GEM: Generative Supervision Helps Embodied Intelligence

Ruowen Zhao[1], Bangguo Li[1], Zuyan Liu[1,2,†], Yinan Liang[1], Junliang Ye[1], Fangfu Liu[1], Diankun Wu[1], Zhengyi Wang[1], Xumin Yu[2], Yongming Rao[2,‡], Han Hu[2], and Jun Zhu[1,‡]

[1] Tsinghua University [2] Tencent Hunyuan

Abstract. Embodied Vision-Language Models (VLMs) have demonstrated impressive performance and generalization in robotics, particularly within Vision-Language-Action frameworks. However, a significant gap remains between the high-level semantic focus of standard text-guided pre-training paradigms and the low-level spatial and physical knowledge critical for execution in embodied environments. In this paper, we introduce **GEM**, a **G**enerative-supervised **Em**bodied vision-language Model designed to bridge this divide. We propose integrating a depth map generation task directly into the VLM pre-training phase. By training this generative objective jointly with the main model, we observe substantial improvements in embodied intelligence, significantly enhancing both semantic understanding and physical operation capabilities. To support this paradigm, we curate and release GEM-4M, a comprehensive large-scale dataset featuring a mixture of grounding, reasoning, and planning data paired with high-quality depth supervision. Extensive experiments demonstrate that GEM achieves state-of-the-art results across diverse embodied benchmarks. Furthermore, our deployed action model, GEM-VLA, exhibits vastly superior task execution abilities in both simulation environments and real-world evaluations. Code, models, and datasets are available at `https://zhaorw02.github.io/GEM/`.

Keywords: Embodied Intelligence · Vision-Language Models · Vision-Language-Action Models

1 Introduction

Recent advancements in Vision-Language Models (VLMs) [2, 4, 35, 44, 66] have unlocked remarkable capabilities in embodied understanding, encompassing critical skills such as spatial recognition, physical grounding, and complex task planning. By effectively aligning visual perception with natural language reasoning, these models [23, 28, 45, 76] have emerged as robust foundation architectures for Vision-Language-Action (VLA) frameworks [5,27,32,64]. Consequently, Embodied VLMs are increasingly being leveraged to drive a massive array of downstream operational tasks, demonstrating potential for generalization and autonomous execution within dynamic, real-world physical environments.

† Project Lead. ‡ Corresponding author.

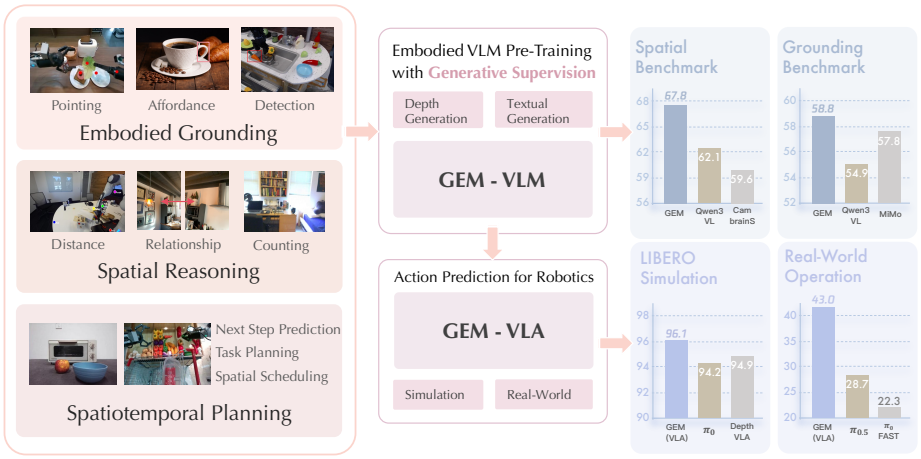

Fig. 1: Overview of GEM. GEM is a generative-supervised embodied VLM that strengthens semantic reasoning and physical grounding by combining language modeling with an auxiliary depth-generation objective (center). Trained on the high-quality, large-scale pre-training datasets spanning diverse embodied tasks (left), GEM achieves strong performance across a wide range of embodied benchmarks. Based on the architecture of GEM, the extending GEM-VLA attains state-of-the-art success rates on LIBERO and generalizes well to real-world robot manipulation (right).

Despite these foundational successes, the predominant paradigm for training embodied VLMs [1, 18, 23, 28, 76] relies heavily on scaling up massive visual question answering datasets [12, 55, 59, 76, 87]. While this approach effectively boosts performance on high-level semantic benchmarks and passive comprehension tasks [61, 65, 78, 87, 99], it inherently creates a disconnect from the physical constraints of real-world applications. Because these datasets primarily emphasize descriptive reasoning over active, physical interaction, a critical bottleneck emerges: superior semantic comprehension does not invariably translate to proficient task execution in complex, real-world environments. Conversely, alternative lines of research [36, 56, 86, 98] attempt to bridge this gap by explicitly integrating spatial, temporal, and low-level physical knowledge directly into downstream VLA models [32, 64] to enhance operational performance. However, these low-level physical priors are typically injected late in the pipeline or treated as separate entities from the broad textual pre-training data. This isolates critical physical grounding from the rich, open-vocabulary semantic guidance of linguistic models, preventing the development of a truly unified, embodied representation. Consequently, a critical and emerging question arises: how can we seamlessly embed essential spatial and physical knowledge directly into the foundational pre-training phase of vision-language models, such that it tangibly elevates both abstract semantic reasoning and actionable, real-world operational intelligence?

To overcome these limitations, we propose **GEM**, a **G**enerative-supervised **Em**bodied vision-language model. To effectively capture fine-grained structural

details and complete spatial and geometric relations within visual scenes, we establish depth map prediction [41] as an intrinsic generative target. This is achieved through a novel hybrid autoregressive-diffusion architecture [11, 70] designed to seamlessly blend generative and representational supervision. Specifically, our approach conditions a diffusion transformer [42,53] on the hidden visual features extracted by an auto-regressive understanding model [2] to synthesize accurate depth maps. To facilitate this integration, we implement a progressive training strategy that initially stabilizes the generation module before jointly optimizing for both depth synthesis and linguistic knowledge acquisition. Furthermore, to synergize with our architectural and training advancements, we introduce GEM-4M, a high-quality, large-scale embodied pre-training dataset. GEM-4M encompasses extensive embodied question-answering pairs that rigorously cover physical grounding, spatial-temporal planning, and physical reasoning tasks. Ultimately, the comprehensive spatial and semantic representations learned by the GEM architecture can be effortlessly extended into a VLA model, denoted as GEM-VLA, facilitating robust, autonomous performance in real-world robotic deployments.

Extensive experimental evaluations demonstrate that GEM and GEM-VLA show remarkable performance under a wide range of benchmarks from recognition to real-world operations. GEM establishes a new state-of-the-art, consistently outperforming leading open-source general-purpose models, as well as spatial and embodied specialists, on key reasoning benchmarks. Specifically, GEM attains the highest overall scores on the challenging spatial-related benchmarks [19, 65, 78, 81] and shows large gains over its initialization backbones. For instance, the VSI-Bench [78] score improves from 50.4 to 62.8 for the 2B model and from 57.9 to 70.6 for the 8B model. On benchmarks that require fine-grained spatial grounding [61, 87, 99], GEM far exceeds the performance of the strong proprietary baseline, Gemini-3-Pro, by 10%. Furthermore, our vision-language-action model, GEM-VLA, achieves a record-breaking 96.1% average success rate on the LIBERO [43] benchmark, outperforming standard VLAs such as π_0 and spatial-enhanced VLAs [56,86]. GEM-VLA also transfers robustly to challenging real-world settings and surpasses recent methods [26,54] with an average success rate of 43%, marking a substantial improvement over the previous state-of-the-art's 28.7%.

