



SpaceTools: Tool-Augmented Spatial Reasoning via Double Interactive RL


Anonymous CVPR submission

Paper ID

 **Question:** which pedal is the smallest and where would you activate it?

One shot answer ❌


 The smallest one in your setup is the mini tuner on the far right. You'd activate it using the footswitch at its bottom.




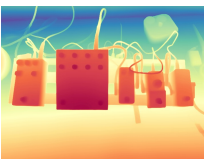
Thinking model answer ❌


Analyze the image:

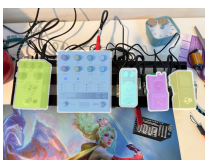
- 1. Identify pedals:** The setup includes a large "Chroma Console" pedal,
- 2. Compare sizes:** The "Lekato AT - 07 Tuner" is noticeably smaller than all other pedals in the image.
- 3. Activation location:** This tuner pedal has a footswitch at its bottom, so it is activated by pressing that switch with the foot.




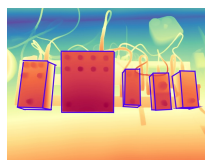
 Use <depth estimator> to understand spatial relationships and see which pedals are closer to the camera.





 Next, I'll apply the <segmentation tool> to mask guitar pedals, isolating each pedal from the other equipment



 Then, for each identified pedal, I'll use <3D bbox> to create a bounding box around it and estimate their volume.



 Now I'll use a <pointing tool> to indicate its switch, showing exactly where to press to activate the pedal.



Tool calling answer ✅

Figure 1. SpaceTools uses multiple computer vision tools to solve complex problems. Shown here is a motivating example.

Abstract

001 Vision Language Models (VLMs) demonstrate strong qual-
 002 itative visual understanding, but struggle with metrically
 003 precise spatial reasoning required for embodied applica-
 004 tions. The agentic paradigm promises that VLMs can use a
 005 wide variety of tools that could augment these capabilities,
 006 such as depth estimators, segmentation models, and pose
 007 estimators. Yet it remains an open challenge how to realize
 008 this vision without solely relying on handcrafted prompting
 009 strategies or enforcing fixed, predefined tool pipelines that
 010 limit VLMs' ability to discover optimal tool-use patterns.
 011 Reinforcement Learning could overcome this gap, but has
 012 so far been limited to reasoning with a single visual tool due
 013 to the large search space in multi-tool reasoning. We intro-
 014 duce Double Interactive Reinforcement Learning (DIRL),
 015 a two-phase training framework where VLMs learn to co-
 016 ordinate multiple tools through interactive exploration and
 017 feedback. In the teaching phase, we combine demonstra-
 018 tions from a single tool specialist trained via interactive RL
 019 with traces from a frontier model using all tools. In the ex-
 020 ploration phase, the model further refines multi-tool coordi-

nation through continued RL. Our model, SpaceTools, with
 tool-augmented spatial reasoning ability, achieves state-of-
 the-art performance on spatial understanding benchmarks
 (RoboSpatial-Home, BLINK, BOP-ASK) and demonstrates
 reliable real-world manipulation using a 7-DOF robot as
 a tool. DIRL provides substantial improvements over the
 vanilla SFT (+12% on RoboSpatial) and RL (+16% on Ro-
 boSpatial) baselines.

1. Introduction

Spatial reasoning—the ability to understand geometric rela-
 tionships between objects and their environment—is an im-
 portant capability for vision-language models (VLMs). It
 enables models to answer geometric questions, such as rel-
 ative positions, spatial configurations, and physical affor-
 dances, which is vital to support the integration of VLMs
 into embodied systems, such as robots. While recent VLMs
 [3, 27, 28, 46] have achieved strong performance on open-
 ended visual questions, their ability to do spatial under-
 standing remains an active field of research [22, 40, 69, 75],

040 particularly in settings that require diverse multi-step reason-
041 ing intertwined with precise geometric perception and
042 3D awareness (see Figure 1). These challenges are ampli-
043 fied in robotics, where perception must seamlessly translate
044 into decision-making and physical action [24].

045 The conventional approach to teach VLMs new capabili-
046 ties involves fine-tuning on task-specific datasets [8, 12, 20,
047 32, 49, 81], an approach limited by the need for large-scale
048 annotations and extensive data engineering. We present a
049 scalable alternative: we empower VLMs to use tools, that is,
050 to call computer vision and robotics modules when needed,
051 and use their outputs to aid in solving the spatial reason-
052 ing task. Such tool use provides access to precise mea-
053 surements and intermediate geometric representations, can
054 leverage computer vision models from VLM-incompatible
055 settings (*e.g.*, dense prediction), and allows combining the
056 strengths of heterogeneous models to augment base-model
057 capability. Recently, ViGoRL [50] demonstrated that re-
058 inforcement learning can enable a VLM to learn grounded
059 reasoning with a single visual tool, namely a cropping op-
060 eration, showing the promise of interactive RL for tool use.
061 However, naive application of RL to many tools creates a
062 prohibitively large search space where exploration fails to
063 discover effective policies.

064 To address this gap, we introduce Double Interactive
065 Reinforcement Learning (DIRL), a two-phase framework
066 where interactive RL is applied twice. The key insight is
067 that RL with a pointing tool is tractable and teaches ground-
068 ing, while multi-tool RL can refine diverse reasoning, but
069 requires good initialization for stable learning. DIRL uses a
070 two-phase training scheme with a teaching phase followed
071 by an exploration phase. In the teaching phase, the model
072 is trained with Supervised Fine-Tuning (SFT) on the basics
073 of tool usage—method signatures, outputs, and information
074 flow using a mix of single-tool Interactive Reinforcement
075 Learning (IRL) traces and multi-tool demonstrations. In
076 the exploration phase, we apply Interactive Reinforcement
077 Learning (IRL) with the full toolset, enabling the model to
078 refine tool coordination for spatial reasoning tasks.

079 Unlike prior work, DIRL allows the model to call tools
080 interactively during training, instead of relying on fixed
081 pipelines or precomputed contexts (Table 1), enabling this
082 behavior at scale requires addressing a key systems chal-
083 lenge: how to efficiently serve diverse, compute-intensive
084 tools during interactive training. To address this, we de-
085 velop *Toolshed*, a platform which hosts computationally in-
086 tensive computer vision tools such as SAM2 [48], Depth
087 Pro [6], RoboRefer [81], and GraspGen [42] as rapid on-
088 demand services during training, decoupling tool resource
089 management from RL or inference workloads, and achiev-
090 ing high tool throughput and utilization. By incorporating
091 real and stochastic tool outputs into the learning loop, DIRL
092 exposes models to actual tool behavior, encouraging reason-

Table 1. Comparison of related work for training supervision and tool-call interactivity during training. ‘-’ indicates that only a single tool is used.

| Method | SFT | RL | Use tools | Non-fixed tool pipeline | Interactive tool call |
|----------------------|-----|----|-----------|-------------------------|-----------------------|
| SpatialVLM [10] | ✓ | × | × | × | × |
| RoboRefer [81] | ✓ | ✓ | × | × | × |
| SpatialPIN [35] | × | × | ✓ | × | × |
| APC [25] | × | × | ✓ | × | × |
| ViGoRL [50] | × | ✓ | - | ✓ | × |
| SpatialReasoner [37] | × | ✓ | × | × | × |
| TIGeR [18] | ✓ | ✓ | ✓ | ✓ | × |
| SpaceTools (ours) | ✓ | ✓ | ✓ | ✓ | ✓ |

ing about tool reliability and discovering improved ways to query the tools.

