaps (inferno for thermal, viridis for mmWave, plasma for acoustic), with fixed normalization ranges determined by the typical operating range of each sensor. Together with the sensor token, the fixed colormap and normalization assign colors to physical semantics. Thermal colors map to temperatures over 0–60°C, while acoustic and mmWave colors encode received power in dB. These mappings remain fixed between training and inference; hardware with a different palette or scale is converted to the same canonical units and range.

A.2 Sensor Heatmap Spatial Alignment

Because the sensors are rigidly mounted and our tabletop workspace has limited depth variation, we perform a lightweight, one-time registration between each sensor heatmap and the RGB frame. The calibrated mapping is fixed during inference and requires no model retraining or online optimization. We use it to retain sensor measurements within the RGB-space target masks when constructing grounded sensor images. Thermal measurements have direct image-plane correspondence, while radar and acoustic maps provide coarser directional grounding to visible regions or proxies, such as containers holding hidden sources.

B Additional Experimental Results

B.1 Memory Scaling with Sensor Count

Table 4: GPU memory as the number of installed sensing modalities increases. Measurements are averaged over 100 samples on one A100 80 GB GPU; training uses batch size 8.

Method	Modalities	Input views	Inference (GB)	Training peak (GB)
Adaptive	Any	2	6.43 ± 0.01	33.68 ± 0.07
AllInput	1	2	6.43 ± 0.01	33.68 ± 0.06
AllInput	2	3	9.64 ± 0.03	37.78 ± 0.10
AllInput	3	4	13.15 ± 0.01	42.50 ± 0.10
AllInput	4	5	16.87 ± 0.01	47.23 ± 0.10
AllInput	8	9	33.50 ± 0.02	66.19 ± 0.10

Adaptive selection keeps two image views—RGB and the selected grounded sensor image—independent of the number of installed modalities. In contrast, AllInput adds one view per modality, causing inference and training memory to grow with the sensor suite.

B.2 Ablation of Physical Sensor Values

We first test whether MuseVLA actually uses the physical values inside the grounded sensor masks, rather than relying on the target-mask shape or language alone. At test time, the flat-color control

replaces all in-mask measurements with a constant white value, while the all-black control removes the sensor value (no color mask). Both controls use the same pretrained checkpoint without further training.

Table 5: Ablation of physical sensor values. We report pooled sensing, manipulation, and end-to-end success across the three modalities.

Condition	Sensing	Manipulation	End-to-end
MuseVLA (no perturbation)	95.8%	77.8%	76.4%
Flat color	47.2%	77.8%	36.1%
All black	47.2%	86.1%	41.7%

Destroying sensor values reduces pooled sensing from 95.8% to 47.2% for both controls (two-sided Fisher's exact $p = 1.09\times10^{-8}$). End-to-end success also drops from 76.4% to 36.1% under flat-color inputs ($p = 1.02\times10^{-4}$) and to 41.7% under all-black inputs ($p = 5.72\times10^{-4}$). This isolates the contribution of the physical sensor measurements: target masks without informative values are insufficient for sensor-guided target selection.

B.3 Robustness to Alignment and Segmentation Errors

Our main evaluation already uses measurements from physical sensors and includes their natural noise and residual sensor-to-RGB alignment errors. It also retains natural SAM3 failures (around or less than 10% of all input frames across tasks) as model inputs. We test more extreme conditions by injecting additional overlay shifts of 30, 50, and 80 pixels for thermal, acoustic, and mmWave, respectively, on top of the existing alignment errors, and by hard-coding a 30% segmentation-failure rate on the thermal task, much higher than the natural failure rate of evaluated object category.

Table 6: Robustness to additional sensor-mask alignment shifts. We report pooled sensing, manipulation, and end-to-end success across the three modalities.

Condition	Sensing	Manipulation	End-to-end
MuseVLA (no perturbation)	95.8%	77.8%	76.4%
Additional shifts	91.7%	75.0%	72.2%

Table 7: Robustness to injected segmentation failures on the thermal task.

Condition	Sensing	Manipulation	End-to-end
MuseVLA (no perturbation)	100%	83.3%	83.3%
30% segmentation failures	83.3%	91.7%	83.3%

Despite the additional errors, MuseVLA retains 91.7% sensing, 75.0% manipulation, and 72.2% end-to-end success under injected shifts. Under the injected 30% segmentation-failure rate, it retains 83.3% end-to-end success. These stress tests show robustness of MuseVLA, demonstrating reliable performance under alignment errors or segmentation failures beyond those naturally present in the main evaluation.