2 Related Work

2.1 Vision-Language Models for Embodied Intelligence

Enhancing the embodied reasoning capabilities of state-of-the-art Vision-Language Models (VLMs) has become a central research focus. A number of data-driven methodologies have emerged to support such reasoning capabilities, including object affordances for manipulation, object counting, spatial relationship understanding, and action planning that determines subsequent steps based on the current states. For instance, some studies [1, 23, 33, 49, 55, 63, 76] contribute curated datasets specifically tailored for embodied tasks, emphasizing multi-modal

understanding and action-aware visual-language alignment. Additionally, other works [18, 28, 88, 93, 99] construct synthetic spatiotemporal reasoning datasets enriched with Chain-of-Thought (CoT) annotations [68] and then incorporate Reinforcement Fine-Tuning (RFT) [60] to further refine reasoning performance of Embodied VLMs. Nevertheless, existing approaches mainly focus on high-level semantic understanding, while overlooking the explicit modeling of fine-grained structural information in visual inputs. As a result, the visual features fail to preserve fine-grained geometric cues, leading to ambiguous spatial relationships. This issue is particularly critical for embodied tasks, where precise perception of object geometry and relative distances is essential for robust manipulation and interaction. In this paper, we imitate this issue by introducing generative supervision to facilitate the fusion of structural and semantic features for more comprehensive embodied reasoning.

2.2 Spatial-Aware Vision-Language-Action Models

Robotic manipulation has evolved from single-task specialists to generalist models trained on broad, diverse datasets. Fueled by advances in VLMs [2, 4, 14, 66], and large-scale robot action datasets [6, 31, 52, 71, 72], this evolution has given rise to the architecture of Vision-Language-Action (VLA) models [5, 10, 27, 32, 39, 46, 47, 64, 69], which integrate the VLM backbone with robot action output head. Inheriting the rich perceptual and linguistic representations of pretrained VLMs, VLA models demonstrate improved adaptability and zero-shot capabilities in interpreting and executing human instructions. Despite their promising performance, current VLAs are primarily confined to 2D observation inputs and lack precise perception and comprehension of the 3D physical world. To bridge this gap, early efforts augmented VLAs with 3D or 2.5D inputs [36,38,89,97,98]. However, such approaches suffer from expensive computational and data acquisition costs. More recent works [37,56,62,73,86] instead explore various implicit enhancement strategies that implicit enhancement strategies that integrate global spatial context into the semantic representations from 2D observations, to inject geometric priors. Nevertheless, these methods mainly rely on simple feature fusion, which limits their ability to substantially improve spatial perception. Other works [9, 24, 29, 40, 50, 67, 91, 92, 94] incorporate generative world models that predict future frames or states to inject world knowledge. Although this improves planning by simulating futures, it contributes little to strengthening the geometric encoding of the current scene. Overall, enhancing VLAs with a robust and physically grounded perception of the real world remains an open and challenging problem.

3 Method

In this section, we detail our design of GEM's overall framework. We elaborate our architecture design in Sec. 3.1 and progressive training pipeline in Sec.3.2. Then we describe the construction of our training dataset GEM-4M in Sec.3.3.

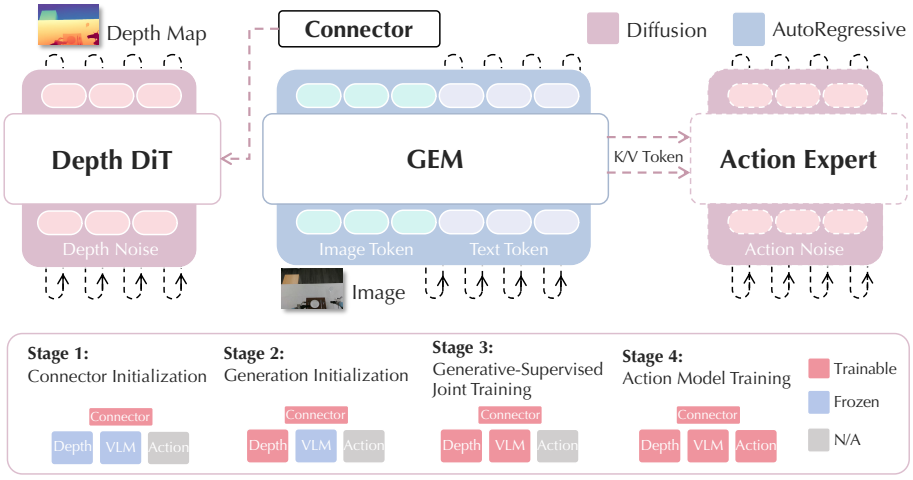

Fig. 2: Architecture of GEM. GEM augments a VLM backbone with a DiT-based depth generator conditioned on the backbone's final-layer visual tokens. We adopt a progressive training paradigm: (i) initialize the connector, (ii) warm up the depth generator, (iii) perform end-to-end joint training, and (iv) train an autoregressive action expert on GEM's multimodal tokens. Building on GEM, the GEM-based VLA predicts continuous actions from these representations, improving robot manipulation.

Finally, we explain how we extend our model to a VLA framework for downstream robot tasks in Sec. 3.4.

3.1 Architecture

In current VLMs, given an instruction l and visual input o, the VLM backbone M_θ encodes them into multimodal token representations $\mathbf{h} = (\mathbf{h}_o, \mathbf{h}_l) = M_\theta(o, l)$ at its final layer. Then they are trained to maximize the likelihood of the target token sequence y, typically using a cross-entropy objective for supervised fine-tuning:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{T} \log p_\theta(y_i | y_{<i}, \mathbf{h}_o, \mathbf{h}_l) \tag{1}$$

This objective helps the models align visual token features with text and perform semantic understanding tasks. Despite demonstrating outstanding performance in various visual tasks, their spatial reasoning ability, particularly in embodied scenarios, is limited because $\mathbf{h}_o$ contains only semantic information from o and lacks sufficient physical structural cues for accurate spatial understanding and manipulation in real-world environments.

To address this, we introduce a depth generative objective for supervision. As illustrated in Figure 2, GEM consists of a VLM backbone M_θ, a lightweight connector C_ϕ and a Diffusion Transformer (DiT)-based depth generative head

G_ψ. In our design, the visual tokens in $\mathbf{h}$, denoted as $\mathbf{h}_o$, are projected into a conditional embedding space via the connector: $\mathbf{c} = C_\phi(\mathbf{h}_o)$. We propose to utilize $\mathbf{c}$ as the condition for the generative head to reconstruct the observation o's depth map d. We then employ a flow matching objective $\mathcal{L}_{\text{flow}}$ to optimize the generative head, which learns the vector field v_t at each timestep $t \sim \mathcal{U}(0,1)$ that transforms a noised distribution $\mathbf{x}_t$ into the ground-truth depth d:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{d,t\sim\mathcal{U}(0,1),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\mathbf{v}_t(\mathbf{x}_t,\mathbf{c}) - \mathbf{u}_t(\mathbf{x}_t \mid d)\|^2\right] \tag{2}$$

where $\mathbf{u}_t(\mathbf{x}_t \mid d)$ is the ground-truth velocity field that transforms $\mathbf{x}_t$ into depth d. We combine this generative supervision loss with $\mathcal{L}_{\text{CE}}$ to allow $\mathbf{h}_o$ to encode adequate structural information for depth generation, as well as sufficient semantic information for inference.