We conduct extensive experiments on a diverse set of spatial reasoning problems, such as determining object-location fit, estimating distances between items, reasoning about occlusions and orientations, pose estimation, and predicting grasping affordances. Our trained model, *SpaceTools*, achieves state-of-the-art performance across multiple spatial reasoning benchmarks, including RoboSpatial-Home [55], BLINK [16], RefSpatial [81], CVBench [81], and BOP-ASK [4]. By integrating a real robot as a tool, *SpaceTools* completes pick-and-place tasks with an 86% success rate, demonstrating effective transfer from spatial reasoning to embodied control and outperforming frontier models equipped with the same tools. In summary, our contributions are:

1. **DIRL**: a novel training paradigm that enables interactive training with a large set of tools.
2. **Toolshed**: an interactive platform for hosting diverse computer vision tools, to be open-sourced.
3. **SpaceTools**: A VLM trained for spatial reasoning via interactive multi-tool use, which achieves state-of-the-art results across spatial reasoning benchmarks and performs robot control via alternating perception and action tool calls.

2. Related Work

Spatial Reasoning with VLMs. Spatial reasoning with VLMs [3, 27–29, 46] refers to understanding geometric relationships among objects and their environment [2, 16, 26, 47, 54, 57, 67]. Recent progress shows that VLMs can increasingly support robots in perceiving and interacting with the physical world [7, 55, 81]. However, VLM spatial reasoning remains insufficient for real-world robotic demands, where multi-step reasoning, precise geometric understanding, and strong 3D awareness are required [44, 69, 75]. Conventional approaches teach VLMs spatial understanding by fine-tuning on task-specific question-answering datasets [9, 12, 31, 49, 50, 55, 65, 81]. Yet these

methods require large-scale data collection and architecture modifications even to introduce a single low-level perceptual capability such as depth [8], pointing [14, 55, 81], and 3D-awareness [36, 38]. Instead of baking all perceptual skills into the model, we propose to enable VLMs to invoke external computer vision and robotics tools as needed, allowing them to solve spatial reasoning tasks and perform real-world manipulation.

Tool-augmented Reasoning. Tool-augmented reasoning aims to enrich model capabilities by supplying additional information from external modules [11, 21, 30, 74]. Typical applications include integrating search engines [11, 15, 21], calculators [43, 77], or code executors [56, 64] into LLMs, and vision tools for VLMs [19, 39, 71]. In the context of spatial reasoning, the community has explored equipping VLMs with vision tools during intermediate reasoning steps. However, most approaches rely on handcrafted prompting strategies [17, 19, 41, 68] or enforce a fixed, predefined tool pipeline [25, 35] in a training-free way, which limits their ability to handle diverse, precise, and 3D-aware reasoning required for robotics. TIGeR [18] is a concurrent work we learned of during the preparation of this manuscript. They focus more on problem-solving via code generation and not interactive learning, deriving their supervision from a predefined synthetic tool pipeline with large-model-based rewriting. In contrast, we enable the model to learn to coordinate a diverse set of vision and robotic tools through both teacher demonstrations involving real tool interactions and self-exploration enabled by interactive RL.

Reinforcement Learning for Reasoning. Reinforcement learning (RL) has been widely applied to enhance the reasoning capabilities of LLMs or VLMs on verifiable tasks such as math [51], coding [13, 61], and general visual question answering (VQA) [70, 73, 76, 78]. Recent work further explores RL for spatial reasoning, enabling models to produce interpretable or grounded reasoning [23, 33, 52, 65, 66]. Some works adopt RL to strengthen chain-of-thought style reasoning before predicting answers [44, 63], while others focus on teaching grounded spatial understanding [50, 65, 80]. Although prior works demonstrate that RL can teach spatial reasoning with use of a single light-weight tool (e.g., cropping), scaling to multiple heterogeneous tools poses a fundamental challenge: with 10+ tools, the action space grows combinatorially, causing naive RL exploration to fail. Our training paradigm decomposes the problem into progressive and tractable phases, enabling the model to learn effective coordination strategies with diverse tools.

3. Problem Formulation

We formulate spatial reasoning as a sequential decision-making problem where a VLM policy π_θ interacts with external tools $\mathcal{Q}_{\text{tools}}$ to respond to a user query \mathcal{I} , which may

Algorithm 1 Spatial Reasoning with Tools

Require: VLM π_θ , User Query \mathcal{I} , Max Turns T_{max}

Ensure: Answer A_{final}

```

1:  $t \leftarrow 1, h_1 \leftarrow \mathcal{I}$  ▷ Initialize dialogue history
2: while  $t \leq T_{\text{max}}$  do ▷  $t$  is a counter
3:    $a_t \leftarrow \pi_\theta(h_t)$  ▷ Generate VLM response
4:    $h_{t+1} \leftarrow h_t \oplus a_t$ 
5:   if <answer> detected in  $a_t$  then
6:      $A_{\text{final}} \leftarrow \text{Parse}(a_t, \text{<answer>}, \text{</answer>})$ 
7:     break ▷ Final turn: task is complete
8:   else if <tool.call> detected in  $a_t$  then
9:      $\mathcal{Q}_{\text{tools}} \leftarrow \text{Parse}(a_t, \text{<tool.call>}, \text{</tool.call>})$ 
10:    for each  $q \in \mathcal{Q}_{\text{tools}}$  do
11:       $h_{t+1} \leftarrow h_{t+1} \oplus \text{CallTool}(q)$  ▷ Execute tool
12:    end for
13:  end if
14:   $t \leftarrow t + 1$ 
15: end while
16: return  $A_{\text{final}}$ 

```

consist of an image-text pair or a robotic manipulation task. The model can reason and interact with tools in multiple turns until it produces a final answer A_{final} or reaches a maximum of T_{max} interaction steps.

At each step t , the VLM receives the historical context h_t , which contains the full dialogue between the user, the VLM, and the tools (initialized as $h_1 = \mathcal{I}$). The model then generates a response a_t according to its policy: If a_t includes tool calls, tools are executed sequentially. Their outputs, together with a_t , are appended to the historical context h_t to form h_{t+1} . The updated context is then used to generate the next-step response.

The complete workflow is outlined in Algorithm 1. The model is required to follow a structured conversational format: reasoning is enclosed within `<think>` tags, tool calls within `<tool.call>` tags, and the final answer within `<answer>` tags.

The goal of this work is to **learn a policy** π_θ that addresses user queries through multi-turn interaction with vision and robotic tools. To achieve this, we propose a new training paradigm accompanied by a novel tool platform.

4. Double Interactive Reinforcement Learning

Training a VLM to reason and act through external tools benefits from both teacher-guided supervision and interactive exploration. We introduce **Double Interactive Reinforcement Learning (DIRL)**, a two-stage framework that unifies these two forms of learning. Enabling DIRL requires seamless communication between the VLM and a diverse set of vision and robotic tools during both data collection and training. We solve this challenging problem by designing **Toolshed**, a distributed infrastructure that manages large-scale tool interaction.

214 **4.1. DIRL**

215 We introduce a new training paradigm that enables VLMs
 216 to effectively use multiple tools. Our approach is motivated
 217 by two observations: (i) naïvely applying IRL (interactive
 218 RL) to all tools at once creates an extremely large search
 219 space, resulting in weak optimization signals, and (ii) pure
 220 SFT on tool-interaction traces yields models that struggle to
 221 coordinate with tools effectively or to go beyond the training
 222 traces. Our method, DIRL, addresses these limitations
 223 and improves the model’s ability to integrate and sequence
 224 multiple tools effectively. DIRL is composed of two phases,
 225 a teaching phase and an exploration phase.

