RynnBrain 1.1: Towards More Capable and Generalizable Embodied Foundation Model

Kehan Li[*,†], **Bohan Hou**[*], **Minghao Zhu**[*], **Tianyi Zhang**[*], **Zesen Cheng**[*], **Zhikai Wang**[*], **Sicong Leng**[*], **Xin Li**[*,†], **Xiao Lin**[*], **Biying Yao**[*], **Minghua Zeng**[*], **Jiangpin Liu**[*], **Ronghao Dang**[*], **Jiayan Guo**[*], **Siteng Huang**, **Haoyu Zhao**, **Heng Ping**, **Yaxi Zhao**, **Kexiang Wang**, **Tong Lu**, **Shengke Xue**, **Jiahao Tang**, **Yulei Wang**, **Zejing Wang**, **Jianwei Gao**, **Shijian Lu**, **Chengju Liu**, **Jianfei Yang**, **Mingxiu Chen**, **Deli Zhao**

[1]DAMO Academy, Alibaba Group [2]Hupan Lab
[*]Core contributors, [†]Project lead

We present RynnBrain 1.1, a family of embodied foundation models spanning 2B, 9B, and 122B-A10B scales. Trained with a unified spatio-temporal and physically grounded framework, RynnBrain 1.1 supports embodied perception, spatial reasoning, localization, and planning. Compared with RynnBrain 1.0, it further introduces contact-point prediction across the model family and native 3D grounding for the 2B and 9B models, yielding representations and outputs that are more directly aligned with robot manipulation. We also develop RynnBrain-VLA with a unified cross-embodiment action space and embodiment-specific masking, and deploy it on Unitree G1, Astribot-S1, and Tianji-Wuji. RynnBrain 1.1 achieves strong results on embodied cognition, localization, and 3D grounding, with the 122B-A10B model outperforming all evaluated proprietary and open-source models on VSI-Bench, MMSI, and RefSpatial-Bench. Real-robot experiments show that RynnBrain-initialized policies outperform Qwen-based and representative generalist VLAs, while joint multi-task and multi-embodiment training improves process scores and success rates over per-task training.

https://alibaba-damo-academy.github.io/RynnBrain
https://github.com/alibaba-damo-academy/RynnBrain
https://huggingface.co/collections/Alibaba-DAMO-Academy/rynnbrain-11
https://modelscope.cn/collections/DAMO_Academy/RynnBrain-11

Date: July 20, 2026

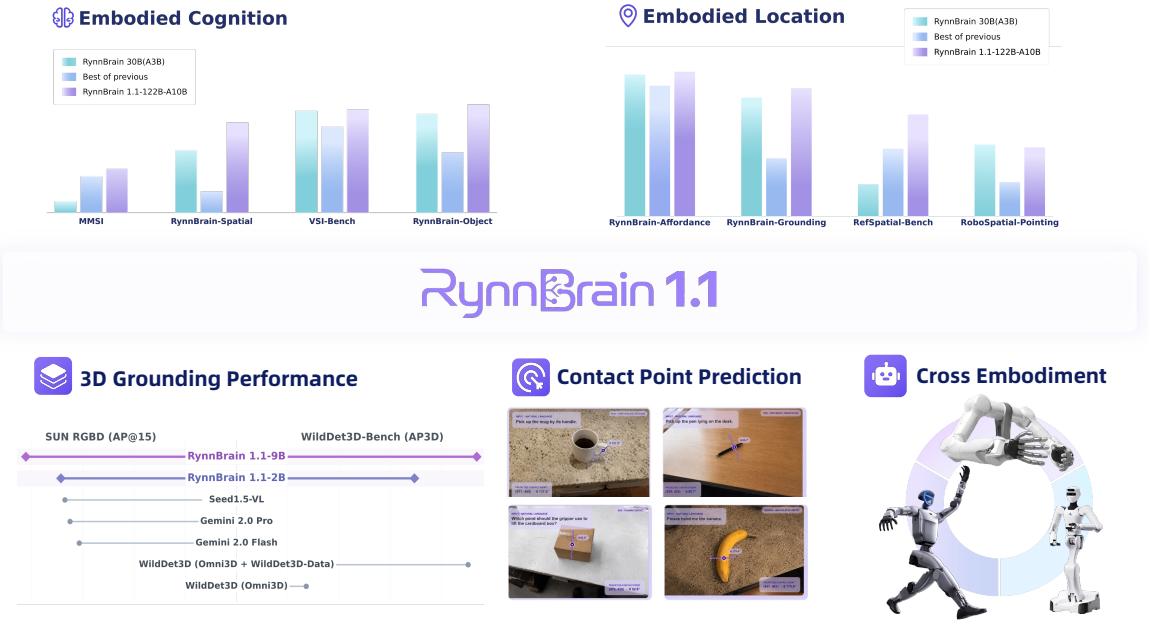

Figure 1 We introduce RynnBrain 1.1, a model family comprising three variants—2B, 9B, and 122B-A10B—together with the post-trained RynnBrain-VLA. RynnBrain 1.1 achieves leading performance across a broad suite of benchmarks, particularly in cognition, localization, 3D grounding, and contact-point prediction. RynnBrain-VLA demonstrates significant promise in enabling action prediction across diverse embodiments within a unified action space.

1 Introduction

Embodied intelligence requires models to go beyond visual recognition and language understanding. A robot operating in the physical world must perceive objects [57, 55], reason about spatial relations [60, 80], identify actionable regions [27], and translate such understanding into executable behaviors [6, 5]. Although recent multimodal large language models have achieved strong performance in image and video question answering tasks [4, 76, 29, 17], real-world robotic systems impose additional requirements: they must understand 3D structure, reason across viewpoints, ground language instructions into physical locations, and support downstream manipulation policies. Therefore, the central challenge for embodied foundation models is not only whether they can understand visual content, but whether visual understanding is grounded in the physical world and, crucially, whether it can be transferred to real-robot action [64].

Recent embodied foundation models [28, 62, 20, 49, 65] have advanced rapidly along different axes, including object and region localization, spatio-temporal understanding, and robot-oriented planning. Specifically, RynnBrain 1.0 [20] explored this direction by introducing a unified embodied multimodal model for egocentric understanding, spatial grounding, physically grounded reasoning, and planning [33, 19, 52], demonstrating the feasibility of combining general and robot-oriented multimodal understanding. Despite its strong performance, several key aspects were not fully explored in RynnBrain 1.0. First, how can the representations and outputs from RynnBrain be brought closer to robot manipulation? Second, how effectively can RynnBrain serve as an initialization for downstream VLA post-training?

Motivated by these two questions, we develop RynnBrain 1.1, a new family of embodied foundation models upgraded from RynnBrain 1.0 [20]. Concretely, to learn representations that are more meaningful for robot manipulation, we add two new training tasks to the embodied pretraining of RynnBrain 1.1: (1) **contact point prediction** and (2) **3D grounding**. The contact point prediction task requires the model to locate the center point and the rotation angle of the gripper for grasping. Compared to the grasp rectangle detection task in [66, 20], contact point prediction can guide the model to provide more precise (i.e., the contact point) and meaningful (i.e., the planar rotation angle of the gripper) features for the downstream robot policy. As for 3D grounding, it introduces explicit and massive 3D modeling during the training of RynnBrain 1.1. Since robots operate in three-dimensional environments [67, 31], equipping embodied foundation models with explicit 3D modeling benefits manipulation through both richer representations and directly usable 3D outputs. In addition, to support the fine-tuning of RynnBrain 1.1 across heterogeneous robot embodiments and control interfaces, we introduce a unified action space that organizes embodiment-specific actions into semantically aligned body-part groups. Embodiment-specific masks are then applied to activate only the dimensions available on each robot. This design allows data from different embodiments to be trained jointly within a single policy without forcing explicit alignment between incompatible action spaces.

Built upon Qwen3.5 [53], RynnBrain 1.1 is available in three model scales—2B, 9B, and 122B-A10B—all trained under a unified capability definition and training framework. This consistent design enables us to systematically study how model capabilities evolve with scale. We evaluate RynnBrain 1.1 from two complementary perspectives. First, we benchmark the base models on a broad suite of embodied multimodal tasks covering spatio-temporal understanding, spatial reasoning, localization, and 3D grounding. The results show an upward trend from 2B to 9B and 122B-A10B, while the 2B and 9B models demonstrate strong language-conditioned 3D grounding after incorporating explicit 3D supervision. Second, we evaluate how large-scale embodied pretraining helps to learn downstream VLA policy. We perform evaluations of RynnBrain-VLA on three different platforms, a Unitree G1 humanoid[1], an Astribot-S1 bimanual robot[2], and a Tianji-Wuji dexterous-hand system[3]. Controlled evaluations on Astribot and Tianji-Wuji show that VLA policies built on our pretrained RynnBrain 1.1 outperform both the Qwen-based VLA policies and the most representative VLA generalists [6, 48]. Notably, joint multi-task and multi-embodiment training further improves the average process score and final success rate over separately fine-tuned per-task policies.

Overall, the contributions of this technical report are as follows:

- We release RynnBrain 1.1, a new family of embodied foundation models spanning 2B, 9B, and 122B-

[1] https://www.unitree.com/g1

[2] https://www.astribot.com/en/product

[3] https://www.tianjizn.com/products/marvin-series/, https://www.wuji.tech/zh/hand