3.2 Progressive Training Recipe

Since there is a gap between the backbone's output space and the DiT's input space, directly training the overall framework end-to-end may cause modality interference between the generative head and the VLM backbone, leading to unstable convergence. To address this, we adopt a progressive training recipe to bridge the gap between the two feature spaces effectively. Specifically, the training pipeline is divided into the following three distinct phases:

Stage 1: Connector Initialization In the first stage, we freeze both the pretrained VLM backbone and the DiT generative head, and only optimize the connector for preliminary feature alignment. The connector projects the backbone's semantic representations into the DiT's input feature space to establish a stable start for later training stages. At this stage, only the generative objective $\mathcal{L}_{\text{flow}}$ is used.

Stage 2: Generative Head Initialization After preliminary feature alignment, the generative head has not yet adapted to the conditioning features from the VLM backbone. Therefore, we freeze the backbone and only optimize both the connector and DiT head to equip the depth generative head with basic image generation ability. At this stage, the generative objective $\mathcal{L}_{\text{flow}}$ is used solely to transform high-level semantic features into fine-grained structure features, building the foundation for subsequent joint training.

Stage 3: Generative-Supervised Joint Training In the final stage, we perform end-to-end generative-supervised joint training. Since the first two stages have established a stable initialization, we unfreeze the trainable parameters of the entire framework, including VLM backbone, connector, and DiT head, to foster synergy between the backbone's semantic understanding and DiT's generative capability. This allows VLM not only to understand semantics but also to refine its representations to be more structure-aware, capturing subtle geometric

cues and spatial relationships. At this stage, both cross-entropy text loss $\mathcal{L}_{\text{CE}}$ and flow-matching generative loss $\mathcal{L}_{\text{flow}}$ supervise the training process, with the total loss defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{flow}}$, where λ is the balancing weight.

3.3 Dataset

To advance the capability of GEM in perception and reasoning real-world scenarios grounded in physical knowledge, we construct a high-quality, large-scale question-answer (QA) dataset, GEM-4M, for supervised fine-tuning. Here we present an overview of the data building engine and sources, while more details about the construction methodologies are provided in the supplementary materials.

Embodied Grounding Data To enhance the model's object recognition and localization capacities in embodied scenarios, we collect 1M high-quality question-answer pairs to support multiple grounding tasks, including open-vocabulary object detection with bounding boxes, localizing objects from instructions, and recognizing object affordances. These data are sourced from several publicly available embodied grounding datasets, such as PACO-LVIS [57], RoboPoint [87], RoboAfford [22], ShareRobot [28], and Roborefit [48]. Additionally, to ensure grounding in physical manipulation scenarios, we generate approximately 100k point and bounding box annotations from open-source robot action datasets [6, 31,52,71] using SAM3 [8]. This combination of open-source and self-curated data covers a wide range of scenarios, enhancing the diversity and generalization of visual grounding in real-world embodied environments. To handle varying image resolutions, both bounding boxes and points are normalized to the range $[0, 1000]$ to ensure consistency.

Physical, Spatial Reasoning Data This category of data aims to help the model build a foundational understanding of the physical world, such as measurement estimation and spatiotemporal reasoning. Specifically, We incorporate open-source spatial datasets, including MindCube [85], ViCA [21], SPAR [90], and VSI-590K [80], to support 3D spatial reasoning and physical attribute perception. Additionally, we also augment these datasets with 100k manually annotated spatial understanding samples from publicly available 3D scene datasets [3, 17,84], following the data processing pipeline proposed in VSI-Bench [80]. To improve spatiotemporal abilities especially in robot tasks, we integrate 1 million question-answer pairs aggregated from multiple publicly available datasets, such as RoboVQA [59], Robo2VLM [12], and RefSpatial [99]. The integration of these diverse, high-quality data sources strengthens the model's spatial awareness and boosts performance in complex embodied reasoning tasks.

Spatiotemporal Planning Data To equip the embodied brain with the ability to plan sub-tasks and forecast the trajectory of each atomic action, we collect

data from public robot datasets [6, 71, 72] with sub-task annotations and construct question-answer pairs. We extract individual frames from entire egocentric videos based on sub-task annotations and identify the manipulated object in each sub-task description using Qwen3 [75]. We then use SAM3 [8] to generate object masks and track their trajectory using CoTracker3 [30]. Finally, based on the sub-task descriptions and visualized trajectories, we create sub-task and trajectory planning question-answer pairs respectively following the RoboVQA [59] and MolmoACT [33] templates, resulting in a dataset of approximately 50K samples. The integration of these spatiotemporal data allows the model to combine basic skills, generalize to new scenarios, and plan actions effectively.

3.4 Expanding to Vision-Language Action Model

We integrate GEM into a VLA framework to evaluate its transfer to robotic manipulation. As illustrated in Figure 2, we integrate a Diffusion Transformer (DiT)-based action expert, denoted as A_ω, to generate continuous actions from multi-modal observations via a diffusion policy. We extract the key–value tokens of the multimodal observation history $\mathcal{O}$ from the attention blocks in backbone M_θ and use it as the conditioning representation $\mathbf{c}_{\text{act}}$ for the action expert A_ω, to bridge high-level reasoning capabilities and low-level action generation. We perform end-to-end joint optimization of the VLM M_θ, the depth generative head G_ψ, and the action expert A_ω using a combination of both depth and action generative objectives. Specifically, the action objective $\mathcal{L}_{\text{action}}$ aims to predict the vector field $\mathbf{v}_t$ at each timestep $t \sim \mathcal{U}(0,1)$ that transforms a noisy action state $\mathbf{a}_t = (1-t)\epsilon + t\mathbf{a}$ into the ground-truth action chunk $\mathbf{a}$:

$$\mathcal{L}_{\text{action}} = \mathbb{E}_{\mathcal{O},\mathbf{a},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t\sim\mathcal{U}(0,1)}\left[\|\mathbf{v}_t(\mathbf{a}_t,\mathbf{c}_{\text{act}}) - \mathbf{u}_t(\mathbf{a}_t \mid \mathbf{a})\|_2^2\right] \tag{3}$$

The total loss is then defined as: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{action}} + \lambda\mathcal{L}_{\text{flow}}$, where λ is the same balancing weight.

4 Experiments

4.1 Implementation Details

We adopt Qwen3-VL [2] as our VLM backbone and Sana [74] as the depth prediction head. We define a light connector comprising 2 layers of MLP that bridge the backbone's output space with the DiT's input space. Since some of our training data lack ground-truth depth annotations, we use DepthAnythingv3 [41] to generate pseudo depth maps for supervision. We train for 500 steps in Stage 1, 4k steps in Stage 2, and 1 epoch in Stage 3. We set $\lambda = 0.1$ to balance structural synthesis with semantic understanding. The training process is performed on 32 NVIDIA A800 GPUs, with a cosine learning rate scheduler from $1e-5$ to $1e-6$. In real-world VLA tasks, we adopt the dedicated action expert from RDT2 [47]. For each specific task, we jointly fine-tune the entire framework for 50k steps on 8 NVIDIA A800 GPUs, with a linear scheduler and a learning rate of $1e-5$. The balancing weight λ in VLA finetuning is also set to 0.1. More implementation details can be seen in the supplementary material.