226 **Teaching phase.** This phase establishes basic tool use ca-
 227 pabilities without the exploration challenges of full multi-
 228 tool RL. We build the teaching dataset from two comple-
 229 mentary sources. First, we apply IRL to train the base
 230 model to use a single pointing tool for spatial reasoning
 231 tasks (*e.g.*, spatial relationship, spatial compatibility, and
 232 relative depth are trained together). This constrained search
 233 space, allows IRL to reliably converge and produce com-
 234 petent behavior. The resulting *IRL-trained teacher* is then
 235 used to generate supervised demonstrations of grounded
 236 reasoning for the first portion of our teaching dataset. Sec-
 237 ond, we prompt a *universal teacher*, which is a frontier
 238 model, to solve spatial reasoning and robot manipulation
 239 tasks with the full set of tools (*e.g.*, pointing, segmentation,
 240 3D bbox, *etc.*), retaining only trajectories that lead to correct
 241 solutions. Finally, we combine both datasets—one part gen-
 242 erated by the IRL-trained teacher and three parts from the
 243 universal teacher—to form the complete teaching dataset.
 244 We then perform supervised fine-tuning (SFT) on the base
 245 model, yielding a policy with initial tool-usage behaviors.

246 **Exploration phase.** This phase refines multi-tool coordi-
 247 nation through interactive exploration. We resume IRL
 248 training on all tasks from the SFT-initialized policy with ac-
 249 cess to all available tools, allowing the model to enhance
 250 tool chaining strategies. The strong initialization prevents
 251 exploration collapse in the large multi-tool action space,
 252 while interactive feedback offers additional refinement of
 253 tool coordinations. These two rounds of IRL give our
 254 method its name, as DIRL involves two IRL phases—one
 255 for teaching and one for exploration.

256 **Policy Update.** We employ Group Relative Policy Op-
 257 timization (GRPO) [51] as our RL training algorithm, as
 258 visualized in Figure 2. For each input \mathcal{I} , in total N roll-
 259 out procedures are launched asynchronously under the cur-
 260 rent policy π_θ . Each rollout proceeds as Algorithm 1,
 261 generating in total N multi-turn rollouts $O_1, O_2, \dots,$
 262 O_N . Their rewards are calculated as r_1, r_2, \dots, r_N , and

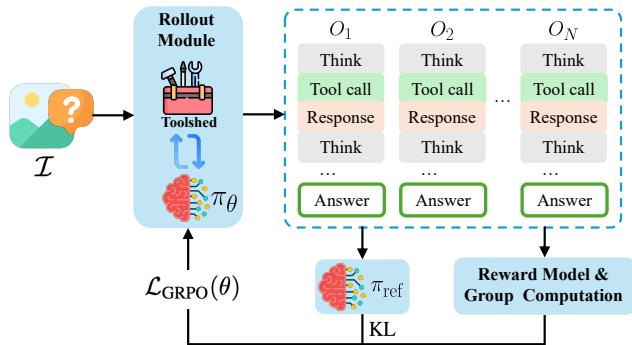


Figure 2. Interactive reinforcement learning (IRL) with Toolshed. The rollout module executes multi-turn trajectories under policy π_θ , alternating between reasoning and tool use before answering. Task rewards are aggregated and used to update π_θ via GRPO with KL regularization against π_{ref} .

the policy is updated by optimizing the GRPO objective $\mathcal{L}_{\text{GRPO}}(r_1, \dots, r_N)$, described in full in the Appendix.

265 **4.2. Toolshed**

266 Our method, DIRL, assumes access to an efficient system
 267 for invoking tools during training. In prior work, tool usage
 268 is either tightly coupled with the training loop, limiting
 269 to simple tools (*e.g.*, cropping [50]), or in case of text-only
 270 tools (*e.g.*, search [21]), highly decoupled via web APIs that
 271 lack the throughput needed for VLM interactive learning
 272 with images. Others side-step the issue altogether by using
 273 pre-computed tool outputs [18], preventing models from
 274 learning interactive, state-dependent tool use.

275 We introduce Toolshed, a scalable framework for deploy-
 276 ing multiple compute-heavy tools alongside training or
 277 inference workloads that mitigates these bottlenecks
 278 through: (1) resource and environment isolation for each
 279 tool instance; (2) decoupled scaling and execution from the
 280 policy’s main inference loop; and (3) asynchronous paral-
 281 lel workers per tool, allowing scaling tool resources inde-
 282 pendently from training resources. Toolshed hosts modular
 283 vision tools (*e.g.*, segmentation, pointing, monocular depth,
 284 3D box fitting, grasp prediction, and various image opera-
 285 tions) and robotic tools (*e.g.*, image capture, grasp execu-
 286 tion, object placement). Implementation details and com-
 287 plete tool APIs are provided in the Appendix.

288 **4.3. Rewards**

289 Reinforcement learning covers spatial reasoning tasks such
 290 as multiple choice question answering, 2D bounding box
 291 localization, pointing, pose, and grasp estimation. We de-
 292 sign normalized, task-specific rewards based on the cor-
 293 rectness of the final answer A_{final} . Each reward measures
 294 the accuracy or geometric consistency of A_{final} against
 295 the ground-truth label or annotation. We additionally experi-

Table 2. Performance comparison across spatial reasoning benchmarks. All values are normalized accuracy (%). **Bold** indicates the best performance within each column, and underline denotes the second-best result. Values of 0 indicate the model either fails to produce valid responses, outputs answers in wrong formats, or produces entirely incorrect predictions, reflecting an inability to handle that task type.

| Model | RoboSpatial | | | BLINK | RefSpatial | CVBench | | BOP-ASK | | |
|-----------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | VQA | Vacant | Overall | Depth | | 2D Rel. | 3D Depth | Pose | Grasp-MACE | Grasp-SR |
| <i>Proprietary Models</i> | | | | | | | | | | |
| Claude Sonnet 4.5 | 75.44 | 23.77 | 57.43 | 78.23 | 7.49 | 89.85 | 78.50 | 1.67 | 40.12 | <u>48.33</u> |
| GPT-4o | 61.61 | 25.10 | 48.88 | 63.71 | 8.48 | 88.77 | 75.50 | 0.00 | 5.50 | 1.67 |
| GPT-5 | 76.50 | 22.17 | 58.39 | 66.13 | 23.10 | <u>95.54</u> | 91.33 | 9.03 | <u>39.59</u> | 41.67 |
| Gemini-ER 1.5 | <u>79.30</u> | 31.10 | <u>62.50</u> | 69.23 | 41.72 | <u>95.54</u> | 90.50 | 0.00 | 30.06 | 23.33 |
| <i>General Open-Source Models</i> | | | | | | | | | | |
| LLaVA-NeXT-8B | 69.31 | 0.00 | 45.15 | 53.23 | 0.78 | 72.15 | 73.67 | 0.00 | 5.04 | 1.67 |
| Qwen2.5-VL-32B | 61.84 | 3.28 | 41.43 | 70.16 | 7.28 | 90.46 | 86.67 | 0.00 | 29.86 | 23.33 |
| Qwen2.5-VL-3B | 53.07 | 0.00 | 35.71 | 70.98 | 0.00 | 70.62 | 65.33 | 0.00 | 6.06 | 0.00 |
| <i>Spatial VLMs</i> | | | | | | | | | | |
| SpaceLLaVA-13B | 61.00 | 2.50 | 40.61 | 51.61 | 3.25 | 61.08 | 62.83 | 0.00 | 0.00 | 0.00 |
| RoboPoint-13B | 70.18 | 19.70 | 52.58 | 54.84 | 15.59 | 74.00 | 76.50 | 0.00 | 0.00 | 0.00 |
| Molmo-7B | 39.92 | 0.82 | 26.29 | 54.03 | 0.00 | 72.15 | 73.33 | 0.00 | 36.74 | 18.33 |
| RoboBrain2.0-7B | 59.64 | 44.35 | 54.31 | 84.68 | 32.50 | 87.23 | 90.00 | 0.00 | 0.00 | 0.00 |
| RoboRefer-8B-SFT | 58.33 | 61.48 | 59.43 | <u>88.71</u> | <u>48.37</u> | 96.31 | 96.50 | 0.00 | 0.00 | 0.00 |
| <i>Tool-free Fine-tuning</i> | | | | | | | | | | |
| Qwen2.5-VL-3B-Tool-free SFT | 66.66 | 41.80 | 58.00 | 80.65 | 20.22 | 91.54 | 83.33 | 2.44 | 39.47 | 35.00 |
| Qwen2.5-VL-3B-Tool-free RL | 67.54 | 28.69 | 54.00 | 80.65 | 23.10 | 87.38 | 70.83 | <u>12.00</u> | 38.79 | 36.67 |
| SpaceTools-3B (Ours) | 79.38 | <u>52.46</u> | 70.00 | 90.32 | 53.07 | 94.92 | <u>96.00</u> | 34.37 | 43.06 | 50.00 |