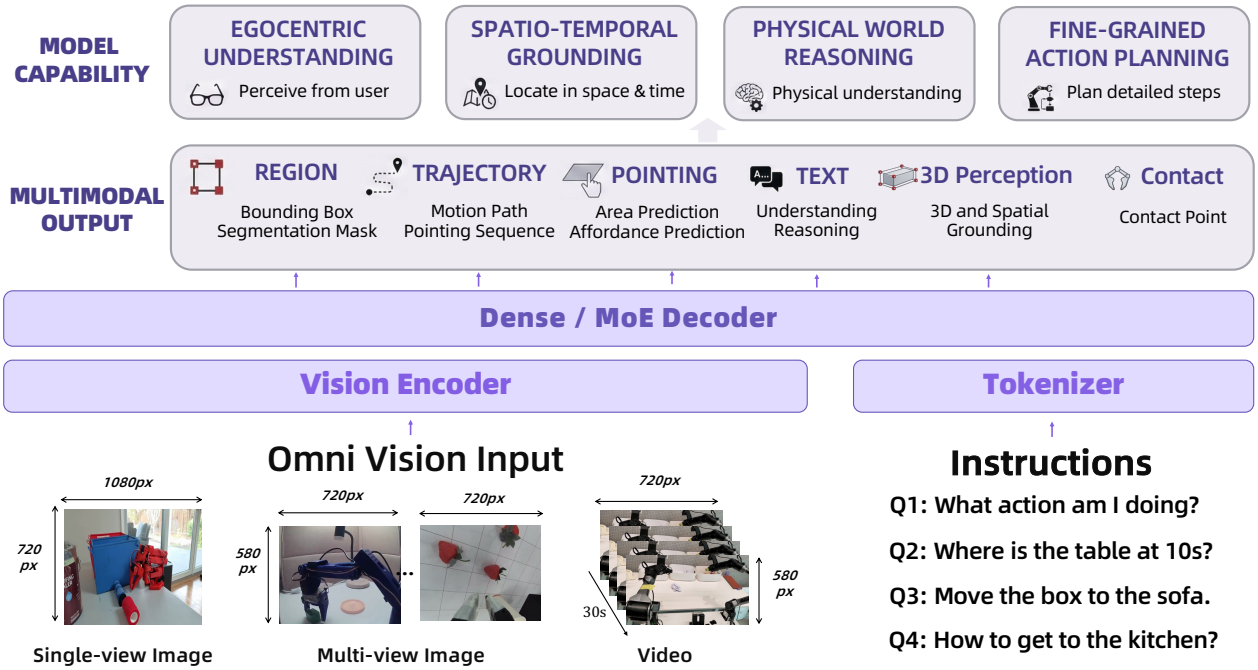

Figure 2 Overview of the RynnBrain 1.1 architecture. Given language instructions and omni-visual inputs—including single-view images, multi-view images, and videos—RynnBrain produces a unified set of multimodal outputs through either a dense or a mixture-of-experts decoder. These outputs cover text, regions, trajectories, pointing, 3D perception, and contact signals, enabling egocentric understanding, spatiotemporal grounding, physically grounded reasoning, and fine-grained action planning in real-world environments.

A10B parameter scales. Extensive evaluations across various benchmarks demonstrate a clear upward trend in model capabilities as the parameter scale increases. Notably, our RynnBrain 1.1-122B-A10B surpasses all proprietary models as well as open-source models on VSI-Bench, MMSI, and RefSpatial-Bench.

- We introduce two new features: (1) contact point prediction and (2) native 3D grounding (available for 2B and 9B-sized models) into embodied pretraining, yielding representations and outputs that are more directly aligned with robot manipulation.
- We develop RynnBrain-VLA, which couples a unified cross-embodiment action space with embodiment-specific masking. Beyond delivering strong real-world performance across heterogeneous embodiments, tasks, and control interfaces, it obtains further gains from joint multi-task and multi-embodiment training over separately fine-tuned per-task policies.

2 Overview

2.1 Model Architecture

An overview of the RynnBrain architecture is shown in Figure 2.

RynnBrain 1.1 adopts a decoder-only vision–language architecture following the design principles of Qwen3.5 [53]. It consists of a vision encoder, a vision–language projector, and a large language model (LLM) backbone initialized from Qwen3.5 variants. In this release, we build three model scales, namely RynnBrain 1.1-2B, RynnBrain 1.1-9B, and RynnBrain 1.1-122B-A10B. All models share the same overall architecture and training formulation, enabling a systematic study of embodied scaling across model sizes. We also adopt DeepStack [45] and Interleaved MRoPE [30] to better integrate multimodal information and support long-context spatio-temporal modeling over images, videos, and embodied observations.

2.2 Training Infrastructure

As described in RynnBrain [20], our training pipeline is designed for large-scale embodied multimodal data with highly heterogeneous sequence lengths. We follow the same infrastructure design, including online sequence-length-aware load balancing, per-sample loss reduction, memory-efficient training with ZeRO and gradient checkpointing, and optimized MoE execution with expert parallelism and efficient token dispatching. These components allow us to train on mixed image, video, text, and spatial-grounding data while mitigating straggler effects caused by long-tail sequence-length distributions.

For RynnBrain 1.1, we extend this infrastructure to the new Qwen3.5-based model family, covering 2B, 9B, and 122B scales. The same training pipeline is used across model sizes to ensure that differences in downstream performance mainly reflect model scaling and data updates rather than changes in infrastructure. For the 122B-A10B model, we adopt sparse MoE training with expert parallelism and communication-efficient token dispatching, following the infrastructure principles established in RynnBrain [20].

3 Pretraining

Similar to RynnBrain [20], our embodied pretraining framework is built around two key principles: (1) spatio-temporal memory and (2) physical-world grounding. Spatio-temporal memory enables the model to aggregate information from past visual observations, including objects, locations, events, and motion dynamics. Physical-world grounding, in turn, encourages the model to represent part of its understanding through explicit spatial outputs—such as objects, regions, affordances, and trajectories—rather than relying solely on unconstrained language generation. RynnBrain 1.1 advances this framework in two ways. First, we apply the same spatio-temporal embodied pretraining paradigm to three model scales—2B, 9B, and 122B-A10B—allowing us to examine how embodied perception, reasoning, localization, and planning change with model capacity. Second, as an initial step toward native 3D grounding, we add large-scale 3D-grounded data to the pretraining mixtures of RynnBrain 1.1-2B and RynnBrain 1.1-9B, enabling us to study the contribution of explicit 3D supervision to physical-space understanding at different model scales.

3.1 Training Recipe

Our pretraining formulation unifies temporal visual modeling and explicit spatial prediction within a single autoregressive framework. Multimodal observations are first encoded into a common token sequence, while both linguistic responses and grounded spatial targets are generated through the same prediction interface. This design equips RynnBrain with temporal memory and physically grounded representations without introducing separate task-specific decoders.

Unified Spatio-temporal Representation. A visual observation is represented as an ordered sequence of frames, $\mathbf{V} = \{I_t\}_{t=1}^{T}$. Static images correspond to the special case $T = 1$, whereas videos contain multiple temporally ordered frames with $T > 1$. For video inputs, we uniformly sample frames over time and encode each frame into visual tokens. Temporal position information is then attached to the resulting tokens so that the model can distinguish frame order and associate observations across time. By casting images and videos into the same sequential form, RynnBrain 1.1 can model object persistence, motion evolution, and long-range visual dependencies using a shared architecture.

Physically Grounded Output Space. Physical grounding is incorporated directly into the model's generation vocabulary. In addition to natural-language tokens, the output sequence may contain discrete representations of bounding boxes $\mathcal{B}$, points $\mathcal{P}$, and trajectory waypoints $\mathcal{T}$. Continuous image coordinates are first mapped to integers in $[0, 1000]$ and are then predicted autoregressively in the same manner as text tokens. As a result, semantic reasoning and spatial localization are learned jointly rather than handled by separate prediction heads. The 2B, 9B, and 122B-A10B models use this shared output specification in the main scaling experiments, while differences among them arise from model capacity and their corresponding training-data mixtures.

In addition, we introduce explicit 3D modeling task into the pretraining of RynnBrain 1.1-2B and RynnBrain 1.1-9B to explore native 3D grounding. These models follow the same instruction-following and autoregressive

training paradigm as the rest of the RynnBrain 1.1 family, while extending the supervision from image-plane grounding toward physical-space understanding.

Specifically, the model takes as input a natural language instruction and camera intrinsics, and is trained to directly predict a 3D bounding box parameterized in the camera coordinate system—including center position $(\mathrm{cx}, \mathrm{cy}, \mathrm{cz})$, box dimensions $(\mathrm{w}, \mathrm{l}, \mathrm{h})$, and orientation $(\mathrm{pitch}, \mathrm{yaw}, \mathrm{roll})$, all in physical units (meters and radians). This 9-dimensional output format follows the design of prior works such as Seed-VL [58] and Qwen3-VL [4], which also adopt a compact set of numerical parameters for spatially grounded predictions. To preserve the unified autoregressive framework, we discretize all continuous 3D quantities into integer tokens within a fixed range, analogous to our 2D spatial tokenization. This design enables the model to perform physically grounded reasoning from a single image, inferring not only what and where in the pixel plane, but also how objects are positioned and oriented in real-world 3D space, directly bridging visual perception and downstream robotic or navigation tasks.

Optimization. All RynnBrain 1.1 variants are optimized end-to-end with a causal autoregressive objective over sequences containing both language and spatial tokens. Given a visual input $\mathbf{V}$ and a target sequence $\mathbf{y} = \{y_i\}_{i=1}^{L}$, the model minimizes the negative log-likelihood

$$\mathcal{L} = -\sum_{i=1}^{L} \log P\left(y_i \mid y_{<i}, \mathbf{V}; \mathbf{\Theta}\right). \tag{1}$$

Here, L denotes the target-sequence length and $\mathbf{\Theta}$ denotes the learnable model parameters. The same optimization objective is used for the entire RynnBrain 1.1 family, including the 3D-grounded 2B and 9B variants. We select scale-specific hyperparameters through preliminary experiments on representative subsets of the pretraining corpus, with the complete configurations summarized in Table 1.

Table 1 Hyperparameters of the pretraining stage for the RynnBrain model series.

Parameter	RynnBrain 1.1-2B	RynnBrain 1.1-9B	RynnBrain 1.1-122B-A10B
Base Model	Qwen3.5-2B	Qwen3.5-9B	Qwen3.5-122B-A10B
Optimizer	AdamW	AdamW	AdamW
Learning Rate	5×10^{-6}	2×10^{-6}	2×10^{-6}
Learning Rate Vision	1×10^{-6}	2×10^{-6}	2×10^{-6}
Global Batch Size	512	1024	1024
Warmup Ratio	0.03	0.03	0.03

3.2 Pretraining Data

RynnBrain 1.1 follows the capability-oriented data organization of RynnBrain [20]. As shown in Table 2, we organize the pretraining mixture into general multimodal data, cognition data, spatio-temporal localization data, 3D-grounded data, robot-oriented perception data, and planning data. For data types already introduced in RynnBrain 1.0, we follow the same instruction-following format and data processing protocol. Specifically, visual inputs are represented as images or videos, and model outputs are converted into mixed sequences of text and spatial tokens, including boxes, points, trajectories, and grasp-related annotations. In this section, we focus on the data updates introduced in RynnBrain 1.1.

General MLLM Data. To preserve broad multimodal understanding, we continue to use a general-purpose image–video–text corpus covering image understanding, video understanding, visual question answering, temporal reasoning, and instruction following. The mixture includes commonly used MLLM datasets such as LLaVA-OV-SI [37], LLaVA-Video [78], ShareGPT-4o-video [13], VideoGPT-plus [43], FineVideo [24], CinePile [56], ActivityNet [8], YouCook2 [81], and LLaVA-SFT [40]. In RynnBrain 1.1, we further incorporate self-collected RynnBrain-Thinking data. RynnBrain-Thinking is included as part of the general mixture to better align the model with the Qwen3.5-Thinking style. All data are converted into the unified instruction-following format before being mixed for pretraining.

Table 2 Pretraining data mixture for RynnBrain 1.1. New or substantially expanded data sources compared with RynnBrain 1.0 are highlighted in blue.