Table 1: Performance on embodied reasoning benchmarks for spatial understanding across different model types. The highest and second-highest accuracy values are highlighted in **bold** and underlined, respectively. GEM-8B achieves state-of-the-art (SOTA) performance and near-SOTA competitive results across general-purpose and spatial specialist models.

Models	CV-Bench	VSI-Bench			MMSI-Bench	EmbSpatial
	All ↑	Abs. Dist.	Rel. Dist.	All ↑	All ↑	All ↑
Gemini-3-Pro [14]	82.5	42.8	56.6	53.0	45.9	81.0
Seed1.8 [58]	86.5	28.0	50.3	47.2	34.6	78.7
GPT-4o [25]	78.6	5.3	37.0	34.0	30.3	71.9
Qwen3-VL-2B [2]	80.0	40.7	49.6	50.4	23.6	69.0
Qwen3-VL-8B [2]	85.1	47.5	58.2	57.9	27.7	77.7
InternVL3.5-8B [66]	81.5	40.9	47.7	54.1	28.4	74.2
LLava-OneVision-7B [35]	61.9	20.2	42.5	32.4	26.6	73.1
SpaceR-7B [51]	74.8	28.6	38.2	35.8	26.4	65.8
VLM-3R [20]	71.8	49.4	65.4	60.7	27.9	68.2
VST-7B [79]	83.5	43.8	60.0	60.6	32.6	73.6
CambrainS-7B [80]	76.9	49.4	66.9	67.5	24.2	70.0
SenseNova-SI-8B [7]	83.2	48.0	64.1	67.9	**43.3**	77.6
Qwen3-VL-2B-SFT	80.7	45.1	62.2	60.0	28.9	72.7
GEM-2B (Ours)	81.4	48.4	64.1	62.8	30.6	73.0
Qwen3-VL-8B-SFT	85.6	53.7	71.4	68.6	32.8	78.3
GEM-8B (Ours)	**86.6**	**56.3**	**72.3**	**70.6**	35.3	**79.4**

4.2 Evaluation on Embodied Reasoning Capacities

We evaluate on public spatiotemporal embodied reasoning benchmarks, including CV-Bench [65], VSI-Bench [78], MMSI-Bench [81] and EmbSpatial [19]. As shown in Table 1, both scales of GEM yield significant improvements across all tasks compared to their base models, Qwen3-VL. For instance, on the challenging VSI-Bench and MMSI-Bench, GEM improves the scores by roughly 10%. In particular, compared with open-source general-purpose baselines [2, 35, 66], the 8B-scale variant of GEM achieves the strongest overall performance on the majority of benchmarks, and remains highly competitive on the remaining. Furthermore, it also achieves superior performance even when compared to spatial specialist models [7,20,51,79,80]. These strong results highlight our model's powerful spatiotemporal reasoning capabilities, which can be effectively transferred to downstream tasks.

We further evaluate our model on benchmarks, including RefSpatial [99], Where2Place [87], and RoboSpatial [61], which focus on object placement and referring in embodied environments. The results summarized in Table 2 show that the 8B variant of GEM achieves the best overall performance compared with both general-purpose and embodied specialist models [1, 23, 28, 49, 77], while the 2B variant also remains highly competitive with many larger-scale models [1, 49, 77]. It is also worth noting that GEM exceeds the strong proprietary baseline, Gemini-3-Pro [63], by about 10% on average across benchmarks.

Table 2: Performance on the object placement and grounding spatial benchmarks across different model types. The highest and second-highest accuracy values are highlighted in **bold** and underlined. * denotes results obtained from their reports. It shows that GEM-8B achieves the best performance, compared with general-purpose and embodied specialist models.

Models	RefSpatial			Where2Place			RoboSpatial
	Loc.	Pla.	All ↑	Seen	Unseen	All ↑	All ↑
Gemini-3-Pro [14]	30.0	35.0	34.3	58.6	43.3	54.0	57.4
Seed1.8 [58]	65.0	41.0	50.2	54.2	53.3	53.8	66.9
GPT-4o [25]	8.00	9.55	8.78	20.3	20.7	20.4	43.5
Qwen3-VL-2B [2]	36.0	29.0	27.4	45.7	43.3	45.0	40.7
Qwen3-VL-8B [2]	54.0	36.7	38.0	61.0	62.2	61.3	65.4
VeBrain-8B [49]	0.03	0.57	0.30	12.3	9.17	11.3	42.5
Magma-8B [77]	1.00	8.00	4.50	9.93	13.1	10.9	33.7
RoboBrain-2.0-7B* [28]	36.0	29.0	32.5	64.3	61.9	63.6	54.2
Mimo-Embodied-7B* [23]	-	-	**48.0**	-	-	63.6	61.8
Cosmos-Reason2-8B [1]	48.0	27.0	33.2	52.9	43.3	50.0	65.7
Qwen3-VL-2B-SFT	37.0	29.0	29.6	56.7	51.4	53.0	44.6
GEM-2B (Ours)	41.0	33.0	32.1	52.8	53.3	53.0	47.4
Qwen3-VL-8B-SFT	53.0	**40.0**	45.8	60.0	62.9	62.0	65.4
GEM-8B (Ours)	**57.0**	38.0	44.4	**65.7**	**63.3**	**65.0**	**66.9**

Such competence highlights GEM's outstanding spatial reasoning capabilities in embodied environments, making it a versatile backbone for embodied AI brains.

Moreover, to assess the impact of depth generative supervision, we exclude this component and fine-tune the base models on the constructed dataset, referred to as Qwen3VL-2B-SFT and Qwen3VL-8B-SFT. As shown in Table 1 and Table 2, the performance of both scales drops compared to the full models. This is because the full model leverages generative supervision to integrate global structural features with semantic features within a shared representation space, which benefits stronger spatial perception capabilities. Notably, on distance-related questions in VSI-Bench [78], the full models significantly outperform their counterparts, demonstrating that depth supervision enhances the model's ability to perceive relative distance and spatial relationships.

4.3 Evaluation on Downstream VLA tasks

Simulation Evaluation We perform evaluation on the widely adopted LIBERO benchmark [43]. It comprises four task suites: Spatial, Object, Goal, and Long, each containing 10 diverse tasks with 50 trials per task. For comprehensive comparison, we select high-performance VLA models, including both standard models [13, 27, 32, 64] and spatially-enhanced VLA models [34, 56, 86, 92, 98] as baselines. Unlike prior models pre-trained on robot data, we fine-tune GEM-2B and its action expert from scratch for action prediction on LIBERO for 20k steps, following the StarVLA implementation [15]. The evaluation results, shown in Table 3, demonstrate that GEM-VLA achieves the highest success rate across all

Table 3: Success rates on the LIBERO benchmark across four task suites. It is demonstrated that GEM-VLA exhibits better performance than all baselines, suggesting that implicit geometry reasoning from depth generative supervision improves generalization across diverse manipulation tasks.

Models	Spatial	Object	Goal	Long	Average ↑
Diffusion Policy [13]	78.3	92.5	68.3	50.5	72.4
Octo-Base [64]	78.9	85.7	84.6	51.1	75.1
OpenVLA [32]	84.7	88.4	79.2	53.7	76.5
π_0 (reported) [27]	96.8	98.8	95.8	85.2	94.2
TraceVLA [98]	84.6	85.2	75.1	54.1	74.8
SpatialVLA [56]	88.2	89.9	78.6	55.5	78.1
MolmoACT [34]	87.0	95.4	87.6	77.2	86.6
DreamVLA [92]	97.5	94.0	89.5	59.5	92.6
DepthVLA [86]	96.4	98.0	95.8	89.2	94.9
Qwen3VL-SFT-VLA	97.2	98.4	95.6	88.4	94.9
GEM-VLA (Ours)	**99.0**	**98.8**	**97.1**	**89.3**	**96.1**

task types. This indicates that despite large-scale action pretraining, standard VLAs still lack sufficient spatial grounding for precise manipulation. Furthermore, GEM-VLA also outperforms other spatially enhanced VLAs, showcasing the strongest physical grounding capacities. Additionally, the performance of the VLA based on GEM also surpasses that based on the standard SFT model, which further underscores the benefit of depth generative supervision for accurate action prediction.