296 mented with a structural *format score* to encourage output
 297 correctness, but found it provided no measurable improve-
 298 ment and excluded it from final training. See details in the
 299 Appendix.

- 300 • **Multiple choice questions.** The reward is binary: $R_B =$
 301 1 if A_{Final} is correct, else 0.
- 302 • **2D bounding boxes.** We compute Mean IoU (MIoU)
 303 between predicted and ground-truth boxes: $R_{\text{MIoU}} =$
 304 $\frac{1}{N} \sum_{i=1}^N \max_j \text{IoU}(\hat{B}_i, B_j)$, where \hat{B}_i and B_j denote
 305 predicted and ground-truth boxes.
- 306 • **Pointing.** For single-point spatial prediction, we use
 307 the Normalized Negative Distance to Centroid (NNDC):
 308 $R_{\text{NNDC}} = \frac{\exp(-5d) - \exp(-5\sqrt{2})}{1 - \exp(-5\sqrt{2})}$, where d is the dis-
 309 tance to the target-region centroid. To emphasize pre-
 310 cision, we clip with the binary accuracy term: $R =$
 311 $\max(R_{\text{NNDC}}, R_B)$.
- 312 • **Pose estimation.** Predicted and ground-truth poses are
 313 converted to eight 2D projected corners. The reward is the
 314 IoU between convex hulls of predicted (\hat{C}) and ground
 315 truth (C) corner sets. $R_{\text{IoU}} = \text{IoU}(\hat{C}, C)$ when both sets
 316 are valid ($|\hat{C}| = |C| = 8$), and 0 otherwise.
- 317 • **Grasp estimation.** We adopt the Normalized Neg-
 318 ative Coordinate Error (NNCE): $R_{\text{NNCE}} = 1 -$
 319 $\frac{1}{\delta_{\text{max}}} \min\left(\delta_{\text{max}}, \frac{1}{N} \sum_{i=1}^N \frac{\|\hat{p}_i - p_i\|_2}{d}\right)$, where \hat{p}_i and p_i are
 320 predicted and ground-truth contact points, d is the gripper
 321 width, N is the number of reference points, and δ_{max} caps
 322 extreme errors. This rewards accurate geometric grasp
 323 alignment. In this work, $\delta_{\text{max}} = 10$.

324 5. Experiments

325 **Dataset.** During the first phase of DURL, we generate a
 326 teaching SFT dataset composed of 8k high-quality tool-use
 327 trajectories: 6k from the universal teacher and 2k from the
 328 IRL-trained teacher. The IRL teacher is trained to use a
 329 Pointing tool (RoboRefer [81]), a common first step be-
 330 fore querying other vision and robotic tools in spatial rea-
 331 soning. For the universal teacher, we use Claude Sonnet
 332 4.5 [1], integrated with Toolshed, consisting of all tools.
 333 Image-question pairs are sampled from RoboSpatial [55],
 334 RefSpatial [81], and BOP-ASK [4]. To extend our setup
 335 to robot manipulation, we augment the HOPE dataset [62]
 336 with grasping and pick-and-place control tasks. Each tra-
 337 jectory in the SFT dataset is a multi-turn dialogue between
 338 the user, assistant, and tools. The same spatial reasoning
 339 image-question pairs in the teaching SFT dataset are used
 340 in the stage-2 IRL.

341 **Training.** We use Qwen2.5-VL-3B-Instruct [3] as our
 342 base model. During training, we use a system prompt that
 343 defines the available tools, their schemas, and usage for-
 344 mats (detailed prompt is in the Appendix). High-level in-
 345 structions within the prompt further specify the model’s
 346 reasoning structure and output format. We use the follow-
 347 ing perception tools: segmentation [48], pointing [14, 81],
 348 depth estimation [6], cuboid fitting, grasp generation [42],
 349 cropping, array indexing, perspective projection, and robot
 350 tools: image capture, depth capture, grasp, place. The first-



Figure 3. Spatial reasoning examples of SpaceTools. It performs diverse spatial reasoning tasks including relative depth, pose, grasp, spatial compatibility, and spatial relationship by interleaving reasoning (gray) and vision tool calls (green) before producing the final answer. Images are taken from BLINK [16], RoboSpatial-Home [55], and BOP-ASK [4].

351 stage SFT starts with the base model and uses cross-entropy
352 next-token prediction loss over all assistant turns in a multi-
353 turn dialogue, training the policy π_θ to imitate the teachers'
354 reasoning processes and tool-use behaviors. In the second-
355 stage IRL, the SFT-initialized model interacts with Tool-
356 shed to refine its tool-augmented spatial reasoning through
357 interactive feedback from task rewards. For SFT, we use
358 the LLaMA-Factory platform [79], and for RL we integrate
359 Toolshed into the VERL framework [53]. More implementa-
360 tion details are in the Appendix.

5.1. Spatial Reasoning Evaluation 361

Benchmarks and Metrics. We evaluate our model 362
363 on a suite of spatial reasoning benchmarks, including
364 RoboSpatial-Home [55] (spatial VQA and vacant space
365 pointing), CVBench [59] (2D relations and 3D relative
366 depth), RefSpatial [81] (placement, location, and unseen),
367 BLINK [16] (relative depth), and BOP-ASK [4]. They
368 cover positional relationship understanding, depth estima-
369 tion, pointing, 3D pose estimation, and robotic grasp predic-
370 tion. We adopt the following metrics: (1) Answer accuracy
371 for multiple-choice and pointing questions. (2) For object

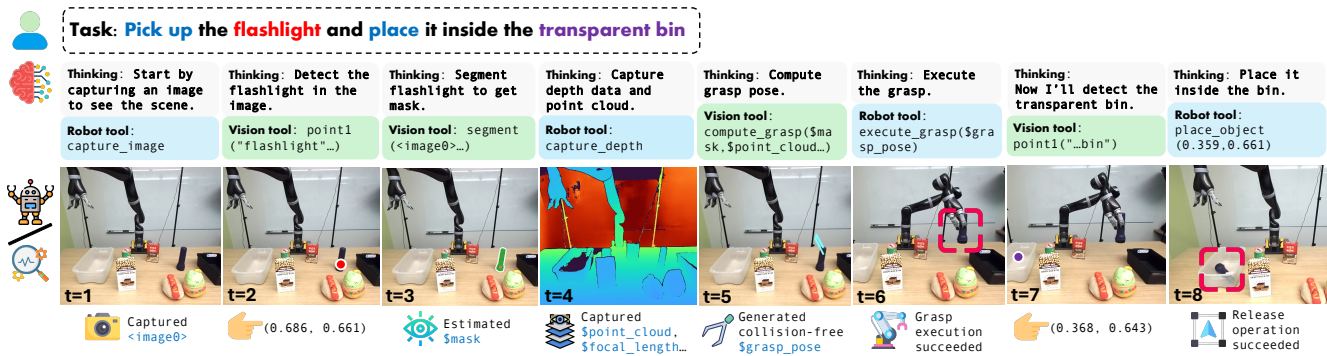


Figure 4. Real-world robot manipulation fully controlled by SpaceTools. The model completes a multi-step task, “picking up the flashlight and placing it in the transparent bin”, via alternating reasoning (gray), vision tools (green) for perception, and robot tools (blue) for action.

pose estimation, we use the normalized Intersection-over-Union (IoU) in range $[0, 100]$ (%). (3) For grasp estimation, which outputs five 2D coordinates of grasp center, and two finger bases and tips, we use the *Mean Angular Coordinate Error (MACE)* to jointly score grasp location and finger-orientation, defined in the Appendix. We report MACE as a normalized score in range $[0, 100]$ (%), and the *Success Rate (SR)* as the percentage of grasps achieving $MACE > 40$.