Category	Sub-Task	Data Sources
General MLLM	General	LLaVA-OV-SI [37], LLaVA-Video [78], ShareGPT-4o-video [13], VideoGPT-plus [43], FineVideo [24], CinePile [56], ActivityNet [8], YouCook2 [81], LLaVA-SFT [40], VL3 [76], RynnBrain-Thinking
Cognition	Object Understanding	RynnBrain-Object, RefCOCO [72], Google Refexp [44], Osprey-724K [74], DAM [39], VideoRefer-700k [75]
	Spatial Understanding	SenseNova-SI-800K [10], VSI-590K [69], VLM-3R [22], RynnBrain-Spatial
	Counting	RynnBrain-Counting, Molmo2 [15]
	OCR	RynnBrain-OCR, Llama-Nemotron-VLM OCR [47]
	Egocentric Task Understanding	EgoRe-5M [52], EgoTaskQA [33], Env-QA [25], QAEgo4D [26], RoboVQA [59], Robo2VLM [12], ShareRobot [32]
Localization	Object Localization	ADE20K [79], COCOStuff [9], Mapillary [46], PACO-LVIS [54], PASCAL-Part [14], VG [35], RoboAfford-Object [27], RynnBrain-Grounding
	Area Localization	RefSpatial [80], RoboAfford-Area [27], Molmo2 [15], RynnBrain-Area, RoboRefer [80],
	Affordance Localization	RynnBrain-Affordance, RoboAfford-Affordance [27]
	Trajectory Prediction	RynnBrain-Trajectory, FSD [73], MolmoAct [36]
	3D Grounding	WildDet3D Essential [31], WildDet3D Synthetic [31], FoundationPose [67]
	Grasp / Contact Point Prediction	Grasp-Anything-Visible [66], GraspClutter6D [3], GraspNet-1B [23], Jacquard V2 [38], GraspFactory [61]
Planning	Manipulation	AgibotWorld [18], Open X-Embodiment [16], RynnBrain-Planning, Galaxea-G0 [34]

Multi-dimensional Cognition Data. For embodied cognition, RynnBrain 1.1 follows the task taxonomy of RynnBrain [20], covering object understanding, spatial understanding, counting, OCR, and egocentric task understanding. Object understanding data include RynnBrain-Object, RefCOCO [72], Google Refexp [44], Osprey-724K [74], DAM [39], and VideoRefer-700K [75], which provide object-centric supervision for fine-grained recognition and reasoning. Spatial understanding data include SenseNova-SI-800K [10], VSI-590K [69], VLM-3R [22], and RynnBrain-Spatial, supporting reasoning over spatial relations, viewpoints, and scene structure. For counting and OCR, we use RynnBrain-Counting, Molmo2 [15], RynnBrain-OCR, and the newly added Llama-Nemotron-VLM OCR data. For egocentric task understanding, we use EgoRe-5M [52], EgoTaskQA [33], Env-QA [25], QAEgo4D [26], RoboVQA [59], Robo2VLM [12], and ShareRobot [32]. These datasets are converted following the RynnBrain instruction format to provide broad supervision for object-centric, spatial, textual, and first-person task understanding.

Spatio-temporal Localization Data. RynnBrain 1.1 inherits the localization formulation of RynnBrain 1.0, where spatial outputs are represented using normalized coordinate tokens. The localization mixture covers object localization, area localization, affordance localization, trajectory prediction, and grasp/contact-related prediction. Object localization uses ADE20K [79], COCOStuff [9], Mapillary [46], PACO-LVIS [54], PASCAL-Part [14], VG [35], RoboAfford-Object [27], and RynnBrain-Grounding. Area localization uses RefSpatial [80], RoboAfford-Area [27], Molmo2 [15], RynnBrain-Area, and the newly added RoboRefer. Affordance localization uses RynnBrain-Affordance and RoboAfford-Affordance [27]. Trajectory prediction uses RynnBrain-Trajectory, FSD [73], and the newly added MolmoAct. For these data sources, we follow the RynnBrain processing protocol and convert different annotations into a shared output space consisting of boxes, points, and trajectories.

3D-grounded Data. To explore native 3D grounding, we conduct a 3D-grounded preview training in the 2B and 9B setting by mixing explicit 3D annotations into the pretraining data. Given an image, a natural language instruction, and camera intrinsics, the model is trained to predict a 3D bounding box in the camera coordinate system. The box is parameterized by its center position $(\mathrm{cx}, \mathrm{cy}, \mathrm{cz})$, dimensions $(\mathrm{w}, \mathrm{l}, \mathrm{h})$, and orientation $(\mathrm{pitch}, \mathrm{yaw}, \mathrm{roll})$, represented in physical units such as meters and radians.

To support this at scale, we construct a multi-source corpus combining large-scale automatically lifted annotations with high-fidelity synthetic data. For broad category and scene coverage, we adopt **WildDet3D-Data** [31], a recently proposed open-vocabulary 3D detection dataset that lifts 2D annotations from COCO, LVIS, Objects365, and V3Det into 3D via a multi-model candidate generation and selection pipeline. We use its two curated subsets after additional filtering: *Essential* (102,979 images, 374K manually verified annotations after filtering out geometrically implausible candidates and small objects, 12,064 categories) and *Synthetic* (896,004 images, 888K automatically selected annotations, 11,896 categories, filtered by a fine-tuned Molmo2 scoring six perceptual criteria with a retention threshold >10), offering unprecedented category diversity but inheriting residual noise from monocular lifting. For each 3D annotation, we leverage Qwen-VL to generate an open-ended referring query based on the corresponding 2D bounding box and category label, incorporating spatial descriptions (e.g., relative positions such as "left," "right," "front," or "behind") and attribute modifiers (e.g., color, size, or material cues) to make the referring expression more discriminative and precise for grounding. To complement this with physically exact supervision, we additionally incorporate **FoundationPose** [67], a synthetic dataset derived from the Google Scanned Objects subset under controlled camera trajectories, which provides 1,849K images with noise-free 3D bounding boxes computed directly from ground-truth object-to-camera transformations, with each asset assigned a natural-language category via Qwen-VL to bridge to real-world semantics. Though its category coverage is limited to 928 instances, its metric-scale accuracy and zero-depth-ambiguity geometry serve as a strong regularizer for depth and orientation prediction, teaching correct inductive biases for camera-to-world projection when mixed with the diverse but noisier WildDet3D annotations.

Contact Point Data. RynnBrain 1.0 represented planar grasps using the four corners of an oriented rectangle. We found this detector-style representation poorly aligned with action grounding in a general embodied model. A target object typically admits a set of functionally valid grasps rather than a unique rectangle. Moreover, unlike an object bounding box, a grasp rectangle has no canonical image-space extent: its width, height, aspect ratio, and boundary placement depend on gripper geometry and dataset-specific annotation conventions. Consequently, IoU against a single reference rectangle may penalize functionally equivalent grasps and is not necessarily correlated with execution success. The four-corner representation also introduces redundant extent variables and geometric consistency constraints into autoregressive generation. RynnBrain 1.1 therefore adopts a compact contact-centered representation $(a = (p, \theta))$, where $(p = (x, y))$ denotes the center of the annotated grasp configuration and (θ) is the in-plane angle of the line connecting the gripper fingers. The coordinates are normalized to [0,1000], and each target is serialized as <grasp pose> (x, y), θ </grasp pose>. This representation factors out annotation- and embodiment-dependent rectangle extents while preserving an action-relevant spatial anchor and orientation cue.

A substantial portion of our source data consists of simulation-native, multidimensional grasp annotations rather than directly usable image–text pairs. We therefore develop a view-conditioned rendering and projection pipeline to convert these annotations into reliable 2D supervision. For each object–grasp configuration, we select a camera pose subject to multiple observability constraints: the target object must remain fully inside the image, occupy an appropriate range of image scales, and avoid severe truncation or occlusion; meanwhile, the projected grasp center and orientation must both remain visible and spatially distinguishable. Views in which the object is excessively near, distant, truncated, or visually ambiguous, or in which the grasp configuration is not observable, are discarded. The retained views are rendered as RGB images, after which the multidimensional grasp annotations are projected into the common $((p, \theta))$ representation. This procedure is essential because unconstrained camera sampling can produce geometrically valid annotations that cannot be inferred from the rendered image.

After rendering, projection, and filtering, the resulting corpus contains 2.6M training samples from GraspAnything [66], Jacquard V2 [38], GraspFactory [61], GraspNet-1B [23], and GraspClutter6D [3]. We treat $((p, \theta))$ as an action-grounding interface rather than a complete executable grasp pose; depth, approach

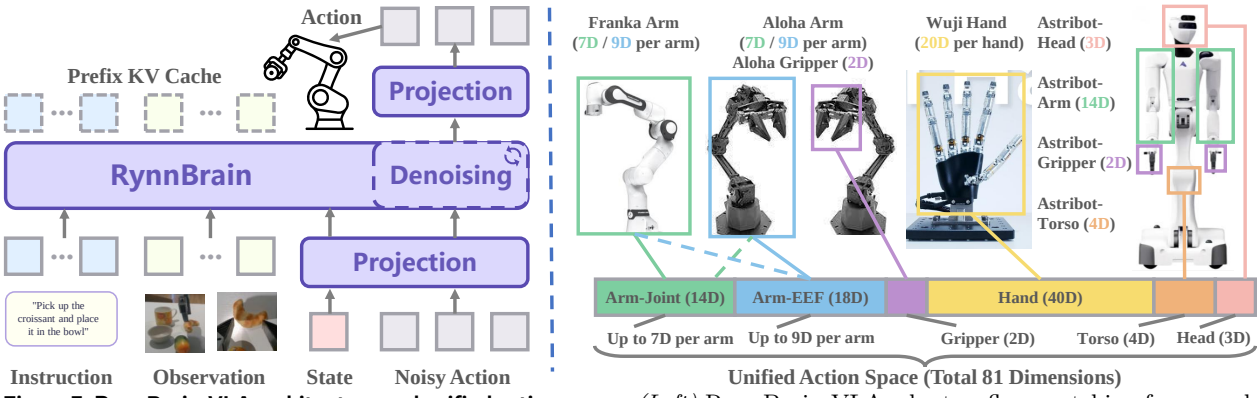

Figure 3 RynnBrain-VLA architecture and unified action space. *(Left)* RynnBrain-VLA adopts a flow-matching framework in which the RynnBrain base model serves as a single-stream Diffusion Transformer (DiT). The model takes a packed sequence of language instruction, multi-camera observations, robot state, and noisy action chunks as input, and iteratively denoises to predict an action chunk. The instruction token prefix is cached via KV cache to reduce inference latency. *(Right)* Actions from heterogeneous robot embodiments are represented within a shared 81-dimensional action space partitioned into semantically aligned body-part groups: Arm-Joint (14D, up to 7D per arm), Arm-EEF (18D, up to 9D per arm), Gripper (2D), Hand (40D), Torso (4D), and Head (3D). Each embodiment activates only its physically available dimensions via an embodiment-specific mask, enabling joint training across robots with incompatible low-level action spaces. In our evaluated embodiments: Unitree G1 activates Hand (14D, 7D per hand) within the unified space alongside a separately predicted 64D SONIC [42] latent token; Tianji-Wuji (Tianji Arm + Wuji Hand) activates Arm-Joint (14D) and Hand (40D); Astribot S1 activates Arm-Joint (14D), Gripper (2D), Head (3D) and Torso (4D). The robots shown in the figure are illustrative examples of the masking mechanism.

direction, gripper aperture, collision constraints, and robot kinematics are resolved by the downstream manipulation policy.