Real-World Evaluation We extend GEM to GEM-VLA following the approach in Sec. 3.4, and deploy it on a UR5 platform to evaluate its performance on real-world manipulation tasks. We compare our model with most advanced baselines π_0-FAST [54] and $\pi_{0.5}$ [26]. We finetune each model on several challenging real-world tasks, including long-horizon tasks (e.g., table bussing) and deformable object manipulation (e.g., folding clothes, unzipping a zipper). Task descriptions are shown in Figure 4.

As summarized in Figure 3, our GEM-VLA demonstrates superior performance across all task and subtask categories. In both deformable manipulation tasks, GEM achieves higher success rates than all baselines [26, 54], particularly on the complex multi-step cloth folding task. For the long-horizon table bussing task, GEM-VLA achieves a significantly higher average progress score, which indicates its stronger long-horizon robustness and stability. These results confirm that GEM effectively transfers its pre-trained physical knowledge to state-of-the-art performance in diverse challenging real-world robotic manipulation.

Moreover, we also investigate the effectiveness of depth generative supervision during VLA fine-tuning. We freeze the depth head and train only the VLM backbone and action expert with the action objective in Eq. 3. As shown in the Figure 3, performance drops on almost all tasks when depth generative supervision is removed, indicating that auxiliary depth prediction enables more accurate

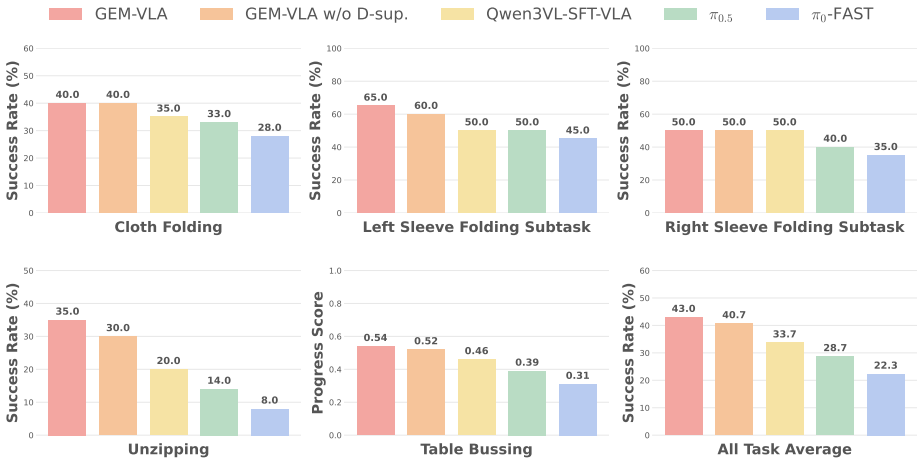

Fig. 3: Comparison of GEM and baseline models on challenging real-world tasks. The progress score refers to the average percentage of sub-tasks completed in the long-horizon task. It is noted that the full GEM-VLA model outperforms previous baselines in success rate and progress score across all tasks and sub-tasks.

manipulation. Notably, the VLA fine-tuned from pre-trained GEM still outperforms its counterpart finetuned on the standard SFT model, which suggests that depth generative supervision facilitates learning physical priors in the embodied VLM, translating into stronger performance on downstream robot tasks.

5 Ablation Studies

5.1 Superiority of Depth Supervision over RGB

We investigate why depth is a more suitable supervisory signal compared to other alternatives. Specifically, we compare our depth-based generation with an RGB-based image generation task, where the model is trained to regenerate the input image. For fair comparison, we fine-tune both models on the open-source VSI-590K dataset [80], keeping all other training hyper-parameters and strategies default. We evaluate each model on the representative spatial reasoning benchmarks, including CV-Bench [65], VSI-Bench [78], and RoboSpatial [61].

The results, summarized in Table 4 (rows 1 and 3), reveal that replacing depth supervision with RGB reconstruction leads to inferior performance, particularly on distance-related questions in VSI-Bench. This suggests depth provides more explicit cues about spatial relationships, such as relative distance, making it a more effective supervisory signal than RGB.

5.2 Effectiveness of Progressive Training Strategy

We evaluate the effectiveness of the proposed progressive three-stage training strategy. To compare, we perform a direct end-to-end training on both the depth

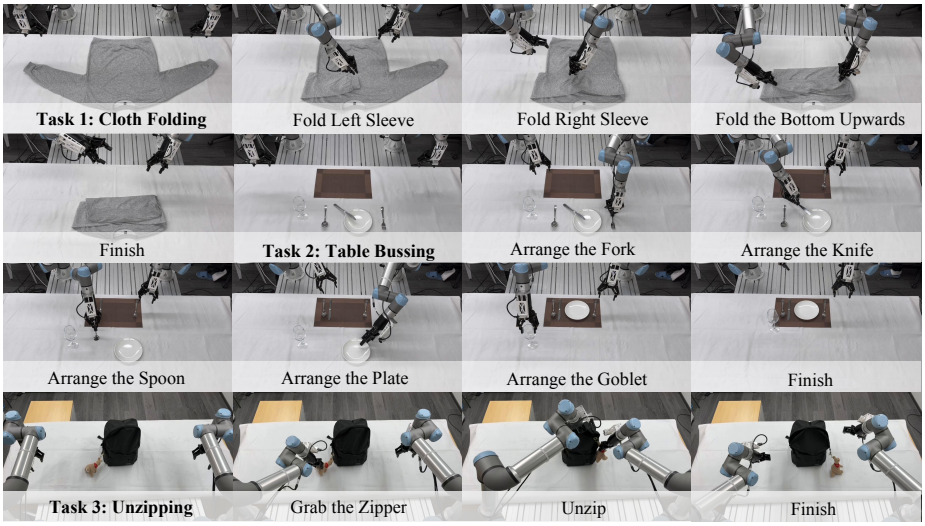

Fig. 4: Demonstrations of three finetuned real-robot tasks, including one long-horizon task **table bussing** and two deformable object manipulation tasks **folding clothes** and **unzipping a backpack's zipper**.

Table 4: Comparison between GEM and different settings. The results show that the default model yields the best performance, which indicates that depth supervision enhances structural learning compared to RGB and direct end-to-end training fails to effectively integrate semantic and structural features.

Models	CV-Bench	VSI-Bench			RoboSpatial
	All ↑	Abs.Dist.	Rel. Dist.	All ↑	All ↑
RGB Supervision	80.9	47.5	62.8	60.0	44.6
Direct End-to-End Co-Training	79.7	42.1	60.0	57.6	44.0
Default Setting (GEM)	**81.1**	**47.8**	**65.2**	**63.0**	**48.9**

generative head and understanding backbone using the VSI-590K dataset [80], and evaluate its performance on the same benchmarks [61, 65, 78] above. The comparison results are shown in Table 4, in the second and third rows. It can be observed that direct end-to-end training underperforms the default three-stage training paradigm. This is because the generative head and connector fail to receive an appropriate initialization, which limits the effective fusion of semantic and structural features. Therefore, direct end-to-end co-training negatively affects the understanding model's performance.

