“Thinking with Images” 汇报 1

尹超

February 6, 2026

Table of Contents

Paper 1：VisualToolBench

- 定位：不是训练新模型，而是提出并评测 tool-enabled 的“think-with-images”基准
- 核心贡献
  - 任务设计覆盖单轮/多轮、多领域
  - Rubric-based 指标体系：可解释、可诊断（部分正确也能量化）
  - 标准化工具集合（图像处理、搜索、计算等）用于统一评测

VisualToolBench：数据与任务结构

数据规模

- 1,204 道开放式题（single-turn 603 / multi-turn 601）
- 2,893 张图像
- 7,777 条 rubrics
- 5 大领域均衡：STEM/Medical/Finance/Sports/Generalist

任务结构

- Single-turn：
  - Region Switch Q&A
  - Hybrid Tool-use
- Multi-turn：
  - Follow-up Test
  - Temporal Reasoning
  - Progressive Reasoning

VisualToolBench：数据

1 Paper 1: VisualToolBench

Statistic	Number
Total questions	1,204
- STEM	238 (19.7%)
- Medical	238 (19.7%)
- Finance	243 (20.2%)
- Sports	240 (20.0%)
- Generalist	245 (20.4%)
Single-turn	603 (50.1%)
- Region Switch Q&A	281 (46.6%)
- Hybrid Tool-use	322 (53.4%)
Multi-turn	601 (49.9%)
- Follow-up Test	198 (32.9%)
- Temporal Reasoning	205 (34.2%)
- Progressive Reasoning	198 (32.9%)
Total number of rubrics	7,777
Total number of images	2,893
Average prompt length	48.41
Average answer length	128.93

Table 2: Statistics of VISUALTOOLBENCH.

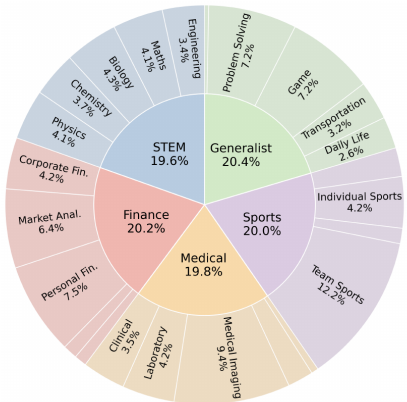

Figure 1: Topic distribution.

VisualToolBench：评测指标（Rubric 体系）

Rubric 维度（5 类）

Visual Understanding / Truthfulness / Instruction Following / Reasoning / Presentation

计分方式（关键要点）

- 每题包含多条 rubric，含权重 $w \in \{1, 2, 3, 4, 5\}$
- Judge 判断每条 rubric 是否满足（0/1），按权重加权归一化得到任务分数
- 可派生 Pass/Fail（关键 rubric 决定）：更接近“能不能用”的产品指标

对我们：怎么用

- rubric score 用于 诊断：错在视觉/真实性/遵循/推理/表达哪一类
- pass/fail 用于 可用性门槛：上线前压测更友好

VisualToolBench： Rubric 评测示意

1 Paper 1: VisualToolBench

Prompt, Image, and Ground Truth

Prompt: How much would it cost to get a gluten-free 14” pizza with provolone cheese, zesty red sauce, and kalamata olives?

Ground Truth: From the menu, the price for one topping 14” pizza is $13. Also $3 should be added for the gluten-free crust. Hence the final price is $13 + $3 = $16.

Key Visual Details

Model Responses via Tool-Use

✦ Gemini-2.5-pro

Tool-Uses: python image process → calculator → ···

Final Response: A 14-inch pizza with one topping is $14, and a gluten-free crust for a 14-inch pizza is an additional $3. The total cost for a gluten-free 14” pizza with provolone cheese, zesty red sauce, and kalamata olives would be $17.

