

Perceive, Interact, Reason: Building Tool-Augmented Visual Agents for Spatial Reasoning

NeurIPS 2026

Changye Li¹, Meng Lu², Yi Wu¹, Ligeng Zhu³

¹Tsinghua University ²Virginia Tech ³NVIDIA

Code: <https://github.com/antoinegg1/PEDIA>
Models & Data: <https://huggingface.co/Changyeli03>

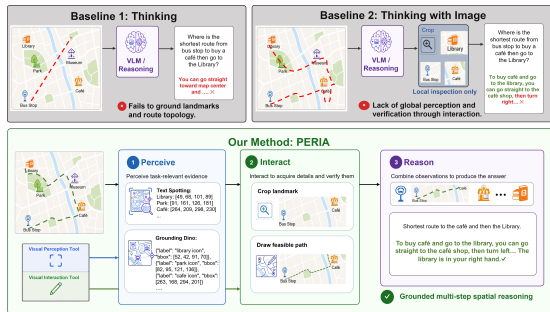
The Challenge: Spatial Reasoning Requires More Than Static Vision

Current VLMs struggle with spatial reasoning:

- Map navigation & route planning
- Visual probing & object tracking
- 3D spatial understanding

Why?

- ✗ Static visual representations
- ✗ Missing fine-grained evidence
- ✗ No multi-step interaction



Key Insight: Active evidence acquisition through tools is essential for spatial reasoning

A tool-augmented visual agent with two key components:

Vision Perception Tools

Extract evidence from images:

- Text spotting & OCR
- Map-optimized reading
- Object segmentation
- Presence verification

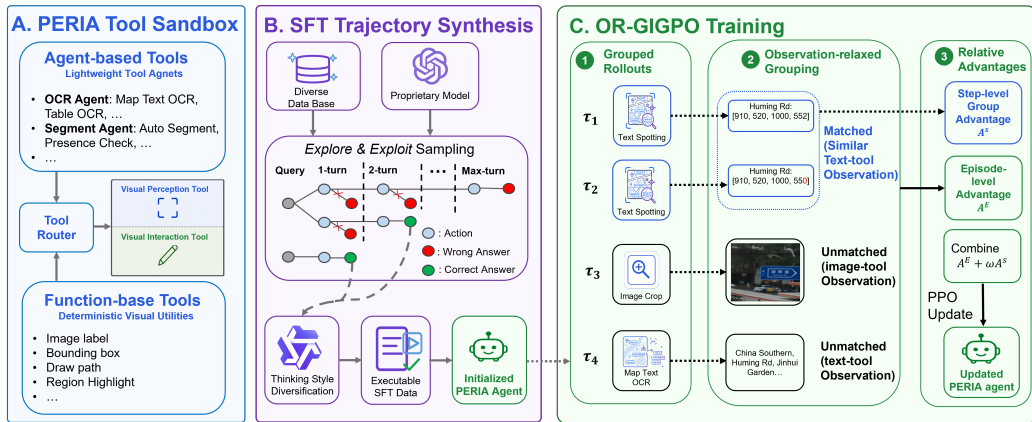
Vision Interaction Tools

Manipulate visual context:

- Path drawing
- Bounding box annotation
- Image cropping
- Region focusing

18 tools + 3 lightweight tool agents organized into a modular sandbox

Training Pipeline: From Tool Sandbox to OR-GIGPO



- ① **Tool Sandbox Construction:** Modular perception & interaction tools
- ② **SFT Trajectory Synthesis:** Explore-exploit with GPT-5 + thinking diversification
- ③ **OR-GIGPO Training:** Group-in-group optimization with relaxed observations

OR-GIGPO: Observation-Relaxed Group-in-Group Policy Optimization

Challenge: Multi-step tool-use trajectories need fine-grained credit assignment

Solution: OR-GIGPO

- **Group-in-Group:** Sample N rollouts per state, group by relaxed observations
- **Relaxed Observations:** Remove tool outputs, keep only actions & thoughts
- **Composite Advantages:** Combine step-level and episode-level signals

$$A_t^{\text{OR-GIGPO}} = \underbrace{\frac{1}{G} \sum_{i=1}^G \left(r_t^{(i)} + \gamma V(s_{t+1}^{(i)}) - V(s_t) \right)}_{\text{Step-level advantage}} + \underbrace{\frac{1}{N} \sum_{j=1}^N \left(R^{(j)} - b \right)}_{\text{Episode-level advantage}}$$

Enables effective multi-tool composition without full observation matching

Results: Strong Performance Across 13 Benchmarks

PERIA-8B vs. State-of-the-Art Models

Model	Size	In-Dist.	Out-of-Dist.	Avg.
Qwen3-VL-Thinking	8B	44.1	31.7	37.4
InternVL3.5	8B	39.2	28.5	33.4
VTool-R1	7B	41.8	29.3	35.0
Mini-o3	7B	42.5	30.8	36.2
PERIA-8B (Ours)	8B	54.1	36.1	44.4
<i>Δ vs. Backbone</i>		<i>+10.0</i>	<i>+4.4</i>	<i>+7.0</i>
Qwen3-VL-Thinking	235B-A22B	56.3	37.2	46.0
GPT-5	-	58.1	38.5	47.5