Physics-aware Planning Data. For manipulation planning, RynnBrain 1.1 follows the structured planning format of RynnBrain [20], where high-level task instructions are paired with grounded sub-task descriptions involving objects, areas, and affordances. The planning mixture includes AgibotWorld [18], Open X-Embodiment [16], RynnBrain-Planning, and the newly added Galaxea-G0. For existing grounded planning data, we follow the same annotation format and conversion protocol as RynnBrain 1.0. For Galaxea-G0, the original annotations provide sub-task-level labels but do not contain explicit high-level task instructions. We therefore construct the corresponding global task descriptions using Gemini 3.1 Pro, conditioned on the annotation IDs and the sequence of annotated sub-tasks. The resulting data are then converted into the same grounded planning format as the rest of the planning mixture.

4 Post-training for Vision-Language-Action Model

We use downstream VLA post-training as both a practical capability and a probe of base-model quality. This section describes the VLA *method*: the policy architecture (Section 4.1), the unified action space design (Section 4.2), the cross-embodiment / cross-driver deployment framework (Section 4.3), Real-Time Chunking for smooth low-latency control (Section 4.4), and the fine-tuning setup shared by all backbones (Section 4.5). The corresponding real-robot results are reported in Section 5.6.

4.1 Model Architecture

As shown in Figure 3, the model architecture of **RynnBrain-VLA** in RynnBrain-1.1 largely follows that in RynnBrain-1.0 [20], where the VLM backbone itself serves as a single-stream diffusion transformer and processes dense interaction among vision, language, and action in each layer. A flow matching framework is adopted to predict an action chunk [5] at each step. To preserve the VLM's inherent instruction-following capabilities, we utilize its native conversation format for organizing the input:

<|im_start|>user
INSTRUCTION:
Pour the wine from the bottle into the wine glass and place the glass on the table.
OBSERVATION:
<camera_1><camera_2><camera_3>
STATE:
<state>
What action should the robot take ?<|im_end|>
<|im_start|>assistant
<action>

Following π_0 [5], actions are positioned at the end of the sequence to enable the KV cache during inference.

4.2 Unified Cross-Embodiment Action Space

As illustrated in Figure 3, RynnBrain-VLA represents actions of heterogeneous robot embodiments within a shared high-dimensional action space. Before training, inspired by [41], we partition the unified action vector into semantically aligned groups corresponding to different body parts and control modes, including left and right arms, parallel grippers, dexterous hands, torso, and head control. Each embodiment activates only the action groups that are physically available on that robot, while all remaining dimensions are suppressed by an embodiment-specific mask.

The action loss is computed only on the active dimensions. For the Unitree G1 humanoid, the policy activates 14 dimensions (out of 40D) of the Hand group within the unified action space. These 14 dimensions correspond to the three-fingered dexterous hands of G1 with 7D per hand. In addition, the policy predicts a separate 64-dimensional SONIC [42] latent motion tokens that are not part of the unified action space. The 14-dimensional hand prediction and the 64-dimensional SONIC token are concatenated into a 78-dimensional representation and then fed to the SONIC whole-body controller to calculate target joint positions for the Unitree G1. For Tianji-Wuji, the policy activates the Arm-Joint group (14D, 7D per arm) and the Hand group (40D, 20D per hand), resulting in 54 active action dimensions. Astribot activates the Arm-Joint (14D, two arms), Gripper (2D), Head (3D), and Torso (4D), while masking the dexterous-hand joints.

This formulation allows trajectories from different robots to be mixed within the same training batch and optimized by a single VLA policy. The shared action components, such as left- and right-arm joint motion, receive supervision from multiple embodiments, whereas the embodiment-specific components are updated only by the corresponding data. Consequently, robot-independent manipulation features—including goal-directed arm motion, grasping, bimanual coordination, and task-level interaction patterns—can be shared across embodiments without requiring explicit alignment of their heterogeneous low-level action spaces. The unified action space not only serves as a common output interface, but also provides a mechanism for cross-embodiment knowledge sharing while preserving embodiment-specific control.

4.3 Cross-Embodiment and Cross-Driver Deployment

Real-world deployment requires a single policy interface to drive robots with very different configurations and low-level controllers. We therefore design a cross-embodiment, cross-driver framework that decouples the VLA policy from embodiment-specific execution, enabling the *same* RynnBrain-VLA recipe to be adapted to heterogeneous robots. The framework has three components:

- **Model layer.** Runs the VLA policy to produce a 32-step action chunk in the standard action space.
- **Control layer.** Two independent loops: a low-frequency loop (30 Hz) coordinating inference timing, the user interface, and action source switching; and a high-frequency loop (200 Hz) that interpolates the model's action chunk to the target control frequency and sends commands to the robot. The two loops communicate via shared memory.
- **Embodiment layer.** Dispatches standard-format actions (left/right arm, gripper, hand, torso) to the corresponding robot actuators and formats the robot's current state back for the model layer. Multi-

camera images are packaged as a dictionary whose keys match the naming convention used during training.

This design allows adding a new embodiment by only implementing the embodiment layer, without modifying the model or control layers. We instantiate this framework on three heterogeneous embodiments—the Unitree G1 humanoid, the Astribot-S1 bimanual manipulator, and the Tianji-Wuji dexterous-hand system—covering distinct action spaces, degrees of freedom, and end-effectors.

4.4 Real-Time Chunking (RTC)

We adopt **Real-Time Chunking (RTC)** [7]. The model predicts a 32-step action chunk but does not wait for the entire chunk to be executed: a new inference is triggered every 5 steps while the current chunk is still running. The unexecuted portion of the previous chunk is then used to guide the denoising of the new chunk, following the action-guidance formulation of Black et al. [7] (Sec. 3.1) with strength $\beta = 10.0$. Each step in the previous chunk is assigned a guidance weight reflecting whether it will be consumed before the new chunk arrives: (i) the initial 5 steps and the steps consumed during inference (estimated as the sliding average of the past ten inference times) receive weight = 1, as they are guaranteed to be executed before the new chunk is ready; (ii) a transition region where the weight decreases smoothly from 1 to 0; and (iii) the tail of the chunk, which receives weight = 0 with no consistency enforced. This ensures smooth chunk-boundary continuity while allowing the model to freely adapt later predictions to new observations.

4.5 Fine-tuning Setup

We collect teleoperation demonstrations on the three target embodiments described above, covering a range of pick-and-place and manipulation tasks with different objects and interaction requirements. Each demonstration contains synchronized visual observations, robot states, and action trajectories, together with a task instruction specifying the target object or placement location.

For the controlled comparison between the Qwen-based VLA and RynnBrain-VLA in Section 5.6, we use the same demonstration data, VLA post-training pipeline, and optimization settings. Both policies are fine-tuned for 60k steps with a learning rate of 2×10^{-5} and a batch size of 32. All input images are proportionally resized such that the short side is 384 pixels. This setup reduces differences introduced by downstream training and allows us to examine how the choice of base model affects the resulting real-robot policy. The fine-tuned policies are evaluated quantitatively in Section 5.6.

5 Evaluation

We organize our evaluation into several parts. We first assess the base models on embodied cognition and localization benchmarks across three model scales. At the 2B scale, we compare RynnBrain 1.1 with RynnBrain 1.0, Cosmos 3-Edge [49], SenseNova-SI-1.1-InternVL3-2B [11], Cosmos-Reason2-2B [2], and Qwen3.5-2B [53]. At the 4B–9B scale, we compare against RynnBrain 1.0, Cosmos 3-Nano [49], RoboBrain-2.5-4B [62], Thinker-4B [51], Molmo2-ER-5B [15], Pelican-VL-7B [77], MiMo-Embodied-7B [28], Cosmos-Reason2-8B [2], SenseNova-SI1.5-8B [11], and Qwen3.5 [53]. At the largest scale, we compare with Cosmos 3-Super, RoboBrain 2.0-32B, Cosmos-Reason2-32B, Pelican-VL-72B [77], Qwen3.5-122B-A10B [53], Gemini 3 Pro [63], GPT 5.4 [50], Claude Sonnet 4.6 [1], Gemini Robotics-ER 1.5 [64], and HY-Embodied 0.5 [65]. We then analyze scaling against the corresponding Qwen3.5 models, evaluate native 3D grounding and contact-point prediction, and finally examine RynnBrain-VLA through controlled real-robot evaluations across three embodiments

5.1 Embodied Cognition Capability

We evaluate RynnBrain 1.1 on a diverse suite of embodied cognition benchmarks covering video-based spatial intelligence, multi-view reasoning, embodied question answering, robot-oriented spatial understanding, and object-centric cognition. The evaluation includes VSI-Bench [68], MMSI [70], ERQA [64], RoboSpatial-VQA [60], MindCube [71], EmbSpatial [21], RynnBrain-Object, and RynnBrain-Spatial. As shown in Tables 3

Table 3 Comparison with lightweight (2B-scale) models on embodied cognition and localization benchmarks. * denotes results from our reproduction.