Response Evaluation

Rubric	Weight	Grade
Rubric 1 The model recognizes that the pizza described in the prompt is a one-topping pizza.	3	Yes (+3)
Rubric 2 (Critical Rubric) The model identifies the price of a one-topping 14” pizza to be $13.	4	No (+0)
Rubric 3 The model adds $3 for the gluten-free crust.	3	Yes (+3)
Rubric 4 (Critical Rubric) The model reports $16 as the price of the pizza.	5	No (+0)
Rubric 5 The model explains how it calculated its final price.	2	Yes (+2)

Rubric Score: (3+3+2)/(3+4+3+5+2) = **0.47**

Pass/Fail: **Fail** (critical rubric fails)

Table of Contents

Paper 2： V-Thinker

- 定位：训练一个通用的 interactive thinking with images 模型
- 两大贡献
  - Data Evolution Flywheel：合成并校验可执行的交互式数据（逐步改图/标注/推理）
  - VTBench：分三层任务的评测集（感知 → 交互 → 推理）

V-Thinker：训练数据与 base model

SFT 数据

- V-Perception-40K（感知对齐）
- V-Interaction-400K（交互对齐）

RL 采样

- 从 We-Math2.0 / MMK12 / ThinkLite / V-Interaction-400K 采样 40K

Base / 对照

- 主要对照基线围绕 Qwen2.5-VL-7B 展开（结果以 Δ 展示）

开源

- GitHub：We-Math/V-Thinker

The Data Evolution Flywheel framework

2 Paper 2: V-Thinker

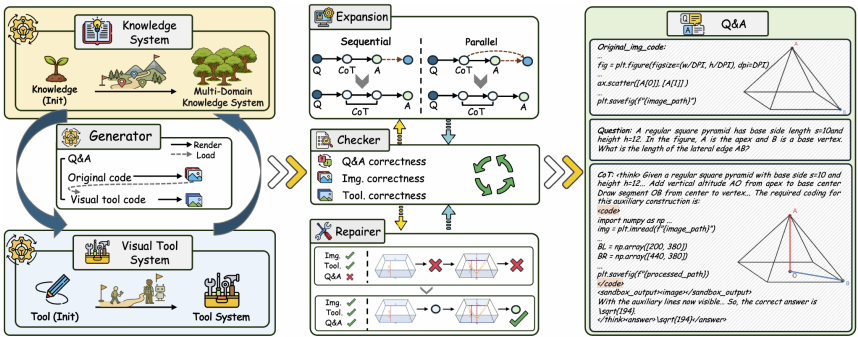

Figure 4. The Data Evolution Flywheel framework: **Left:** knowledge-driven evolution mechanism. **Middle:** coordinated calibration and progressive expansion stages. **Right:** representative synthetic QA instances generated through the flywheel.

Figure: Data Evolution Flywheel 框架

Performance on VTBench and general reasoning

Method	VTBench				General Reasoning			
	Perception	*Instruct. Interaction*	*Interactive Reasoning*	*Avg.*	*MathVision Acc.*	*We-Math Acc.*	*VisuLogic Acc.*	*Avg.*
GPT-4o	12.6	26.0	36.4	25.0	43.8	68.8	26.3	46.3
InternVL3-78B	13.8	19.0	43.4	25.4	43.1	64.2	27.7	45.0
InternVL3-8B	10.4	6.8	33.8	17.0	29.3	58.8	24.9	37.7
LLaVA-OV-1.5-8B	12.2	12.2	30.2	18.2	25.6	56.7	23.7	35.3
InternVL3-2B	3.0	3.4	22.0	9.5	23.3	41.7	24.3	29.8
Qwen2.5-VL-7B	12.6	8.8	31.8	17.7	23.0	61.7	26.0	36.9
V-Thinker-7B	**18.6**	**31.6**	**40.4**	**30.2**	**29.3**	**62.8**	**26.6**	**39.6**
Δ (*vs* Qwen2.5-VL-7B)	***+6.0***	***+22.8***	***+8.6***	***+12.5***	***+6.3***	***+1.1***	***+0.6***	***+2.7***

Table 1. Overall performance on VTBench (left) and general reasoning (right). (*Instruct. Interaction* denotes *Instruction-Guided Interaction*.)