Baselines. We compare our model (SpaceTools) against four categories of baselines. (1) **Proprietary models** include Claude Sonnet 4.5 [1], GPT-4o [45], GPT-5 [46], and Gemini-ER 1.5 [60], which represent state-of-the-art commercial vision-language systems. (2) **General open-source models** include LLaVA-NeXT-8B [28] and Qwen2.5-VL-32B [3], which serve as publicly available multimodal foundations without spatial specialization. (3) **Spatial VLMs** include SpaceLLaVA-13B [10], RoboPoint-13B [72], Molmo-7B [14], RoboBrain2.0-7B [58], and RoboRefer-8B-SFT, which are trained with additional spatial reasoning or robotic data. (4) **Tool-free fine-tuning** contains variants of the same base model (Qwen2.5-VL-3B) trained without tool use, only on the 8k source question and answer samples¹ from DURL’s stage-1: (4a) *Tool-free SFT* is a supervised fine-tuning baseline. (4b) *Tool-free RL* applies reasoning RL à la Deepseek R1 [13] without tool use.

5.2. Spatial Reasoning Results

As shown in Table 2, SpaceTools achieves state-of-the-art results on nearly all benchmarks, surpassing proprietary, open-source, and spatial VLM baselines. SpaceTools outperforms Gemini-ER 1.5 by +7.5% on RoboSpatial, exceeds Claude Sonnet 4.5 by +24.4% on pose estimation, and surpasses GPT-5 by +8.3% on grasp prediction. Moreover, tool-augmented training yields substantially stronger results on spatial reasoning than tool-free fine-tuning of the same base model on the the same 8k VQA pairs regardless of

¹no tool calling

Table 3. Real-world robotic manipulation performance of SpaceTools and zero-shot VLM baselines equipped with Toolshed. Values are success rates (%) for *Pick* and *Relation Pick* tasks, partial success rates (%) for *Pick & Place*, and seconds for Time-to-First-Movement (TTFM).

| Model | Real Robot Manipulation Tasks | | | |
|------------------------------|-------------------------------|-----------------|-------------------|------|
| | Pick | Rel. Pick | Pick & Place | TTFM |
| $\pi_{0.5}$ | 0 (0/7) | 0 (0/6) | 0 (0/14) | 1s |
| GPT-5 + Toolshed | 71 (5/7) | 33 (2/6) | 65 (9/14) | 36s |
| Claude Sonnet 4.5 + Toolshed | 86 (6/7) | 50 (3/6) | 79 (11/14) | 30s |
| Qwen2.5-VL-3B + Toolshed | 0 (0/7) | 0 (0/6) | 0 (0/14) | - |
| SpaceTools (Ours) | 86 (6/7) | 83 (5/6) | 86 (12/14) | 10s |

learning technique. SpaceTools-3B achieves higher accuracy on all tasks, notably +12% and +16% on RoboSpatial, than tool-free SFT and RL respectively.

Figure 3 shows qualitative examples. We find that SpaceTools dynamically adapts its reasoning and tool-use strategies to each task. For example, it primarily relies on pointing for tasks such as spatial compatibility and relationship; it invokes depth estimation for relative-depth queries; and it composes multiple tools for more advanced reasoning like pose or grasp prediction. Moreover, SpaceTools has learned corrective behaviors, such as falling back to self-estimation when a tool fails, or switching to alternative pointing tools to refine uncertain detections. Therefore, the model has *learned internal procedures* for tool selection, ordering, and error recovery, rather than relying on hand-crafted pipelines in prior works [25, 34].

5.3. Experiments on Real Robot Manipulation

In order to validate SpaceTools we conduct an experiment where robotics controls are presented as tools, see Figure 4. The robot arm serves as an action tool, complementing vision-based perception tools. By alternating between perception (pointing, segmentation, depth, grasp estimation) and action (capture, grasp) tools, the VLM orchestrates a closed perception–action loop fully guided by language rea-

Table 4. Ablation on training recipes. IRL-T denotes the IRL-trained teacher; Univ-T denotes the universal (frontier-model) teacher; S2-IRL denotes the Stage-2 interactive RL phase. Checkmarks indicate which components are included.

| Variant | IRL-T | Univ-T | S2-IRL | RoboSpatial | RefSpatial | Pose | Mean |
|-------------------------------|-------|--------|--------|-------------|------------|-------|-------|
| <i>with Interactive RL</i> | | | | | | | |
| SpaceTools (Ours) | ✓ | ✓ | ✓ | 70.00 | 53.07 | 34.37 | 52.48 |
| w/o IRL Teacher | × | ✓ | ✓ | 61.14 | 29.60 | 34.29 | 41.68 |
| w/o Univ. Teacher | ✓ | × | ✓ | 65.14 | 54.51 | 8.92 | 42.86 |
| w/o Stage 2 IRL | ✓ | ✓ | × | 67.71 | 51.98 | 33.28 | 50.99 |
| <i>without Interactive RL</i> | | | | | | | |
| Tool SFT | × | ✓ | × | 59.71 | 24.91 | 32.94 | 39.19 |
| Tool NIRL | × | ✓ | × | 55.14 | 28.16 | 30.89 | 38.06 |

soning, in contrast to prior work where robot action is an external process to model reasoning [55]. We evaluate SpaceTools, Claude Sonnet 4.5, and GPT-5 in this tool-augmented system as well as comparing with a strong vision-language-action model, $\pi_{0.5}$ [5]. We focus on three type of tasks; pick, relational pick, and pick & place, results from this experiments are presented in Table 3. During the experiments we observed that SpaceTools is better grounded in spatial reasoning as well as being capable of orchestrating multiple tools, whereas other methods, like GPT-5, fail to chain tools coherently, sometimes inventing grasp poses or camera intrinsics instead of reusing computed values. Please consult the Appendix for further details.

5.4. Ablation Study

To analyze the contribution of each component in the DIRL framework, we perform systematic ablations on spatial reasoning benchmarks by removing (1) the *IRL-trained teacher* (IRL-T), (2) the *universal teacher* (Univ-T), and (3) the *Stage 2 IRL* phase (S2-IRL).

In addition, to evaluate the importance of interactive reinforcement learning, we compare DIRL with two classic non-interactive training schemes: (a) *Tool SFT* with the universal teacher, where the model is trained on multi-turn tool-use traces through direct supervision, and (b) *Tool Non-Interactive Reinforcement Learning (Tool NIRL)*, which follows the conventional tool-learning setup in large language models (LLM) [77]. In Tool NIRL, ground-truth tool call traces are required, and the reward is based on the correctness of tool names, tool arguments, and answers. Detailed configurations are provided in the Appendix.