5.3 Effect of Generative Supervision on Structural Priors Learning

To investigate whether standard SFT visual representations contain sufficient geometric information for embodied reasoning, we feed the final-layer visual

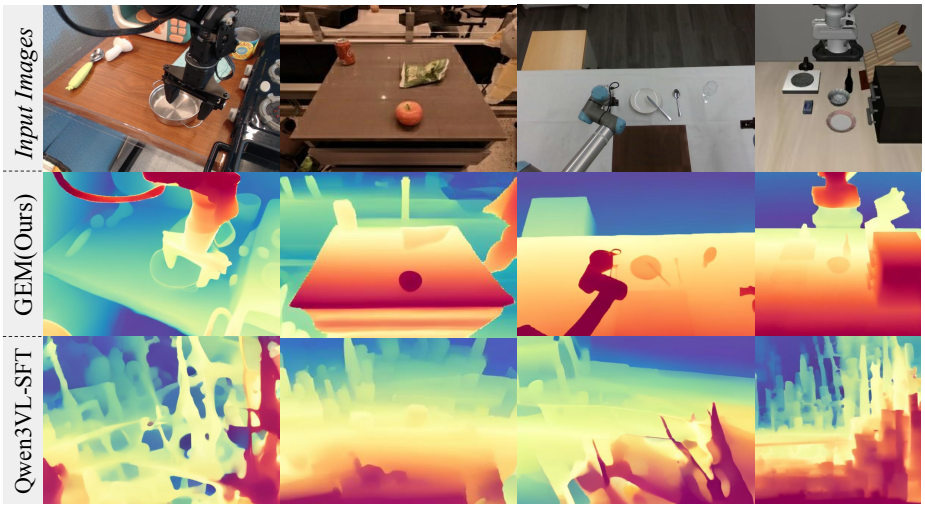

Fig. 5: Qualitative visualization of generation results from visual features in GEM and Qwen3-VL-SFT. Results from standard SFT features exhibit limited structural details, suggesting insufficient geometric cues in semantic-dominated representations. GEM features show clearer structural patterns, indicating that depth generative supervision helps capture spatial cues beneficial for embodied perception.

token features from Qwen3-VL-SFT into GEM's depth generator and use the reconstructed depth maps as a qualitative visualization of the encoded geometric cues. As shown in Figure 5, the depth maps generated from Qwen3-VL-SFT features exhibit limited structural details, suggesting that standard SFT representations are mainly dominated by high-level semantic signals and may lack explicit low-level geometric information. Such feature-level misalignment motivates the design of our depth generation module, which introduces depth generative supervision to encourage geometry-aware visual representations. In contrast, GEM produces depth maps with clearer structural patterns, indicating that the proposed supervision helps preserve spatial and geometric information from 2D inputs.

6 Conclusion

In this paper, we introduce GEM, a novel Generative-supervised Embodied vision-language framework designed to bridge the gap between high-level semantic reasoning and low-level physical grounding by learning depth generation as an intrinsic target for scene geometry. A progressive training recipe is adopted to optimize depth synthesis and language objectives to better fuse structural and semantic representations. Beyond architectural and training advancements, we also construct a large-scale dataset that covers various embodied tasks to support training. Extensive evaluations demonstrate that GEM achieves strong

performance across a variety of embodied benchmarks. Furthermore, the VLA model built on GEM sets new records on simulation benchmarks and shows robust generalization in real-world robotic tasks.

References

1. Azzolini, A., Bai, J., Brandon, H., Cao, J., Chattopadhyay, P., Chen, H., Chu, J., Cui, Y., Diamond, J., Ding, Y., et al.: Cosmos-reason1: From physical common sense to embodied reasoning. arXiv preprint arXiv:2503.15558 (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., et al.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. arXiv preprint arXiv:2111.08897 (2021)
4. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al.: Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726 (2024)
5. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022)
6. Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., He, X., Huang, X., et al.: Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE (2025)
7. Cai, Z., Wang, R., Gu, C., Pu, F., Xu, J., Wang, Y., Yin, W., Yang, Z., Wei, C., Sun, Q., Zhou, T., Li, J., Pang, H.E., Qian, O., Wei, Y., Lin, Z., Shi, X., Deng, K., Han, X., Chen, Z., Fan, X., Deng, H., Lu, L., Pan, L., Li, B., Liu, Z., Wang, Q., Lin, D., Yang, L.: Scaling spatial intelligence with multimodal foundation models. arXiv preprint arXiv:2511.13719 (2025)
8. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2025), `https://arxiv.org/abs/2511.16719`
9. Cen, J., Yu, C., Yuan, H., Jiang, Y., Huang, S., Guo, J., Li, X., Song, Y., Luo, H., Wang, F., et al.: Worldvla: Towards autoregressive action world model. arXiv preprint arXiv:2506.21539 (2025)
10. Cheang, C., Chen, S., Cui, Z., Hu, Y., Huang, L., Kong, T., Li, H., Li, Y., Liu, Y., Ma, X., et al.: Gr-3 technical report. arXiv preprint arXiv:2507.15493 (2025)
11. Chen, J., Xue, L., Xu, Z., Pan, X., Yang, S., Qin, C., Yan, A., Zhou, H., Chen, Z., Huang, L., et al.: Blip3o-next: Next frontier of native image generation. arXiv preprint arXiv:2510.15857 (2025)
12. Chen, K., Xie, S., Ma, Z., Sanketi, P.R., Goldberg, K.: Robo2vlm: Visual question answering from large-scale in-the-wild robot manipulation datasets. arXiv preprint arXiv:2505.15517 (2025)

13. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research **44**(10-11), 1684–1704 (2025)
14. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 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)
15. starVLA Contributors: Starvla: A lego-like codebase for vision-language-action model developing. GitHub repository (1 2025). `https://doi.org/10.5281/zenodo.18264214`, `https://github.com/starVLA/starVLA`
16. Cui, R., Zhang, Z., Pang, J., Chi, H., Guo, J., Zhang, S., Xie, S., Jin, X., Mu, Y., Yang, J., Yao, G., Zhan, X., Zhang, Y.Q., Zhao, H.: Libero-safety: A comprehensive benchmark for physical and semantic safety in vision-language-action models (2026), `https://arxiv.org/abs/2606.23686`
17. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
18. Dang, R., Guo, J., Hou, B., Leng, S., Li, K., Li, X., Liu, J., Mao, Y., Wang, Z., Yuan, Y., et al.: Rynnbrain: Open embodied foundation models. arXiv preprint arXiv:2602.14979 (2026)
19. Du, M., Wu, B., Li, Z., Huang, X.J., Wei, Z.: 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). pp. 346–355 (2024)
20. Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Qu, H., Wang, D., Yan, Z., et al.: Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. arXiv preprint arXiv:2505.20279 (2025)
21. Feng, Q.: Visuospatial cognitive assistant. arXiv preprint arXiv:2505.12312 (2025)
22. Hao, X., Tang, Y., Zhang, L., Ma, Y., Diao, Y., Jia, Z., Ding, W., Ye, H., Chen, L.: Roboafford++: A generative ai-enhanced dataset for multimodal affordance learning in robotic manipulation and navigation. arXiv preprint arXiv:2511.12436 (2025)
23. Hao, X., Zhou, L., Huang, Z., Hou, Z., Tang, Y., Zhang, L., Li, G., Lu, Z., Ren, S., Meng, X., et al.: Mimo-embodied: X-embodied foundation model technical report. arXiv preprint arXiv:2511.16518 (2025)
24. Hu, Y., Guo, Y., Wang, P., Chen, X., Wang, Y.J., Zhang, J., Sreenath, K., Lu, C., Chen, J.: Video prediction policy: A generalist robot policy with predictive visual representations. arXiv preprint arXiv:2412.14803 (2024)
25. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
26. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: $pi_{0.5}$: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054 (2025)
27. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: π_0: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054 (2025)
28. Ji, Y., Tan, H., Shi, J., Hao, X., Zhang, Y., Zhang, H., Wang, P., Zhao, M., Mu, Y., An, P., et al.: Robobrain: A unified brain model for robotic manipulation from abstract to concrete. arXiv preprint arXiv:2502.21257 (2025)