Figure: Overall performance on VTBench and general reasoning

VTBench：任务 → 输出 → 判分

2 Paper 2: V-Thinker

VTBench 三层任务（每层 500，合计 1500 QA）

- Perception：细粒度定位/坐标
- Instruction-Guided Interaction：按指令画线/标注/改图
- Interactive Reasoning：交互支撑的推理问答

输出与判分口径

- Perception：输出 Python code 画点/坐标；执行后图像 vs 标注图，LMM-as-judge
- Interaction：输出 Python code 执行交互；输出图 vs 标注图，LMM-as-judge
- Reasoning：输出 final answer；LLM-as-judge 判正确

Table of Contents

3 Paper 3: DeepEyes

Paper 3：DeepEyes

- **定位**：用强化学习直接激励“thinking with images”
- **核心思路**：用最小工具闭环（zoom-in）把“看不清”变成可学习动作
- **训练算法**：GRPO（group relative policy optimization）

意义

主动感知 (Active Perception)：模型自己决定“看哪里”，而不是被动接受固定输入。
多步规划 (Multi-step Planning)：将复杂问题拆解为定位、放大、对比、判断四个阶段。
工具使用 (Tool Use)：将“放大/裁剪”作为一种外部工具来增强自身的感知上限。

DeepEyes：方法概览

3 Paper 3: DeepEyes

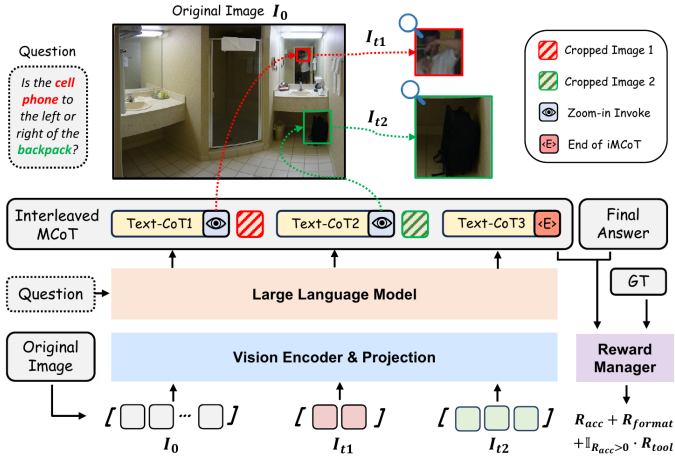

Figure 2: **Overview of *DeepEyes*.** Our model itself decides whether to perform a second perception via zoom-in by generating grounding coordinates and cropping relevant regions, or to answer directly.

DeepEyes：训练数据与 Base model

3 Paper 3: DeepEyes

训练数据三块

- Fine-grained：V* training set（小目标/细字/局部细节）
- Chart：ArxivQA（图表/曲线等结构化视觉）
- Reason：ThinkLite-VL（推理广度）

Base / 对照

- 明确围绕 Qwen2.5-VL-7B 做数据选择与对照

开源

- GitHub：Visual-Agent/DeepEyes

DeepEyes：Reward

三项 reward 组成（工程要点）

- Accuracy reward：最终答案正确性
- Formatting reward：输出格式可解析（防止 policy 崩坏）
- Conditional tool bonus：仅当答案正确且至少使用一次工具才给 bonus

$$R(\tau) = R_{\text{acc}}(\tau) + R_{\text{format}}(\tau) + \mathbb{I}_{R_{\text{acc}}(\tau)>0} \cdot R_{\text{tool}}(\tau)$$

为什么重要

- 防止“乱用工具”刷分；鼓励“正确且有效 的工具调用”
- 特别适合做：工具策略是否真实提升成功率 的验证闭环

DeepEyes：评测覆盖面（用于综合回归）

3 Paper 3: DeepEyes

高分辨/细粒度

V*Bench、HR-Bench（4K/8K）——验证 zoom-in 是否提升细节感知

Table 1: **Results on High-Resolution Benchmarks.** E2E indicates whether the model is end-to-end, requiring no manually defined workflow. * denotes the results are reproduced by ourselves.

Model	E2E	Param Size	V^* Bench [41]			HR-Bench 4K [59]			HR-Bench 8K [59]		
			Attr	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	Overall
GPT-4o [60]	✓	-	-	-	66.0	70.0	48.0	59.0	62.0	49.0	55.5
o3 [8]	✓	-	-	-	95.7	-	-	-	-	-	-
SEAL [41]	✗	7B	74.8	76.3	75.4	-	-	-	-	-	-
DyFo [44]	✗	7B	80.0	82.9	81.2	-	-	-	-	-	-
ZoomEye [61]	✗	7B	93.9	85.5	90.6	84.3	55.0	69.6	88.5	50.0	69.3
LLaVA-OneVision [62]	✓	7B	75.7	75.0	75.4	72.0	54.0	63.0	67.3	52.3	59.8
Qwen2.5-VL* [58]	✓	7B	73.9	67.1	71.2	85.2	52.2	68.8	78.8	51.8	65.3
Qwen2.5-VL* [58]	✓	32B	87.8	88.1	87.9	89.8	58.0	73.9	84.5	56.3	70.4
DeepEyes	✓	7B	91.3	88.2	90.1	91.3	59.0	75.1	86.8	58.5	72.6
Δ (*vs* Qwen2.5-VL 7B)	-	-	+17.4	+21.1	+18.9	+6.1	+6.8	+6.3	+10.0	+6.8	+7.3