Quantitative results are in Table 4, with our main findings summarized: (1) Removing the IRL-trained teacher leads to a sharp performance drop, particularly on tasks requiring fine spatial grounding such as RefSpatial and RoboSpatial. (2) Removing the universal teacher also degrades performance, especially on pose tasks that require multi-tool composition (*e.g.*, segmentation + depth + 3D bbox). (3) Stage 2 IRL provides the final boost of tool-augmented reasoning. Eliminating the Stage 2 IRL phase affects per-

Table 5. Comparison of proprietary models with and without the Toolshed enhancement across robotic spatial reasoning benchmarks. Values are normalized accuracy (%).

| Model | RoboSpatial | BLINK | RefSpatial | BOP-ASK | |
|------------|-------------|---------|------------|---------|--------------|
| | | | | Pose | Grasp (MACE) |
| GPT-5 | 58.39 | 66.13 | 23.10 | 9.03 | 39.59 |
| + Toolshed | 55.14 | 90.32 ↑ | 36.10 ↑ | 15.00 ↑ | 41.49 ↑ |
| Claude | 57.43 | 78.23 | 7.49 | 1.67 | 40.12 |
| + Toolshed | 52.86 | 75.00 | 27.80 ↑ | 25.00 ↑ | 44.19 ↑ |

formance across RoboSpatial, RefSpatial, and pose tasks. (4) Both Tool SFT and Tool NIRL baselines underperform DIRL by a large margin (+13.4 and +14.4 mean improvement, respectively). This suggests that interactive RL is key to teaching VLMs consistent reasoning over complex tool sequences.

6. Discussion & Conclusion

Agentic VLMs hold the promise of reasoning through arbitrary external tools. Motivated by this, we examine whether large VLMs can improve their spatial reasoning by leveraging vision tools in a fully zero-shot setting. As shown in Table 5, tool integration yields clear gains on tasks requiring precise spatial grounding or explicit geometric reasoning. For example, GPT-5 with Toolshed improves on RefSpatial (from 23.1 to 36.1) and pose estimation (from 9.0 to 15.0), suggesting that tool feedback mitigates limitations in spatial grounding and 3D understanding. In contrast, high-level tasks such as RoboSpatial and BLINK show mixed trends, as models tend to overuse tools and struggle to correctly interpret nuanced tool outputs. We also find that IRL improves out-of-domain generalization. When a model is trained to use a single powerful tool such as pointing [81], it not only performs better on its in-domain benchmark but also transfers unexpectedly well. For instance, a model trained only on RoboSpatial [55] reaches 72.3% accuracy on that benchmark and still achieves 34.3% on RefSpatial—where other fine-tuning approaches score zero. These results highlight the promise of agentic VLMs and their ability to acquire new skills through tool use.

In conclusion, we introduce DIRL, a new method for training tool-augmented VLMs through progressive and interactive learning. To support this, we built Toolshed, a system for deploying diverse tools at scale for online interaction during training. Our experiments show that our trained model, SpaceTools, achieves state-of-the-art performance on multiple spatial reasoning benchmarks and exhibits strong out-of-distribution generalization, including the ability to use a robot as a tool. This work demonstrates that VLMs can acquire complex spatial reasoning capabilities through learned tool coordination rather than architectural modification or large-scale data-driven fine-tuning.

511 **7. Acknowledgments**

512 The authors would like to thank Vineet Bha, Alex Zook,
513 Stephen Tyree, and Huijie Zhang for their inputs and com-
514 ments on this manuscript.

515 **References**

- 516 [1] Anthropic. Claude sonnet 4.5 system card. Technical report,
517 Anthropic, 2025. Accessed: 2025-11-09. 5, 7
- 518 [2] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki
519 Kawanabe. Scanqa: 3d question answering for spatial scene
520 understanding. In *Proceedings of the IEEE/CVF Conference*
521 *on Computer Vision and Pattern Recognition (CVPR)*, 2022.
522 2
- 523 [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin
524 Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun
525 Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhao-
526 hai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu,
527 Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng,
528 Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin.
529 Qwen2.5-vl technical report, 2025. 1, 2, 5, 7
- 530 [4] Vineet Bhat, Sungsu Kim, Valts Blukis, Greg Heinrich,
531 Prashanth Krishnamurthy, Ramesh Karri, Stan Birchfield,
532 Farshad Khorrani, and Jonathan Tremblay. Bop-ask:
533 Object-interaction reasoning for vision-language models,
534 2025. 2, 5, 6
- 535 [5] Kevin Black, Noah Brown, James Darpinian, Karan Dha-
536 balia, Danny Driess, Adnan Esmail, Michael Robert Equi,
537 Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya
538 Ghosh, Lachy Groom, Karol Hausman, brian ichter, Szymon
539 Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc,
540 Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj
541 Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura
542 Smith, Jost Tobias Springenberg, Kyle Stachowicz, James
543 Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan
544 Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-
545 action model with open-world generalization. In *Proceed-*
546 *ings of The 9th Conference on Robot Learning*, pages 17–40.
547 PMLR, 2025. 8
- 548 [6] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain,
549 Marcel Santos, Yichao Zhou, Stephan R. Richter, and
550 Vladlen Koltun. Depth pro: Sharp monocular metric depth in
551 less than a second. In *International Conference on Learning*
552 *Representations*, 2025. 2, 5
- 553 [7] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi
554 Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Pre-
555 cise spatial understanding with vision language models. In
556 *2025 IEEE International Conference on Robotics and Au-*
557 *tomation (ICRA)*, pages 9490–9498, 2025. 2
- 558 [8] Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory
559 Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas
560 Chandra, and Yangyang Shi. Depthlm: Metric depth from
561 vision language models, 2025. 2, 3
- 562 [9] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa
563 Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endow-
564 ing vision-language models with spatial reasoning capabili-
565 ties. In *Proceedings of the IEEE/CVF Conference on Com-*
puter Vision and Pattern Recognition (CVPR), pages 14455–
14465, 2024. 2
- [10] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa
Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endow-
ing vision-language models with spatial reasoning capabili-
ties. In *Proceedings of the IEEE/CVF Conference on Com-*
puter Vision and Pattern Recognition, pages 14455–14465,
2024. 2, 7
- [11] Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun,
Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen
Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng
Chen. Research: Learning to reason with search for llms via
reinforcement learning, 2025. 3
- [12] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Rui-
han Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatial-
rgpt: Grounded spatial reasoning in vision-language mod-
els. In *Advances in Neural Information Processing Systems*
(NeurIPS), 2024. 2
- [13] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capa-
bility in llms via reinforcement learning, 2025. 3, 7
- [14] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tri-
pathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi,
Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu,
Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo,
Yen-Sung Chen, Ajay Patel, Mark Yatskar, Chris Callison-
Burch, Andrew Head, Rose Hendrix, Favien Bastani, Eli
VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda,
Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron
Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper
Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt,
Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin
Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna,
Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross
Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo
and pixmo: Open weights and open data for state-of-the-art
vision-language models. In *Proceedings of the IEEE/CVF*
Conference on Computer Vision and Pattern Recognition
(CVPR), 2025. 3, 5, 7
- [15] Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yu-
jia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and
Wanjun Zhong. Retool: Reinforcement learning for strategic
tool use in llms, 2025. 3
- [16] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang,
Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and
Ranjay Krishna. Blink: Multimodal large language models
can see but not perceive. *arXiv preprint arXiv:2404.12390*,
2024. 2, 6
- [17] Tanmay Gupta and Aniruddha Kembhavi. Visual Program-
ming: Compositional visual reasoning without training. In
CVPR, pages 14953–14962, 2023. 3
- [18] Yi Han, Cheng Chi, Enshen Zhou, Shanyu Rong, Jingkun
An, Pengwei Wang, Zhongyuan Wang, Lu Sheng, and
Shanghang Zhang. Tiger: Tool-integrated geometric reason-
ing in vision-language models for robotics. *arXiv preprint*
arXiv:2510.07181, 2025. 2, 3, 4
- [19] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Osten-
dorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna.