29. Jiang, Y., Huang, S., Xue, S., Zhao, Y., Cen, J., Leng, S., Li, K., Guo, J., Wang, K., Chen, M., et al.: Rynnvla-001: Using human demonstrations to improve robot manipulation. arXiv preprint arXiv:2509.15212 (2025)
30. Karaev, N., Makarov, Y., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6013–6022 (2025)
31. Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M.K., Chen, L.Y., Ellis, K., et al.: Droid: A large-scale in-the-wild robot manipulation dataset. arXiv preprint arXiv:2403.12945 (2024)
32. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
33. Lee, J., Duan, J., Fang, H., Deng, Y., Liu, S., Li, B., Fang, B., Zhang, J., Wang, Y.R., Lee, S., Han, W., Pumacay, W., Wu, A., Hendrix, R., Farley, K., VanderBilt, E., Farhadi, A., Fox, D., Krishna, R.: Molmoact: Action reasoning models that can reason in space (2025), https://arxiv.org/abs/2508.07917
34. Lee, J., Duan, J., Fang, H., Deng, Y., Liu, S., Li, B., Fang, B., Zhang, J., Wang, Y.R., Lee, S., et al.: Molmoact: Action reasoning models that can reason in space. arXiv preprint arXiv:2508.07917 (2025)
35. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
36. Li, C., Wen, J., Peng, Y., Peng, Y., Zhu, Y.: Pointvla: Injecting the 3d world into vision-language-action models. IEEE Robotics and Automation Letters **11**(3), 2506–2513 (2026)
37. Li, F., Song, W., Zhao, H., Wang, J., Ding, P., Wang, D., Zeng, L., Li, H.: Spatial forcing: Implicit spatial representation alignment for vision-language-action model. arXiv preprint arXiv:2510.12276 (2025)
38. Li, X., Heng, L., Liu, J., Shen, Y., Gu, C., Liu, Z., Chen, H., Han, N., Zhang, R., Tang, H., et al.: 3ds-vla: A 3d spatial-aware vision language action model for robust multi-task manipulation. In: 9th Annual Conference on Robot Learning (2025)
39. Li, X., Liu, M., Zhang, H., Yu, C., Xu, J., Wu, H., Cheang, C., Jing, Y., Zhang, W., Liu, H., et al.: Vision-language foundation models as effective robot imitators. arXiv preprint arXiv:2311.01378 (2023)
40. Liao, Y., Zhou, P., Huang, S., Yang, D., Chen, S., Jiang, Y., Hu, Y., Cai, J., Liu, S., Luo, J., et al.: Genie envisioner: A unified world foundation platform for robotic manipulation. arXiv preprint arXiv:2508.05635 (2025)
41. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
42. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
43. Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: Libero: Benchmarking knowledge transfer for lifelong robot learning. Advances in Neural Information Processing Systems **36**, 44776–44791 (2023)
44. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
45. Liu, H., Li, X., Li, P., Liu, M., Wang, D., Liu, J., Kang, B., Ma, X., Kong, T., Zhang, H.: Towards generalist robot policies: What matters in building vision-language-action models (2025)

46. Liu, J., Chen, H., An, P., Liu, Z., Zhang, R., Gu, C., Li, X., Guo, Z., Chen, S., Liu, M., et al.: Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. arXiv preprint arXiv:2503.10631 (2025)
47. Liu, S., Li, B., Ma, K., Wu, L., Tan, H., Ouyang, X., Su, H., Zhu, J.: Rdt2: Exploring the scaling limit of umi data towards zero-shot cross-embodiment generalization. arXiv preprint arXiv:2602.03310 (2026)
48. Lu, Y., Fan, Y., Deng, B., Liu, F., Li, Y., Wang, S.: Vl-grasp: a 6-dof interactive grasp policy for language-oriented objects in cluttered indoor scenes. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 976–983. IEEE (2023)
49. Luo, G., Yang, G., Gong, Z., Chen, G., Duan, H., Cui, E., Tong, R., Hou, Z., Zhang, T., Chen, Z., et al.: Visual embodied brain: Let multimodal large language models see, think, and control in spaces. arXiv preprint arXiv:2506.00123 (2025)
50. Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., Song, H., Chen, Q., Deng, X., Pang, J.: F1: A vision-language-action model bridging understanding and generation to actions. arXiv preprint arXiv:2509.06951 (2025)
51. Ouyang, K., Liu, Y., Wu, H., Liu, Y., Zhou, H., Zhou, J., Meng, F., Sun, X.: Spacer: Reinforcing mllms in video spatial reasoning. arXiv preprint arXiv:2504.01805 (2025)
52. O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 6892–6903. IEEE (2024)
53. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
54. Pertsch, K., Stachowicz, K., Ichter, B., Driess, D., Nair, S., Vuong, Q., Mees, O., Finn, C., Levine, S.: Fast: Efficient action tokenization for vision-language-action models. arXiv preprint arXiv:2501.09747 (2025)
55. Qu, D., Song, H., Chen, Q., Chen, Z., Gao, X., Ye, X., Lv, Q., Shi, M., Ren, G., Ruan, C., et al.: Eo-1: Interleaved vision-text-action pretraining for general robot control. arXiv preprint arXiv:2508.21112 (2025)
56. Qu, D., Song, H., Chen, Q., Yao, Y., Ye, X., Ding, Y., Wang, Z., Gu, J., Zhao, B., Wang, D., et al.: Spatialvla: Exploring spatial representations for visual-language-action model. arXiv preprint arXiv:2501.15830 (2025)
57. Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., et al.: Paco: Parts and attributes of common objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7141–7151 (2023)
58. Seed, B.: Seed1. 8 model card: Towards generalized real-world agency. Tech. rep., 2025a. Technical Report
59. Sermanet, P., Ding, T., Zhao, J., Xia, F., Dwibedi, D., Gopalakrishnan, K., Chan, C., Dulac-Arnold, G., Maddineni, S., Joshi, N.J., et al.: Robovqa: Multimodal long-horizon reasoning for robotics. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 645–652. IEEE (2024)
60. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
61. Song, C.H., Blukis, V., Tremblay, J., Tyree, S., Su, Y., Birchfield, S.: Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics.