- **+10.0%** on in-distribution tasks
- **+4.4%** on out-of-distribution tasks

- **7.0%–14.8%** better than SOTA at similar size
- Approaches 235B models performance

Performance Across Task Families

Methods (↓)	In-Distribution Dataset							Out-of-Distribution Dataset							Overall
Datasets (→)	ReasonMap Plus	ReasonMap	VisualProbe Easy	VisualProbe Medium	VisualProbe Hard	MapTrace	Avg.	MapEval	BabyVision	Ball Tracing	Paper Folding	Cube 3-View Reasoning	Real-world Spatial Reasoning	VStar	Avg.
Qwen3-8B	65.7	12.3	54.6	31.3	26.6	74.1	44.1	54.0	13.0	17.8	11.0	21.2	27.7	77.5	31.7
PERIA-8B(Ours)	72.6 ^{+6.9}	27.1 ^{+14.8}	61.7 ^{+7.1}	41.2 ^{+9.9}	44.3 ^{+17.7}	77.7 ^{+3.6}	54.1 ^{+10.0}	55.5 ^{+1.5}	15.5 ^{+2.5}	18.8 ^{+1.0}	15.7 ^{+4.7}	36.3 ^{+15.1}	29.1 ^{+1.4}	82.2 ^{+4.7}	36.1 ^{+4.4}
Qwen3-235B	76.7	30.7	58.8	32.0	27.3	72.9	49.7	69.0	21.9	20.9	6.0	38.5	34.5	86.9	39.6
GPT-5	89.7	53.8	62.4	21.6	25.4	82.2	55.8	72.7	22.9	25.9	11.0	32.9	38.3	70.1	39.1

Strong results across diverse spatial reasoning tasks:

- **Map Reasoning:** ReasonMap, ReasonMap-Plus, MapTrace, MapEval
- **Visual Probing:** Multi-object tracking and localization
- **3D Reasoning:** Cube views, paper folding, real-world spatial
- **Vision Reconstruction:** Ball tracking, BabyVision, V*

Ablation Studies: What Matters?

Training Recipe:

- **w/o RL:** 40.4% → **Full:** 44.4%
 - RL adds **+4.0%** overall
 - Especially strong on ID tasks (+7.2%)
- **w/o SFT:** 40.7% (RL from scratch)
 - SFT establishes tool competence
 - RL refines multi-step reasoning

OR-GIGPO Components:

- **vs. PPO:** +2.8% improvement
- **vs. GRPO:** +1.5% improvement
- **Relaxed Obs:** Critical for grouping
- **Composite Adv:** Balances short/long term

Both SFT & RL stages
are essential

Qualitative Example: Multi-Step Tool Use

Task: Plan a subway route from Station A to Station D

- ① **Perception:** Use `map_text_ocr` to extract all station names
 - Result: Identified 15 stations and 3 transfer points
- ② **Interaction:** Use `draw_path` to trace potential routes
 - Result: Generated 2 candidate paths
- ③ **Reasoning:** Compare route lengths and transfer counts
 - Result: Route via Station B (1 transfer, shorter)
- ④ **Verification:** Use `bounding_box` to verify connections
 - Result: Confirmed all connections are valid

PERIA correctly solves the task through active tool composition

Key Contributions

- ❶ **PERIA Framework:** A modular tool-augmented visual agent for spatial reasoning
 - 18 perception & interaction tools
 - Unified POMDP formulation
- ❷ **OR-GIGPO Algorithm:** Novel RL method for multi-tool composition
 - Observation-relaxed grouping
 - Composite advantage estimation
- ❸ **Strong Empirical Results:** State-of-the-art on 13 benchmarks
 - +10.0% over backbone on in-distribution tasks
 - Approaches 30 $\times$ larger models
- ❹ **Open Resources:** Code, models, and datasets released
 - Reproducible training pipeline
 - Extensible tool sandbox

Limitations and Future Directions

Current Limitations:

- Tool sandbox currently focused on 2D spatial reasoning
- Limited to 18 tools (though easily extensible)
- OR-GIGPO requires significant compute for RL training

Future Directions:

- **Broader Tool Coverage:** 3D manipulation, video understanding
- **More Efficient Training:** Distillation, data efficiency improvements
- **Compositional Generalization:** Zero-shot tool composition
- **Multi-Modal Extensions:** Audio, text, embodied environments

Vision: General-purpose tool-augmented visual agents for any reasoning task

PERIA: Tool-Augmented Visual Agents for Spatial Reasoning

What we built:

- Modular tool sandbox
- SFT trajectory synthesis
- OR-GIGPO training algorithm
- 8B model $\approx$ 235B baseline

What we learned:

- Tools enable evidence acquisition
- SFT + RL both essential
- Relaxed grouping helps credit
- Strong generalization

Thank You!

Paper: <https://arxiv.org/abs/2606.12830>

Code: <https://github.com/antoinegg1/PEDIA>

Project Page: <https://antoinegg1.github.io/PEDIA/>