- 623 Visual sketchpad: Sketching as a visual chain of thought for
624 multimodal language models. *Advances in Neural Informa-*
625 *tion Processing Systems*, 37:139348–139379, 2024. 3
- [20] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan
626 Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao
627 Mu, Pengju An, et al. Robobrain: A unified brain model for
628 robotic manipulation from abstract to concrete. In *Proceed-*
629 *ings of the Computer Vision and Pattern Recognition Con-*
630 *ference*, pages 1724–1734, 2025. 2
- [21] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan
631 Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-
632 r1: Training llms to reason and leverage search engines with
633 reinforcement learning. *arXiv preprint arXiv:2503.09516*,
634 2025. 3, 4
- [22] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What’s
635 “up” with vision-language models? investigating their strug-
636 gle with spatial reasoning. In *Proceedings of the Confer-*
637 *ence on Empirical Methods in Natural Language Processing*
638 *(EMNLP)*, 2023. 1
- [23] Dongyoung Kim, Huiwon Jang, Sumin Park, Jaehyung Kim,
639 Younggyo Seo, and Jinwoo Shin. Robot-r1: Reinforcement
640 learning for enhanced embodied reasoning in robotics. In
641 *The Thirty-ninth Annual Conference on Neural Information*
642 *Processing Systems*, 2025. 3
- [24] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao,
643 Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan
644 Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An
645 open-source vision-language-action model. *arXiv preprint*
646 *arXiv:2406.09246*, 2024. 2
- [25] Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina
647 Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-
648 aware reasoning in vision-language models via mental im-
649 agery simulation. *arXiv preprint arXiv:2504.17207*, 2025.
650 2, 3, 7
- [26] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia,
651 Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Im-
652 agine while reasoning in space: Multimodal visualization-of-
653 thought, 2025. 2
- [27] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov,
654 Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi,
655 and Song Han. Vila: On pre-training for visual language
656 models. *arXiv preprint arXiv:2312.07533*, 2023. 1, 2
- [28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee.
657 Improved baselines with visual instruction tuning. In *Pro-*
658 *ceedings of the IEEE/CVF Conference on Computer Vision*
659 *and Pattern Recognition (CVPR)*, pages 26296–26306, 2024.
660 1, 7
- [29] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan
661 Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Im-
662 proved reasoning, ocr, and world knowledge. Online blog
663 / model documentation, 2024. Improved Vision-Language
664 Model over LLaVA. 2
- [30] Weiwen Liu, Xu Huang, Xingshan Zeng, Xinlong Hao,
665 Shuai Yu, Dexun Li, Shuai Wang, Weinan Gan, Zhengying
666 Liu, Yuanqing Yu, et al. Toolace: Winning the points of llm
667 function calling. *arXiv preprint arXiv:2409.00920*, 2024. 3
- [31] Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang
668 Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang,
669 Shuang Wu, Tongtong Cao, Guowei Huang, et al. Spatialcot:
670 Advancing spatial reasoning through coordinate alignment
671 and chain-of-thought for embodied task planning. *arXiv*
672 *preprint arXiv:2501.10074*, 2025. 2
- [32] Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han
673 Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. Ssr:
674 Enhancing depth perception in vision-language models via
675 rationale-guided spatial reasoning. In *The Thirty-ninth An-*
676 *nuual Conference on Neural Information Processing Systems*,
677 2025. 2
- [33] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang
678 Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-
679 rft: Visual reinforcement fine-tuning. *arXiv preprint*
680 *arXiv:2503.01785*, 2025. 3
- [34] Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and
681 Andrew Markham. Spatialpin: Enhancing spatial reasoning
682 capabilities of vision-language models through prompting
683 and interacting 3d priors. *arXiv preprint arXiv:2403.13438*,
684 2024. 7
- [35] Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and
685 Andrew Markham. Spatialpin: Enhancing spatial reason-
686 ing capabilities of vision-language models through prompt-
687 ing and interacting 3d priors. *Advances in neural information*
688 *processing systems*, 37:68803–68832, 2024. 2, 3
- [36] Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso
689 de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: To-
690 wards explicit and generalizable 3d spatial reasoning. *arXiv*
691 *preprint arXiv:2504.20024*, 2025. 3
- [37] Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso
692 de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: To-
693 wards explicit and generalizable 3d spatial reasoning. *arXiv*
694 *preprint arXiv:2504.20024*, 2025. 2
- [38] Wufei Ma, Luoxin Ye, Celso M de Melo, Jieneng Chen, and
695 Alan Yuille. Spatialllm: A compound 3d-informed design
696 towards spatially-intelligent large multimodal models. *arXiv*
697 *preprint arXiv:2505.00788*, 2025. 3
- [39] Zixian Ma, Weikai Huang, Jieyu Zhang, Tanmay Gupta, and
698 Ranjay Krishna. m&m’s: A benchmark to evaluate tool-use
699 for multi-step multi-modal tasks, 2024. 3
- [40] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav
700 Putta, Sriram Yenamandra, Mikael Henaff, Sneha Sil-
701 wal, Paul Mccvay, Oleksandr Maksymets, Sergio Arnaud,
702 Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma,
703 Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk,
704 Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris
705 Paxton, Sasha Sax, and Aravind Rajeswaran. Openqa: Em-
706 bodied question answering in the era of foundation models.
707 In *Conference on Computer Vision and Pattern Recognition*
708 *(CVPR)*, 2024. 1
- [41] Damiano Marsili, Rohun Agrawal, Yisong Yue, and Geor-
709 gia Gkioxari. Visual agentic ai for spatial reasoning with a
710 dynamic api. *arXiv preprint arXiv:2502.06787*, 2025. 3
- [42] Adithyavairavan Murali, Balakumar Sundaralingam, Yu-Wei
711 Chao, Wentao Yuan, Jun Yamada, Mark Carlson, Fabio
712 Ramos, Stan Birchfield, Dieter Fox, and Clemens Eppner.
713 Graspgen: A diffusion-based framework for 6-dof grasping
714 with on-generator training, 2025. 2, 5