In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15768–15780 (2025)
62. Song, W., Zhou, Z., Zhao, H., Chen, J., Ding, P., Yan, H., Huang, Y., Tang, F., Wang, D., Li, H.: Reconvla: Reconstructive vision-language-action model as effective robot perceiver. arXiv preprint arXiv:2508.10333 (2025)
63. Team, G.R., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., et al.: Gemini robotics: Bringing ai into the physical world. arXiv preprint arXiv:2503.20020 (2025)
64. Team, O.M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al.: Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213 (2024)
65. Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024)
66. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
67. Wang, Y., Li, X., Wang, W., Zhang, J., Li, Y., Chen, Y., Wang, X., Zhang, Z.: Unified vision-language-action model. arXiv preprint arXiv:2506.19850 (2025)
68. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
69. Wen, J., Zhu, Y., Li, J., Zhu, M., Tang, Z., Wu, K., Xu, Z., Liu, N., Cheng, R., Shen, C., et al.: Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. IEEE Robotics and Automation Letters (2025)
70. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., Liu, Z., Xia, Z., Li, C., Deng, H., Wang, J., Luo, K., Zhang, B., Lian, D., Wang, X., Wang, Z., Huang, T., Liu, Z.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
71. Wu, K., Hou, C., Liu, J., Che, Z., Ju, X., Yang, Z., Li, M., Zhao, Y., Xu, Z., Yang, G., et al.: Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. arXiv preprint arXiv:2412.13877 (2024)
72. Wu, S., Liu, X., Xie, S., Wang, P., Li, X., Yang, B., Li, Z., Zhu, K., Wu, H., Liu, Y., et al.: Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. arXiv preprint arXiv:2511.17441 (2025)
73. Wu, W., Lu, F., Wang, Y., Yang, S., Liu, S., Wang, F., Ma, S., Sun, H., Wang, Y., Qiu, Z., Xiong, H., Wang, Z., Zhou, S., Ren, Y., Zhang, K., Yu, H., Zhao, J., Zhu, Q., Cheng, R., Li, Y.L., Huang, Y., Zhu, X., Shen, Y., Zheng, K.: A pragmatic vla foundation model. arXiv preprint arXiv:2601.18692v1 (2026)
74. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., et al.: Sana: Efficient high-resolution image synthesis with linear diffusion transformers. arXiv preprint arXiv:2410.10629 (2024)
75. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
76. Yang, G., Zhang, T., Hao, H., Wang, W., Liu, Y., Wang, D., Chen, G., Cai, Z., Chen, J., Su, W., et al.: Vlaser: Vision-language-action model with synergistic embodied reasoning. arXiv preprint arXiv:2510.11027 (2025)
77. Yang, J., Tan, R., Wu, Q., Zheng, R., Peng, B., Liang, Y., Gu, Y., Cai, M., Ye, S., Jang, J., et al.: Magma: A foundation model for multimodal ai agents. In:

Proceedings of the computer vision and pattern recognition conference. pp. 14203–14214 (2025)
78. Yang, J., Yang, S., Gupta, A., Han, R., Fei-Fei, L., Xie, S.: Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. arXiv preprint arXiv:2412.14171 (2024)
79. Yang, R., Zhu, Z., Li, Y., Huang, J., Yan, S., Zhou, S., Liu, Z., Li, X., Li, S., Wang, W., et al.: Visual spatial tuning. arXiv preprint arXiv:2511.05491 (2025)
80. Yang, S., Yang, J., Huang, P., Brown, E., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., Lu, D., Fergus, R., LeCun, Y., Fei-Fei, L., Xie, S.: Cambrian-s: Towards spatial supersensing in video. arXiv preprint arXiv:2511.04670 (2025)
81. Yang, S., Xu, R., Xie, Y., Yang, S., Li, M., Lin, J., Zhu, C., Chen, X., Duan, H., Yue, X., et al.: Mmsi-bench: A benchmark for multi-image spatial intelligence. arXiv preprint arXiv:2505.23764 (2025)
82. Ye, J., Huang, Z., Qu, Y., Wang, C., Yang, Y., Li, Y., Luo, Y., Chen, Z., Lu, S., Zhu, J., et al.: Universe3d: Emerging properties of unified multimodal models in 3d understanding and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 613–623 (2026)
83. Ye, J., Wang, Z., Zhao, R., Xie, S., Zhu, J.: Shapellm-omni: A native multimodal llm for 3d generation and understanding. arXiv preprint arXiv:2506.01853 (2025)
84. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
85. Yin, B., Wang, Q., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., et al.: Spatial mental modeling from limited views. In: Structural Priors for Vision Workshop at ICCV'25 (2025)
86. Yuan, T., Liu, Y., Lu, C., Chen, Z., Jiang, T., Zhao, H.: Depthvla: Enhancing vision-language-action models with depth-aware spatial reasoning. arXiv preprint arXiv:2510.13375 (2025)
87. Yuan, W., Duan, J., Blukis, V., Pumacay, W., Krishna, R., Murali, A., Mousavian, A., Fox, D.: Robopoint: A vision-language model for spatial affordance prediction for robotics. arXiv preprint arXiv:2406.10721 (2024)
88. Yuan, Y., Cui, H., Huang, Y., Chen, Y., Ni, F., Dong, Z., Li, P., Zheng, Y., Hao, J.: Embodied-r1: Reinforced embodied reasoning for general robotic manipulation. arXiv preprint arXiv:2508.13998 (2025)
89. Ze, Y., Zhang, G., Zhang, K., Hu, C., Wang, M., Xu, H.: 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. arXiv preprint arXiv:2403.03954 (2024)
90. Zhang, J., Chen, Y., Zhou, Y., Xu, Y., Huang, Z., Mei, J., Chen, J., Yuan, Y.J., Cai, X., Huang, G., et al.: From flatland to space: Teaching vision-language models to perceive and reason in 3d. arXiv preprint arXiv:2503.22976 (2025)
91. Zhang, J., Guo, Y., Hu, Y., Chen, X., Zhu, X., Chen, J.: Up-vla: A unified understanding and prediction model for embodied agent. arXiv preprint arXiv:2501.18867 (2025)
92. Zhang, W., Liu, H., Qi, Z., Wang, Y., Yu, X., Zhang, J., Dong, R., He, J., Wang, H., Zhang, Z., Yi, L., Zeng, W., Jin, X.: Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. CoRR **abs/2507.04447** (2025). https://doi.org/10.48550/ARXIV.2507.04447, https://doi.org/10.48550/arXiv.2507.04447
93. Zhang, Y., Liu, C., Ren, X., Ni, H., Zhang, S., Ding, Z., Hu, J., Shan, H., Niu, Z., Liu, Z., et al.: Pelican-vl 1.0: A foundation brain model for embodied intelligence. arXiv preprint arXiv:2511.00108 (2025)

94. Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1702–1713 (2025)
95. Zhao, R., Wang, Z., Wang, Y., Zhou, Z., Zhu, J.: Flexidreamer: Single image-to-3d generation with flexicubes. arXiv preprint arXiv:2404.00987 (2024)
96. Zhao, R., Ye, J., Wang, Z., Liu, G., Chen, Y., Wang, Y., Zhu, J.: Deepmesh: Auto-regressive artist-mesh creation with reinforcement learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10612–10623 (2025)
97. Zhen, H., Qiu, X., Chen, P., Yang, J., Yan, X., Du, Y., Hong, Y., Gan, C.: 3d-vla: A 3d vision-language-action generative world model. arXiv preprint arXiv:2403.09631 (2024)
98. Zheng, R., Liang, Y., Huang, S., Gao, J., Daumé III, H., Kolobov, A., Huang, F., Yang, J.: Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. arXiv preprint arXiv:2412.10345 (2024)
99. Zhou, E., An, J., Chi, C., Han, Y., Rong, S., Zhang, C., Wang, P., Wang, Z., Huang, T., Sheng, L., et al.: Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. arXiv preprint arXiv:2506.04308 (2025)