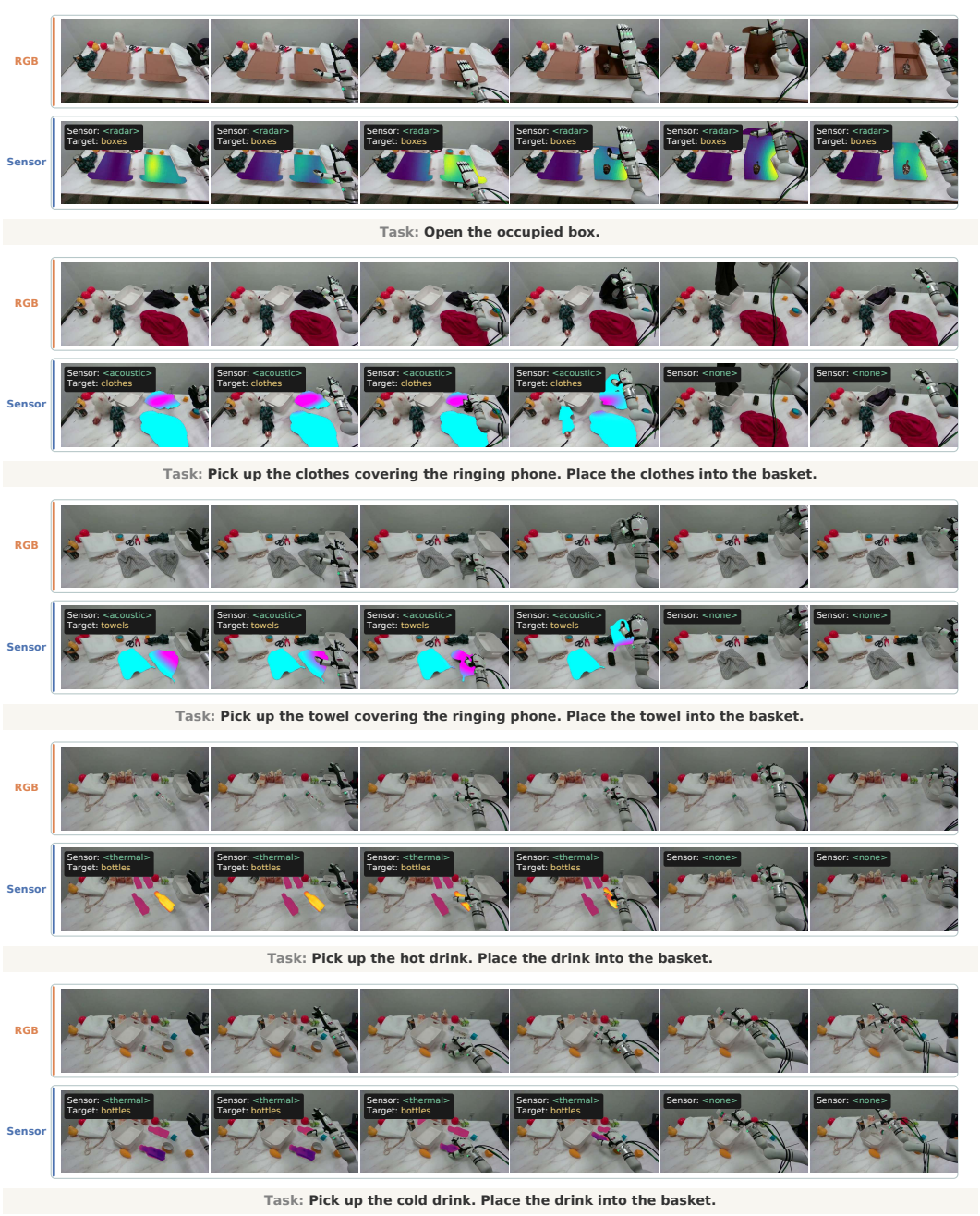


Figure 7: MuseVLA execution trajectories of training task evaluation.

C Real-Robot Execution Trajectories



Figure 8: MuseVLA execution trajectories of multi-stage task evaluation.



Figure 9: MuseVLA execution trajectories of zero-shot unseen task evaluation.