Benchmark	RynnBrain 1.1 2B	RynnBrain 2B	Cosmos 3 -Edge 2B	SenseNova-SI-1.1 InternVL3 2B	Cosmos -Reason2 2B	Qwen3.5 2B
VSI-Bench	**72.9**	70.5	59.2	63.7*	46.3*	44.1
MMSI	**40.5**	34.1	32.3	34.2	27.8*	28.3
ERQA	**42.3**	42.3	42.0	28.7*	37.8	33.0
RoboSpatial-Pointing	59.6	**65.6**	-	0.0*	11.5*	-
RoboSpatial-VQA	60.4	64.9	-	24.8*	**65.3***	-
RoboSpatial-Overall	62.1	**65.2**	-	16.1*	47.4*	41.5
MindCube	**61.7**	50.1	-	41.8	33.7*	37.5
EmbSpatial	73.9	**79.6**	-	63.0*	65.6*	67.7*
RynnBrain-Object	**74.6**	70.7	24.0	20.7	25.8	37.1*
RynnBrain-Spatial	**63.8**	57.2	22.6	13.0	24.3	15.6*
RefSpatial-Bench	**58.5**	52.7	48.4	5.6*	23.6*	30.0
RynnBrain-Grounding	**84.1**	79.1	51.6	12.6	14.5	24.2*
RynnBrain-Area	**57.2**	54.6	39.1	8.6	22.1	8.1*
RynnBrain-Affordance	**90.2**	89.4	77.6	57.8	55.5	80.4*
RynnBrain-Trajectory	**68.8**	66.6	58.7	49.0	43.7	28.3*

and 4, RynnBrain 1.1 improves substantially over the previous RynnBrain models on reasoning-intensive and spatially demanding tasks. At the 2B scale, the improvements are task-dependent: MMSI increases from 34.1 to 40.5, MindCube from 50.1 to 61.7, and RynnBrain-Spatial from 57.2 to 63.8, while several RoboSpatial and EmbSpatial results remain below RynnBrain-2B. The gains become more systematic at the 9B scale, where RynnBrain 1.1 outperforms RynnBrain-8B on eight of ten shared benchmarks. In particular, MindCube improves from 56.6 to 86.9, MMSI from 39.6 to 47.0, and RynnBrain-Spatial from 59.9 to 67.9, indicating stronger multi-view and spatial reasoning. Compared with other models at similar scales, RynnBrain 1.1 achieves leading results on multiple cognition benchmarks at both the 2B and 9B scales. RynnBrain 1.1-122B-A10B further extends this advantage to the large-model regime, with performance increasing from 40.5 to 47.0 and 52.0 on MMSI, and from 61.7 to 86.9 and 89.6 on MindCube, when scaling from 2B to 9B and 122B-A10B. These results show that model scaling is particularly beneficial for reasoning-intensive embodied cognition.

5.2 Embodied Localization Capability

We evaluate embodied localization across six spatial grounding tasks: RefSpatial-Bench [80],RoboSpatial-Pointing [60], RynnBrain-Grounding, RynnBrain-Area, RynnBrain-Affordance, and RynnBrain-Trajectory. These benchmarks cover language-guided spatial reference, object and region localization, affordance grounding, and trajectory prediction. RynnBrain 1.1 improves over the previous RynnBrain models on location benchmarks at both the 2B and 9B scales. At 2B, RefSpatial-Bench increases from 52.7 to 58.5 and RynnBrain-Grounding from 79.1 to 84.1. At 9B, the gains become larger on relational and motion-oriented tasks: RefSpatial-Bench improves from 59.2 to 67.2, while RynnBrain-Trajectory increases from 64.5 to 70.0. The relatively small gain on RynnBrain-Affordance reflects its already high performance in RynnBrain 1.0. Both RynnBrain 1.1-2B and RynnBrain 1.1-9B achieve the best results among models in their respective scale groups on most of these location benchmarks. At the large scale, RynnBrain 1.1-122B-A10B achieves leading performance on RefSpatial-Bench, RynnBrain-Grounding, and RynnBrain-Affordance. The clearest scaling trend appears on RefSpatial-Bench, which improves from 58.5 at 2B to 67.2 at 9B and 79.1 at 122B-A10B, showing that relational spatial grounding benefits strongly from increased model capacity.

Table 4 Comparison with 4B–9B scale models on embodied cognition and localization benchmarks. * denotes results from our reproduction.

Benchmark	RynnBrain 1.1 9B	RynnBrain 8B	Cosmos 3 -Nano 8B	RoboBrain-2.5 4B	Thinker 4B	Molmo2-ER 5B	Pelican-VL 7B	MiMo-Embodied 7B	Cosmos -Reason2 8B	SenseNova -SI1.5 8B	Qwen3.5 9B
VSI-Bench	**74.9**	70.9	54.9	41.7	65.4	74.5	52.8	48.5	53.7*	67.3*	61.0
MMSI	**47.0**	39.6	36.2	20.5	27.7*	43.8	26.2*	30.2*	31.3*	38.3*	13.4
ERQA	**47.5**	46.8	46.0	43.3	42.5*	46.8	39.8*	46.8	46.0*	40.0*	41.5*
RoboSpatial-Pointing	**69.7**	60.1	33.0*	-	-	32.0	-	-	24.6	29.3*	48.7
RoboSpatial-VQA	68.0	**80.0**	77.6*	-	-	73.4	-	-	76.3	61.4*	57.0
RoboSpatial-Overall	69.1	**73.1**	62.1*	62.3	70.8	59.0*	57.5	61.8	59.0*	40.1*	54.1*
MindCube	86.9	56.6	40.5*	26.9	39.0*	57.0	33.7*	43.1*	43.8*	**92.1**	33.1
EmbSpatial	81.9	80.4	81.4*	76.9	80.2	78.8	76.6	76.2	77.0*	80.3	**83.0**
RynnBrain-Object	**75.9**	71.2	39.8	36.4	26.5	17.1	30.8	39.0	37.2	16.8	45.6
RynnBrain-Spatial	**67.9**	59.9	26.1	32.6	17.3	19.7	20.5	28.3	31.4	19.9	28.4
RefSpatial-Bench	**67.2**	59.2	53.1	56.0	61.0	52.5	22.3	48.0	33.1*	8.6*	37.3
RynnBrain-Grounding	**85.4**	81.6	72.9	5.4	2.4	4.2	3.5	49.8	60.0	44.7	33.0
RynnBrain-Area	**59.6**	56.2	52.0	42.4	33.8	30.6	46.5	49.4	37.6	11.3	32.9
RynnBrain-Affordance	**90.7**	90.4	85.1	69.1	59.9	81.7	81.4	84.4	83.9	64.9	60.5
RynnBrain-Trajectory	**70.0**	64.5	67.9	63.0	49.2	59.3	59.2	61.3	64.0	53.9	45.0

5.3 Scaling Analysis

A central contribution of RynnBrain 1.1 is studying how embodied capability evolves with model scale under a unified training recipe. We compare RynnBrain 1.1 (2B, 9B, 122B-A10B) against the corresponding raw Qwen3.5 baselines at matched model sizes, grouping benchmarks into three categories: *general embodied cognition* (visual recognition, spatial understanding, and object-centric reasoning), *reasoning-intensive cognition* (multi-view spatial reasoning and video-based temporal inference), and *embodied localization* (language-guided spatial grounding and trajectory prediction). Figure 4 summarizes the scaling trends across these groups.

Embodied capabilities exhibit non-uniform scaling laws. Unlike standard vision–language benchmarks where scaling consistently improves performance, we observe three distinct scaling regimes in the embodied domain. (1) For *general embodied cognition*, both RynnBrain 1.1 and Qwen3.5 improve monotonically with scale, and the performance gap steadily narrows. This mirrors conventional MLLM scaling—larger models reliably improve on recognition and comprehension tasks, and general model scaling can partially compensate for the lack of embodied supervision. (2) For *reasoning-intensive cognition*, the two models diverge fundamentally: RynnBrain 1.1 improves steadily (+38.6%), while Qwen3.5 exhibits *negative scaling* (−39.2%), with the gap widening from 18.2 to 50.8. This indicates that multi-view and temporal reasoning capabilities are not emergent from general VLM scaling but require explicit embodied supervision as a prerequisite. Without spatiotemporal grounding, stronger language priors in larger models may override weak visual-spatial signals, producing more confident but less accurate responses. (3) For *embodied localization*, both models improve substantially, yet Qwen3.5 at its largest scale still underperforms RynnBrain 1.1 at its smallest. This suggests that for spatial-temporal pointing tasks, *data scaling*—training on explicit coordinate supervision—is more effective than model scaling alone. RynnBrain's embodied pretraining equips even its compact model with structured spatial output capabilities that pure language modeling must recover implicitly, requiring disproportionately more parameters to approximate.

Embodied pretraining as a scaling enabler. The decomposition in Figure 4(d) further illustrates this contrast. RynnBrain 1.1 consistently establishes a higher performance floor across all categories, reflecting the value of embodied data even at compact scales. More critically, on reasoning-intensive tasks, embodied pretraining determines whether scaling *helps or hurts*—it transforms a capability that degrades under pure VLM scaling into one that benefits consistently. These findings indicate that for the embodied domain, data-driven spatial supervision and model scaling are complementary rather than substitutable: scaling amplifies the benefit of embodied pretraining on reasoning-heavy tasks, while embodied supervision ensures that scaling produces

Table 5 Comparison with large-scale open-source and proprietary models on embodied cognition and localization benchmarks. * denotes results from our reproduction.

Benchmark	RynnBrain 1.1	Cosmos 3 Super	RoboBrain 2.0	Cosmos-Reason2	Pelican-VL	Qwen3.5	Gemini 3 Pro	GPT 5.4	Claude Sonnet 4.6	Gemini Robotics -ER 1.5	HY-Embodied 0.5
	122B-A10B	32B	32B	32B	72B	122B-A10B	-	-	-	-	A32B
VSI-Bench	**75.0**	60.9	42.7	67.0*	57.3	66.6*	48.8*	49.2*	44.4*	45.8	68.3
MMSI	**52.0**	41.8	28.5*	32.1*	30.7*	9.2*	49.2	37.5*	33.0*	-	39.2
ERQA	54.3	51.2	46.0	45.3	43.0	62.0	**70.5**	47.5*	44.8*	54.8	62.3
RoboSpatial-Pointing	**69.7**	24.2*	-	34.4	-	59.8	27.0*	20.8*	5.8*	31.1	-
RoboSpatial-VQA	70.6	74.1*	-	67.1	-	68.9	66.7*	75.0*	55.3*	**79.3**	-
RoboSpatial-Overall	70.3	56.7*	**72.4**	55.4	55.4	65.7*	53.0*	56.1*	38.0*	62.6	-
MindCube	**89.6**	40.2*	29.2	39.4*	32.5*	30.8*	70.8	10.4*	21.0*	54.7	69.2
EmbSpatial	83.7	**83.8***	73.8*	78.5*	76.6	80.5	83.6	81.8*	61.5*	78.4	-
RynnBrain-Object	**77.0**	48.3	26.2	36.8*	42.2	55.4	58.4	55.5	33.5	-	-
RynnBrain-Spatial	**70.0**	34.5	11.6	38.4*	32.2	31.0	38.5	43.3	36.7	-	-
RefSpatial-Bench	**79.1**	57.0	54.0	45.5*	49.5	69.3	65.5	26.7*	14.6*	48.5	57.2
RynnBrain-Grounding	**86.6**	74.4	0.0	60.0	10.8	66.6	62.1	24.2	3.3	-	-
RynnBrain-Area	60.7	53.0	45.3	37.6	53.2	41.9	**61.5**	37.1	15.8	-	-
RynnBrain-Affordance	**91.3**	84.6	76.1	83.9	87.3	79.0	86.0	83.4	60.0	-	-
RynnBrain-Trajectory	69.7	69.3	60.3	64.0	64.1	62.8	72.0	**72.8**	56.7	-	-