- 738 [43] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, 795
739 Long Ouyang, Christina Kim, Christopher Hesse, Shan- 796
740 tanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, 797
741 Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin But- 798
742 ton, Matthew Knight, Benjamin Chess, and John Schulman. 799
743 Webgpt: Browser-assisted question-answering with human 800
744 feedback, 2022. 3
- 745 [44] NVIDIA. Cosmos-reason1: From physical common sense to 801
746 embodied reasoning, 2025. 2, 3
- 747 [45] OpenAI. Gpt-4 technical report, 2024. 7
- 748 [46] OpenAI. Gpt-5 system card. Technical report, OpenAI, 802
749 2025. Accessed: 2025-11-09. 1, 2, 7
- 750 [47] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp 803
751 Kraehenbuehl, and Vladlen Koltun. Does spatial cognition 804
752 emerge in frontier models? In *International Conference on* 805
753 *Learning Representations*, 2025. 2
- 754 [48] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang 806
755 Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman 807
756 Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junt- 808
757 ing Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao- 809
758 Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feicht- 810
759 enhofer. Sam 2: Segment anything in images and videos. 811
760 *arXiv preprint arXiv:2408.00714*, 2024. 2, 5
- 761 [49] Arijit Ray, Jiafei Duan, Ellis L Brown II, Reuben Tan, Dina 812
762 Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kem- 813
763 bhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, 814
764 and Kate Saenko. SAT: Dynamic spatial aptitude training 815
765 for multimodal language models. In *Second Conference on* 816
766 *Language Modeling*, 2025. 2
- 767 [50] Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush 817
768 Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragki- 818
769 adaki. Vigorl: Visually grounded reinforcement learning. 819
770 *arXiv preprint arXiv:2505.23678*, 2025. 2, 3, 4
- 771 [51] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao 820
772 Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, 821
773 Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of 822
774 mathematical reasoning in open language models, 2024. 3, 4
- 775 [52] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo 823
776 Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, 824
777 Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm- 825
778 r1: A stable and generalizable r1-style large vision-language 826
779 model, 2025. 3
- 780 [53] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, 827
781 Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and 828
782 Chuan Wu. Hybridflow: A flexible and efficient rlhf frame- 829
783 work. *arXiv preprint arXiv: 2409.19256*, 2024. 6
- 784 [54] Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, 830
785 Reza Haf, and Yuan-Fang Li. An empirical analysis on spa- 831
786 tial reasoning capabilities of large multimodal models. In 832
787 *Proceedings of the 2024 Conference on Empirical Meth-* 833
788 *ods in Natural Language Processing*, pages 21440–21455, 834
789 Miami, Florida, USA, 2024. Association for Computational 835
790 Linguistics. 2
- 791 [55] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen 836
792 Tyree, Yu Su, and Stan Birchfield. RoboSpatial: Teaching 837
793 spatial understanding to 2D and 3D vision-language models 838
794 for robotics. In *Proceedings of the IEEE/CVF Conference* 839
840 *on Computer Vision and Pattern Recognition (CVPR)*, 2025. 841
842 Oral Presentation. 2, 3, 5, 6, 8
- [56] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: 843
Visual inference via python execution for reasoning. In *Pro-* 844
ceedings of the IEEE/CVF international conference on com- 845
puter vision, pages 11888–11898, 2023. 3
- [57] Emilia Szymanska, Mihai Dusmanu, Jan-Willem BURLAGE, 846
Mahdi Rad, and Marc Pollefeys. Space3D-Bench: Spatial 847
3D Question Answering Benchmark. In *European Confer-* 848
ence on Computer Vision (ECCV) Workshops, 2024. 2
- [58] BAAI RoboBrain Team. Robobrain 2.0 technical report, 849
2025. 7
- [59] CVBench Team. Cvbench: A benchmark for cross-video 850
multimodal reasoning, 2025. 6
- [60] Gemini Robotics Team. Gemini robotics 1.5: Pushing the 851
frontier of generalist robots with advanced embodied reason-
ing, thinking, and motion transfer, 2025. 7
- [61] Kimi Team. Kimi k1.5: Scaling reinforcement learning with 812
llms, 2025. 3
- [62] Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, 814
Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose 815
estimation of household objects for robotic manipulation: 816
An accessible dataset and benchmark. In *International Con-* 817
ference on Intelligent Robots and Systems (IROS), 2022. 5
- [63] Peiyao Wang and Haibin Ling. Svqa-r1: Reinforcing spatial 819
reasoning in mllms via view-consistent reward optimization. 820
arXiv preprint arXiv:2506.01371, 2025. 3
- [64] Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, 822
Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions 823
elicit better llm agents. In *Proceedings of the 41st Interna-* 824
tional Conference on Machine Learning. JMLR.org, 2024. 825
3
- [65] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, 827
Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial 828
reasoning in vision-language models with interwoven think- 829
ing and visual drawing, 2025. 2, 3
- [66] Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, 831
Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, 832
and Klara Nahrstedt. Vtool-r1: Vllms learn to think with 833
images via reinforcement learning on multimodal tool use, 834
2025. 3
- [67] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, 836
Li Dong, Lei Cui, and Furu Wei. Mind’s eye of llms: 837
visualization-of-thought elicits spatial reasoning in large lan- 838
guage models. In *Proceedings of the 38th International* 839
Conference on Neural Information Processing Systems, Red 840
Hook, NY, USA, 2024. Curran Associates Inc. 2
- [68] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan 842
Li, and Jianfeng Gao. Set-of-mark prompting unleashes ex- 843
traordinary visual grounding in gpt-4v, 2023. 3
- [69] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng 845
Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, 846
Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and 847
Tong Zhang. Embodiedbench: Comprehensive benchmark- 848
ing multi-modal large language models for vision-driven em- 849
bodied agents. In *Forty-second International Conference on* 850
Machine Learning, 2025. 1, 2

- 852 [70] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan
853 Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun
854 Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision:
855 Advancing generalized multimodal reasoning through cross-
856 modal formalization. *arXiv preprint arXiv:2503.10615*,
857 2025. 3
- 858 [71] Shaofeng Yin, Ting Lei, and Yang Liu. Toolvqa: A dataset
859 for multi-step reasoning vqa with external tools. In *Proceed-*
860 *ings of the IEEE/CVF International Conference on Com-*
861 *puter Vision (ICCV)*, pages 4424–4433, 2025. 3
- 862 [72] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay,
863 Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousa-
864 vian, and Dieter Fox. Robopoint: A vision-language model
865 for spatial affordance prediction for robotics. In *Conference*
866 *on Robot Learning (CoRL)*, 2024. Also available as arXiv
867 preprint arXiv:2406.10721. 7
- 868 [73] Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao,
869 Fan Yang, Ming Tang, and Jinqiao Wang. Vision-r1: Evolv-
870 ing human-free alignment in large vision-language models
871 via vision-guided reinforcement learning, 2025. 3
- 872 [74] Jianguo Zhang, Tian Lan, Ming Zhu, Zuxin Liu, Thai Hoang,
873 Shirley Kokane, Weiran Yao, Juntao Tan, Akshara Prab-
874 hakar, Haolin Chen, et al. xlam: A family of large ac-
875 tion models to empower ai agent systems. *arXiv preprint*
876 *arXiv:2409.03215*, 2024. 3
- 877 [75] Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze
878 Huang, Jilin Mei, Junhui Chen, Yujie Yuan, Xinyue Cai,
879 Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang.
880 From flatland to space: Teaching vision-language models to
881 perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*,
882 2025. 1, 2
- 883 [76] Jingyi Zhang, Jiaying Huang, Huanjin Yao, Shunyu Liu,
884 Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning
885 to reason with multimodal large language models via step-
886 wise group relative policy optimization. In *Proceedings of*
887 *the IEEE/CVF International Conference on Computer Vision*
888 *(ICCV)*, 2025. 3
- 889 [77] Shaokun Zhang, Yi Dong, Jieyu Zhang, Jan Kautz, Bryan
890 Catanzaro, Andrew Tao, Qingyun Wu, Zhiding Yu, and
891 Guilin Liu. Nemotron-research-tool-n1: Exploring tool-
892 using language models with reinforced reasoning, 2025. 3,
893 8
- 894 [78] Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xi-
895 aowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia,
896 Song-Chun Zhu, and Qing Li. Chain-of-focus: Adaptive
897 visual search and zooming for multimodal reasoning via rl.
898 *ArXiv*, abs/2505.15436, 2025. 3
- 899 [79] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye,
900 and Zheyuan Luo. LlamaFactory: Unified efficient fine-tuning
901 of 100+ language models. In *Proceedings of the 62nd An-*
902 *ual Meeting of the Association for Computational Linguis-*
903 *tics (Volume 3: System Demonstrations)*, pages 400–410,
904 Bangkok, Thailand, 2024. Association for Computational
905 Linguistics. 6
- 906 [80] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao,
907 Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deep-
908 eyes: Incentivizing “thinking with images” via reinforc-
909 e-ment learning, 2025. 3
- [81] 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. 2, 3, 5, 6, 8 910
911
912
913
914