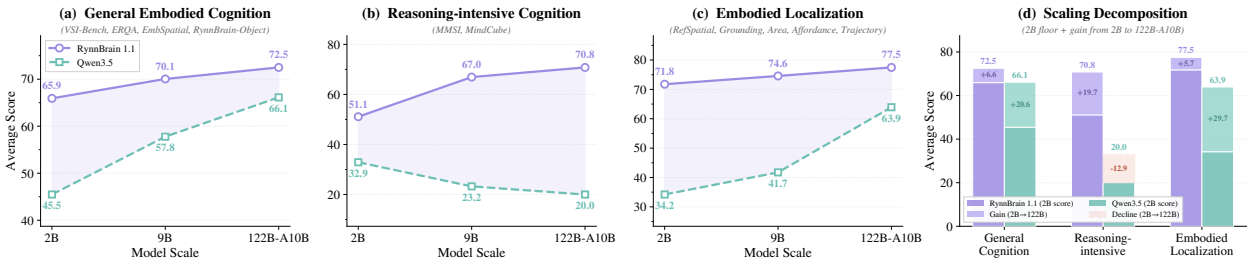

Figure 4 Scaling analysis of RynnBrain 1.1 vs. Qwen3.5 from 2B to 122B-A10B. (a)–(c) Average scores across three capability groups: general embodied cognition, reasoning-intensive cognition, and embodied localization. (d) Scaling decomposition showing the 2B performance floor and the gain from scaling to 122B-A10B.

reliable gains rather than regression.

5.4 3D Grounding

We evaluate the native 3D grounding capability of RynnBrain 1.1 at both the 2B and 9B scales. On SUN RGB-D, RynnBrain 1.1-2B achieves 34.28 AP@15, already outperforming general-purpose VLMs such as Seed1.5-VL (33.5) and Gemini 2.0 Pro (32.5). Scaling the model to 9B further improves the result to **41.12 AP@15**, substantially narrowing the gap to the closed-source Gemini Robotics-ER (48.3). These results show that explicit 3D-grounded training can equip even compact models with metric spatial understanding, while larger model capacity further strengthens depth, size, and orientation reasoning.

A similar scaling trend is observed on WildDet3D-Bench. RynnBrain 1.1-2B obtains 17.36 AP3D, while RynnBrain 1.1-9B reaches **23.44 AP3D**, surpassing the specialized WildDet3D detector trained with additional in-domain data (22.6). The improvement from 2B to 9B indicates that native 3D grounding benefits from model scaling in addition to explicit 3D supervision. Overall, RynnBrain 1.1 demonstrates strong language-conditioned 3D grounding across model sizes, with the 9B model achieving competitive performance against both specialized detectors and substantially larger proprietary systems.

In addition to the quantitative evaluation, we visualize representative predictions in Figure 6. RynnBrain 1.1 can localize target objects and estimate their spatial extent and orientation from a single image, including large furniture such as beds, tables, and chairs. These examples illustrate the model's ability to produce physically grounded 3D predictions across diverse viewpoints and scene layouts.

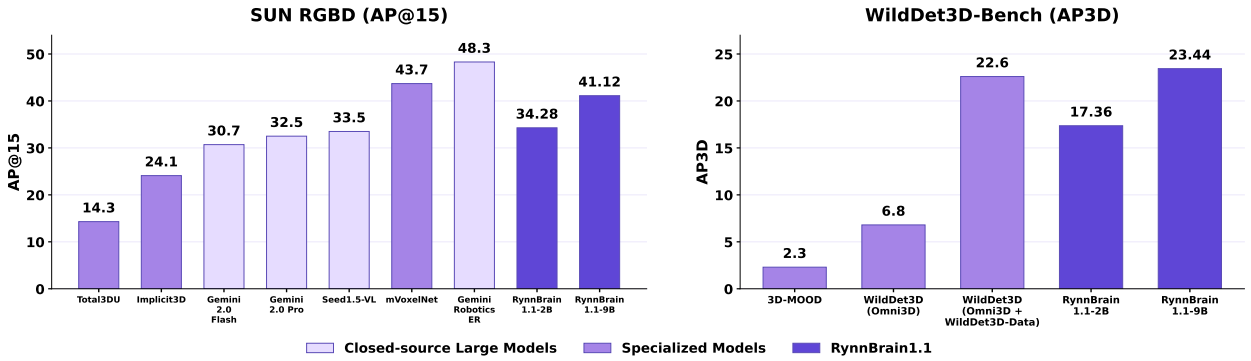

Figure 5 Performance Comparison on SUN RGB-D and WildDet3D-Bench with RynnBrain 1.1-2B and 9B.

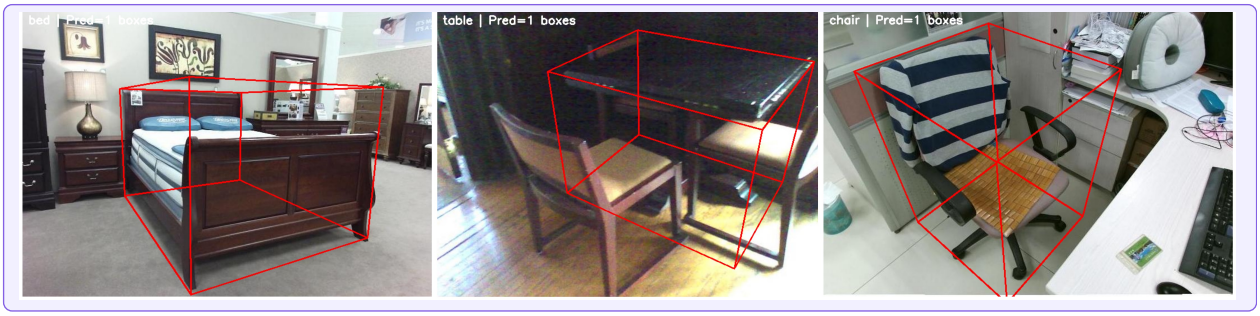

Figure 6 Qualitative examples of language-conditioned 3D grounding by RynnBrain 1.1. Given an image and a target object description, the model predicts an oriented 3D bounding box in the camera coordinate system. The examples cover objects with different scales, shapes, viewpoints, and levels of occlusion in indoor environments.

5.5 Contact Point Prediction

Evaluating contact point prediction remains challenging because there is currently no standardized metric that reliably reflects the functional validity of a predicted contact. Pixel-level distance to a single annotation may penalize alternative contact points that are equally valid for manipulation, while conventional grasp-box metrics are not directly applicable to our contact-centered representation. We therefore provide a qualitative evaluation in Figure 7 to illustrate the capability of RynnBrain 1.1 across diverse objects and instructions. Here we show the cases produced by RynnBrain 1.1-9B.

The visualized cases show that RynnBrain 1.1 can identify task-relevant contact locations rather than merely predicting object centers. For example, the model selects the handle when asked to pick up a mug, adapts its prediction when the mug is overturned, and produces appropriate contact locations and orientations for thin objects, boxes, and irregularly shaped items. It also distinguishes the instructed target in cluttered or referential settings. These results demonstrate that RynnBrain 1.1 can jointly reason about object identity, geometry, state, and intended interaction when generating contact-oriented spatial predictions.

5.6 Real-Robot Evaluation

We evaluate RynnBrain-VLA on real robots to answer three questions: (i) can RynnBrain-VLA achieve stable performance on real-world manipulation tasks? (ii) does a stronger embodied foundation model lead to a better VLA policy? and (iii) can joint training on data from heterogeneous robot embodiments provide mutual benefits across embodiments?

5.6.1 Experimental Setup

Evaluation Protocol. We deploy RynnBrain-VLA on three heterogeneous embodiments, as summarized in Table 6 and Figure 8. For each task, we conduct 20 trials with randomized initial object placements and report the final task success rate under a fixed success criterion. Since some long-horizon tasks can be

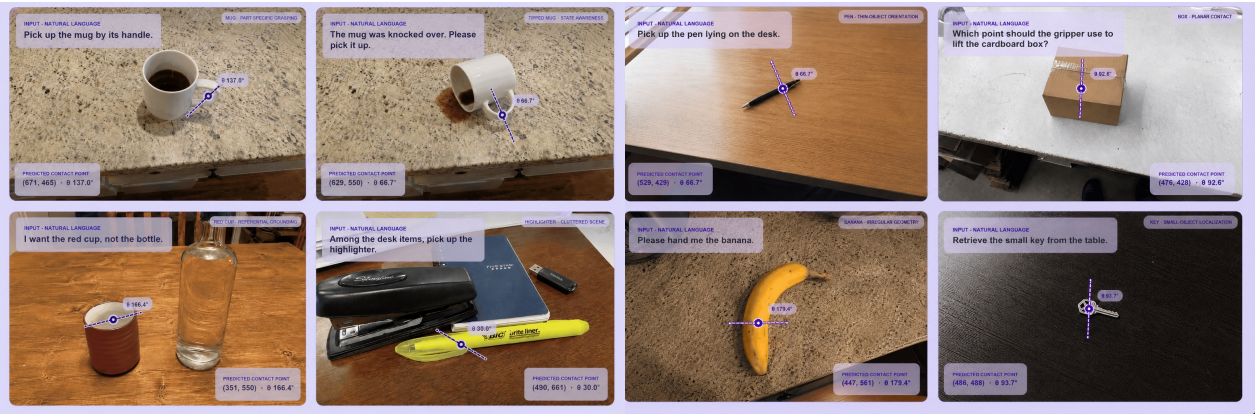

Figure 7 Qualitative results of contact point prediction by RynnBrain 1.1. Given an image and a natural-language instruction, the model predicts an action-relevant contact point together with its in-plane grasp orientation. The examples cover diverse object geometries, object states, target parts, and cluttered scenes, demonstrating instruction-conditioned contact grounding beyond object-level localization. Coordinates are normalized to $[0, 1000]$.

naturally decomposed into multiple sub-tasks, final success alone does not fully reflect the execution progress of a policy. We therefore additionally report a process score that measures the fraction of completed sub-tasks. For a task with K sub-tasks, the process score is defined as

$$\text{ProcessScore} = \frac{1}{20K} \sum_{i=1}^{20} \sum_{j=1}^{K} s_{i,j}, \tag{2}$$

where $s_{i,j} \in \{0, 1\}$ indicates whether the j-th sub-task is successfully completed in the i-th trial. Sprinkle the Petals, Pour the Wine, and Grab the Spatulas contain 4, 5, 3 subtasks, respectively. This metric provides a more fine-grained measure of task progress, especially for long-horizon manipulation tasks in which a policy may complete several intermediate stages without reaching the final goal.

Deployment and Action Space. For all tasks, inference is performed locally on a workstation equipped with a single NVIDIA GeForce RTX 4090 GPU. The workstation is directly connected to the robot and transmits the predicted actions for real-time execution. For tasks involving the Unitree G1 humanoid, we employ SONIC as the whole-body controller. The model predicts a 64-dimensional SONIC latent motion token alongside a 14-dimensional dexterous-hand prediction corresponding to the G1's three-finger-joint hand, which is represented within the unified action space's Hand group. These two outputs are concatenated into a 78-dimensional action representation and decoded by the SONIC controller into target positions for the actuated joints of the Unitree G1. The 64-dimensional SONIC token is a G1-specific output that lies outside the 81-dimensional unified action space and is used only for the Unitree G1. For tasks involving the Tianji-Wuji system (Tianji Arm and Wuji Hand), the policy activates the Arm-Joint group (14D, two arms) and the Hand group (40D, two dexterous hands), yielding a 54-dimensional active action representation that is sent directly to the Tianji-Wuji controller. For tasks involving the Astribot S1, the policy activates the Arm-Joint (14D), Gripper (2D), Head (3D) and Torso (4D), with the dexterous-hand dimensions masked out.

Baselines and Training Settings. We select GR00T N1.7 [48] and π0.5 [6] as representative generalist VLA baselines. We additionally construct Qwen-Based-VLA by applying the same VLA post-training recipe used for RynnBrain-VLA to the Qwen base model. For RynnBrain-VLA, we consider two training settings. In the single-task setting, a separate policy is fine-tuned for each individual task. In the generalist setting, we jointly fine-tune a single policy on multi-task and multi-embodiment data using our unified action space, and denote the resulting model as RynnBrain-VLA$_{\text{Generalist}}$. To obtain the best deployment performance for each policy, we use method-specific action-chunk lengths rather than enforcing a shared chunk size across all models. Specifically, we use an action-chunk length of 40 for GR00T N1.7, 50 for π0.5, and 32 for RynnBrain-VLA, including both the single-task and generalist settings, and Qwen-Based-VLA.

Tasks. We design three challenging long-horizon manipulation tasks for quantitative comparison: *Sprinkle*

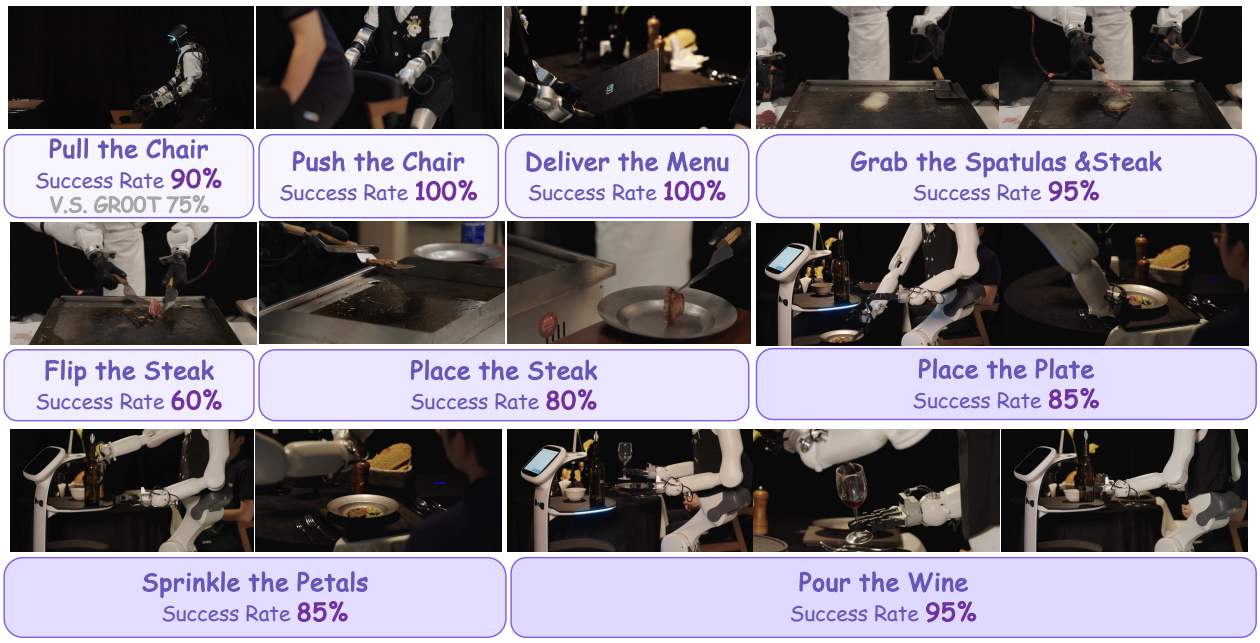

Figure 8 Real-robot evaluation across all tasks considered in our experiments. The tasks are performed in diverse real-life environments and cover whole-body manipulation, bimanual coordination, dexterous-hand operation, object placement, furniture interaction, and object delivery. Representative execution snapshots and success rates over 20 trials are shown for each task.

the Petals, *Pour the Wine*, and *Grab the Spatulas*. Each task consists of multiple temporally ordered subtasks and requires the policy to complete a sequence of perception, reaching, grasping, manipulation, and placement operations. The task setups are illustrated in Figure 8. The first two tasks are performed on the Astribot robot with gripper-based whole-body control, where the robot needs to actively change its viewpoint during execution. In *Sprinkle the Petals*, the policy must continuously observe the container and determine whether the petals have been fully poured out before proceeding, requiring closed-loop visual observation and task-progress understanding. *Pour the Wine* requires the robot to precisely align the bottle opening with the target glass and execute a controlled pouring motion, placing stronger demands on fine-grained affordance understanding and precise manipulation. *Grab the Spatulas* is performed with Tianji-Wuji dexterous hands and requires fine-grained grasping and hand control to reliably manipulate the tools. Together, these tasks cover complementary challenges in long-horizon reasoning, visual feedback, precise affordance-guided manipulation, and dexterous interaction, and serve as the main benchmark for comparing different VLA methods and training settings. Beyond the three main tasks, we further evaluate RynnBrain-VLA on a series of additional robot tasks across heterogeneous embodiments, including object placement, flipping, furniture interaction, and object delivery. These evaluations broaden the coverage of embodiments and manipulation primitives and provide an additional assessment of policy generality.

5.6.2 Stable Real-World Performance

As shown in Figure 8 and Table 6, RynnBrain-VLA achieves stable performance across multiple robot embodiments, including the Astribot bimanual platform, the Tianji-Wuji dexterous-hand system, and the Unitree G1 humanoid. The policy performs reliably under substantially different action spaces and control interfaces, covering gripper-based whole-body manipulation, dexterous grasping, bimanual coordination, and humanoid whole-body control. In particular, the Astribot tasks require the robot to actively change its viewpoint during execution and continuously update its actions from visual feedback. The strong results on these tasks indicate that RynnBrain-VLA remains robust to viewpoint changes rather than relying on a fixed observation configuration.

RynnBrain-VLA also performs well on the Unitree G1 humanoid. On *Pull the Chair*, which involves coordinated humanoid whole-body motion, RynnBrain-VLA achieves a 90% success rate, compared with 75% for

Table 6 Real-robot evaluation results. Each task reports the process score and final success rate (%). The process score measures the average fraction of completed sub-tasks, while a trial is counted as successful only when all sub-tasks of the corresponding task are completed.

Model	Sprinkle the Petals *Astribot*		Pour the Wine *Astribot*		Grab the Spatulas *Tianji-Wuji*		Average –	
	Proc.	**Succ.**	**Proc.**	**Succ.**	**Proc.**	**Succ.**	**Proc.**	**Succ.**
Qwen-Based-VLA	75.00	65.00	75.00	65.00	55.00	50.00	68.33	60.00
GR00T N1.7	86.25	75.00	77.00	65.00	86.67	80.00	83.31	73.33
$\pi_{0.5}$	65.00	60.00	69.00	65.00	83.33	70.00	72.44	65.00
RynnBrain-VLA	87.50	**85.00**	88.00	80.00	**98.33**	**95.00**	91.28	86.67
RynnBrain-VLA$_{\text{Generalist}}$	**88.75**	**85.00**	**97.00**	**95.00**	96.67	**95.00**	**94.14**	**91.67**

GR00T N1.7 combined with the same SONIC controller. Since GR00T N1.7 with SONIC is a natural baseline for predicting SONIC motion tokens, this comparison provides additional evidence that RynnBrain-VLA can produce reliable actions even for challenging whole-body control tasks. Overall, the results demonstrate that its real-world performance is stable not only across tasks, but also across robot morphologies, viewpoints, and control configurations.

5.6.3 A Stronger Embodied Base Yields a Better VLA

We next investigate whether an embodied foundation model provides a stronger initialization for VLA post-training. RynnBrain-VLA and Qwen-Based-VLA follow the same VLA post-training recipe and use the same action-chunk length, while being initialized from RynnBrain and Qwen Based Model, respectively. As shown in Table 6, RynnBrain-VLA consistently outperforms Qwen-Based-VLA across all three long-horizon tasks. Its average process score improves from 68.33% to 91.28%, while the average final success rate increases from 60.00% to 86.67%. The advantage is particularly pronounced on *Grab the Spatulas*, where RynnBrain-VLA achieves a 95% success rate compared with 50% for Qwen-Based-VLA. These results suggest that the capabilities acquired during embodied foundation-model training transfer effectively to downstream VLA post-training. Compared with directly post-training a general-purpose VLM as a VLA policy, initializing from RynnBrain leads to more reliable execution of multi-stage manipulation tasks and substantially fewer failures during intermediate stages. This comparison provides direct evidence that a stronger embodied base model can serve as a better foundation for real-world VLA policies.

5.6.4 Cross-Embodiment Generalization

To examine whether the capability of RynnBrain-VLA is tied to a specific robot embodiment, we adapt it to three morphologically and mechanically distinct robot platforms through the deployment framework described in Section 4.3. As shown in Table 6 and Figure 8, RynnBrain-VLA achieves stable performance across humanoid robots, dexterous-hand manipulation, whole-body control, and bimanual grasping settings. These embodiments differ substantially in morphology, action dimensionality, and low-level control interfaces. The consistent performance across these settings suggests that the embodied representations learned by RynnBrain capture task- and interaction-level knowledge that is not tightly coupled to a particular robot morphology or action parameterization, allowing the same embodied foundation model to serve as a shared starting point for heterogeneous robot policies.

More interestingly, we find that jointly training multiple tasks and embodiments within a single VLA does not introduce obvious interference. With the unified action-space strategy in Section 4.2, RynnBrain-VLA$_{\text{Generalist}}$ improves the average process score from 91.28% to 94.14% and the final success rate from 86.67% to 91.67% compared with separately fine-tuned single-task policies. We hypothesize that different embodiments share common interaction structures—such as object affordances, task progress, and reach–grasp–manipulate patterns—even though their executable actions are different. Joint training therefore allows these shared structures to reinforce a common visual–action representation, while unified action-space

preserves embodiment-specific control. This suggests that cross-embodiment data can act as complementary supervision: the key benefit of a unified VLA may lie not only in supporting more robots with one model, but also in allowing different embodiments to teach each other how the physical world can be interacted with.

6 Conclusion and Future Work

We presented **RynnBrain 1.1**, an embodied foundation model family spanning 2B, 9B, and 122B-A10B scales. RynnBrain 1.1 studies embodied capability scaling under a unified training framework and demonstrates clear improvements in multi-view reasoning, spatial grounding, and robot-oriented understanding. We further introduce explicit 3D-grounded supervision for the 2B and 9B models, enabling native language-conditioned 3D grounding, together with a compact contact-centered representation for interaction grounding.

We also develop RynnBrain-VLA to connect embodied understanding with real-world action. Under the same post-training recipe, RynnBrain-VLA substantially outperforms a Qwen-based VLA on long-horizon manipulation tasks. Through a unified action space with embodiment-specific masking, the policy is deployed across a Unitree G1 humanoid, an Astribot bimanual robot, and a Tianji-Wuji dexterous-hand system. Joint multi-task and multi-embodiment training further improves real-robot performance, providing initial evidence that heterogeneous robot data can contribute complementary supervision to a shared policy.

Looking forward, we aim to extend RynnBrain toward a unified multimodal model that combines perception, understanding, generation, spatial grounding, and action within a common framework. An important direction is to develop RynnBrain into the cognitive core of a general embodied agent, integrating long-term memory, world modeling, task planning, active perception, and closed-loop interaction. Such an agent could continuously reason about changes in the physical world, generate and revise plans from feedback, and acquire new skills through interaction across diverse tasks and robot embodiments.

References

[1] Anthropic. Introducing claude sonnet 4.6, February 2025. https://www.anthropic.com/news/claude-sonnet-4-6.

[2] Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

[3] Seunghyeok Back, Joosoon Lee, Kangmin Kim, Heeseon Rho, Geonhyup Lee, Raeyoung Kang, Sangbeom Lee, Sangjun Noh, Youngjin Lee, Taeyeop Lee, et al. Graspclutter6d: A large-scale real-world dataset for robust perception and grasping in cluttered scenes. *IEEE Robotics and Automation Letters*, 2025.

[4] 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, 2025. https://arxiv.org/abs/2511.21631.

[5] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[6] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y Galliker, et al. $\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization. In *9th Annual Conference on Robot Learning*, 2025.

[7] Kevin Black, Manuel Galliker, and Sergey Levine. Real-time execution of action chunking flow policies. *Advances in Neural Information Processing Systems*, 38:33383–33407, 2026.

[8] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[9] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[10] Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, et al. Scaling spatial intelligence with multimodal foundation models. *arXiv preprint arXiv:2511.13719*, 2025.

[11] Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[12] Kaiyuan Chen, Shuangyu Xie, Zehan Ma, Pannag R Sanketi, and Ken Goldberg. Robo2vlm: Visual question answering from large-scale in-the-wild robot manipulation datasets. *arXiv preprint arXiv:2505.15517*, 2025.

[13] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024.

[14] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1978, 2014.

[15] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.

[16] Open X-Embodiment Collaboration, Abby O'Neill, and et.al. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

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

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

[19] Ronghao Dang, Yuqian Yuan, Wenqi Zhang, Yifei Xin, Boqiang Zhang, Long Li, Liuyi Wang, Qinyang Zeng, Xin Li, and Lidong Bing. Ecbench: Can multi-modal foundation models understand the egocentric world? a holistic embodied cognition benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[20] Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, Jiangpin Liu, Yunxuan Mao, Zhikai Wang, Yuqian Yuan, Minghao Zhu, Xiao Lin, Yang Bai, Qian Jiang, Yaxi Zhao, Minghua Zeng, Junlong Gao, Yuming Jiang, Jun Cen, Siteng Huang, Liuyi Wang, Wenqiao Zhang, Chengju Liu, Jianfei Yang, Shijian Lu, and Deli Zhao. Rynnbrain: Open embodied foundation models. *arXiv preprint arXiv:2602.14979*, 2026. https://arxiv.org/abs/2602.14979.

[21] Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing 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, 2024.

[22] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

[23] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020.

[24] Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024.

[25] Difei Gao, Ruiping Wang, Ziyi Bai, and Xilin Chen. Env-qa: A video question answering benchmark for comprehensive understanding of dynamic environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1675–1685, October 2021.

[26] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

[27] Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang, Yanbiao Ma, Yunfeng Diao, Ziyu Jia, Wenbo Ding, Hangjun Ye, and Long Chen. Roboafford++: A generative ai-enhanced dataset for multimodal affordance learning in robotic manipulation and navigation. *arXiv preprint arXiv:2511.12436*, 2025.

[28] Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, Guang Li, Zheng Lu, Shuhuai Ren, Xianhui Meng, et al. Mimo-embodied: X-embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*, 2025.

[29] Bohan Hou, Jiuning Gu, Jiayan Guo, Ronghao Dang, Sicong Leng, Xin Li, Xuemeng Song, and Jianfei Yang. Interlv-search: Benchmarking interleaved multimodal agentic search. *arXiv preprint arXiv:2605.07510*, 2026.

[30] Jie Huang, Xuejing Liu, Sibo Song, Ruibing Hou, Hong Chang, Junyang Lin, and Shuai Bai. Revisiting multimodal positional encoding in vision-language models. *arXiv preprint arXiv:2510.23095*, 2025.

[31] Weikai Huang, Jieyu Zhang, Sijun Li, Taoyang Jia, Jiafei Duan, Yunqian Cheng, Jaemin Cho, Matthew Wallingford, Rustin Soraki, Chris Dongjoo Kim, Shuo Liu, Donovan Clay, Taira Anderson, Winson Han, Ali Farhadi, Bharath Hariharan, Jason Ren, and Ranjay Krishna. Wilddet3d: Scaling promptable 3d detection in the wild. *arXiv preprint arXiv:2604.08626*, 2026.

[32] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:2502.21257*, 2025.

[33] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35:3343–3360, 2022.

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

[35] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[36] 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. *arXiv preprint arXiv:2508.07917*, 2025.

[37] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[38] Qiuhao Li and Shenghai Yuan. Jacquard v2: Refining datasets using the human in the loop data correction method. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7932–7938. IEEE, 2024.

[39] Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, et al. Describe anything: Detailed localized image and video captioning. *arXiv preprint arXiv:2504.16072*, 2025.

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[41] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, pages 29982–30009, 2025.

[42] Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025.

[43] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024.

[44] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[45] Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems*, 37:23464–23487, 2024.

[46] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017.

[47] NVIDIA. Llama-nemotron-vlm-dataset-v1. https://huggingface.co/datasets/nvidia/Llama-Nemotron-VLM-Dataset-v1, 2025.

[48] NVIDIA, Johan Bjorck, Nikita Cherniadev Fernando Castañeda, 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. In *ArXiv Preprint*, March 2025.

[49] NVIDIA et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

[50] OpenAI. Introducing GPT-5.4, March 2025. https://openai.com/index/introducing-gpt-5-4/.

[51] Baiyu Pan, Daqin Luo, Junpeng Yang, Jiyuan Wang, Yixuan Zhang, Hailin Shi, and Jichao Jiao. Thinker: A vision-language foundation model for embodied intelligence. *https://arxiv.org/abs/2601.21199*, 2025.

[52] Baoqi Pei, Yifei Huang, Jilan Xu, Yuping He, Guo Chen, Fei Wu, Yu Qiao, and Jiangmiao Pang. Egothinker: Unveiling egocentric reasoning with spatio-temporal cot. *arXiv preprint arXiv:2510.23569*, 2025.

[53] Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

[54] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023.

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

[56] Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024.

[57] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the" edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024.

[58] Bytedance Seed. Seed1.8 model card: Towards generalized real-world agency, 2026. https://arxiv.org/abs/2603.20633.

[59] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 645–652. IEEE, 2024.

[60] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15768–15780, 2025.

[61] Srinidhi Kalgundi Srinivas, Yash Shukla, Adam Arnold, and Sachin Chitta. Graspfactory: A large object-centric grasping dataset. *arXiv preprint arXiv:2509.20550*, 2025.

[62] Huajie Tan, Enshen Zhou, Zhiyu Li, Yijie Xu, Yuheng Ji, Xiansheng Chen, Cheng Chi, Pengwei Wang, Huizhu Jia, Yulong Ao, et al. Robobrain 2.5: Depth in sight, time in mind. *arXiv preprint arXiv:2601.14352*, 2026.

[63] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[64] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[65] Tencent Robotics X, HY Vision Team, Xumin Yu, Zuyan Liu, Ziyi Wang, He Zhang, Yongming Rao, Fangfu Liu, Yani Zhang, Ruowen Zhao, Oran Wang, Yves Liang, Haitao Lin, Minghui Wang, Yubo Dong, Kevin Cheng, Bolin Ni, Rui Huang, Han Hu, Zhengyou Zhang, Linus, and Shunyu Yao. Hy-embodied-0.5: Embodied foundation models for real-world agents. *arXiv preprint arXiv:2604.07430*, 2026.

[66] An Dinh Vuong, Minh Nhat Vu, Hieu Le, Baoru Huang, Huynh Thi Thanh Binh, Thieu Vo, Andreas Kugi, and Anh Nguyen. Grasp-anything: Large-scale grasp dataset from foundation models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14030–14037. IEEE, 2024.

[67] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[68] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.

[69] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025.

[70] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025.

[71] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

[72] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pages 69–85. Springer, 2016.

[73] Yifu Yuan, Haiqin Cui, Yibin Chen, Zibin Dong, Fei Ni, Longxin Kou, Jinyi Liu, Pengyi Li, Yan Zheng, and Jianye Hao. From seeing to doing: Bridging reasoning and decision for robotic manipulation, 2025. https://arxiv.org/abs/2505.08548.

[74] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28202–28211, 2024.

[75] Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18970–18980, 2025.

[76] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[77] Yi Zhang, Che Liu, Xiancong Ren, Hanchu Ni, Shuai Zhang, Zeyuan Ding, Jiayu Hu, Hanzhe Shan, Zhenwei Niu, Zhaoyang Liu, et al. Pelican-vl 1.0: A foundation brain model for embodied intelligence. *arXiv preprint arXiv:2511.00108*, 2025.

[78] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[79] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[80] 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.

[81] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.