Figure: 高分辨率基准测试结果

DeepEyes：评测覆盖面（用于综合回归）

3 Paper 3: DeepEyes

Grounding / Hallucination

refCOCO/refCOCO+/refCOCOg、ReasonSeg、POPE ——验证定位与幻觉抑制

Table 2: **Results on Grounding and Hallucination Benchmarks.** We compare *DeepEyes* with open-source MLLMs on several grounding and hallucination benchmarks. * denotes the results are reproduced by ourselves.

Model	Param Size	refCOCO	refCOCO+	refCOCOg	ReasonSeg	POPE			
						Adversarial	Popular	Random	Overall
LLaVA-OneVision [62]	7B	-	-	-	-	-	-	-	88.4
Qwen2.5-VL [58]	7B	90.0	84.2	87.2	-	-	-	-	-
Qwen2.5-VL* [58]	7B	89.1	82.6	86.1	68.3	85.9	86.5	87.2	85.9
DeepEyes	7B	89.8	83.6	86.7	68.6	84.0	87.5	91.8	87.7
Δ (*vs* Qwen2.5-VL 7B)	-	+0.7	+1.0	+0.6	+0.3	-1.9	+1.0	+4.6	+1.8

Figure: 在定位（Grounding）与幻觉（Hallucination）基准测试上的结果

DeepEyes：评测覆盖面（用于综合回归）

3 Paper 3: DeepEyes

推理任务

MathVista/MathVision/MathVerse/We-Math 等——验证泛化到推理 QA

Table 3: **Results on Multimodal Reasoning Benchmarks.** We evaluate our model on several multimodal reasoning benchmarks. * denotes the results are reproduced by ourselves, and † represents the results are copied from [63].

Model	Param Size	Math Vista [64]	Math Verse [65]	Math Vision [66]	We Math [67]	Dyna Math [68]	Logic Vista [69]
LLaVA-OneVision [62]	7B	58.6†	19.3†	18.3†	20.9†	-	33.3†
Qwen2.5-VL [58]	7B	68.2	49.2	25.1	35.2†	-	44.1†
Qwen2.5-VL* [58]	7B	68.3	45.6	25.6	34.6	53.3	45.9
DeepEyes	7B	70.1	47.3	26.6	38.9	55.0	47.7
Δ (*vs* Qwen2.5-VL 7B)	-	+1.9	+1.7	+1.0	+4.3	+1.7	+1.8

Figure: 多模态推理基准测试结果

Table of Contents

Paper 4：SAM-R1

4 Paper 4: SAM-R1

- 定位：多模态分割 / 指代分割（referring segmentation）的 RL 对齐
- 核心范式：MLLM 先输出 bbox/point 等中间表示，再交给 SAM2-Large 生成 mask
- 关键贡献：用 SAM 的 mask 质量构造 reward，把分割质量显式反馈给 MLLM

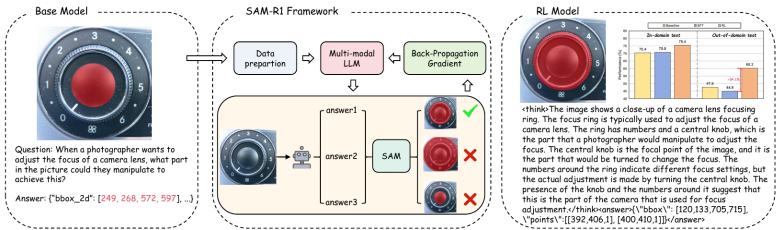

Figure: SAM-R1 方法概览

SAM-R1：方法概览

4 Paper 4: SAM-R1

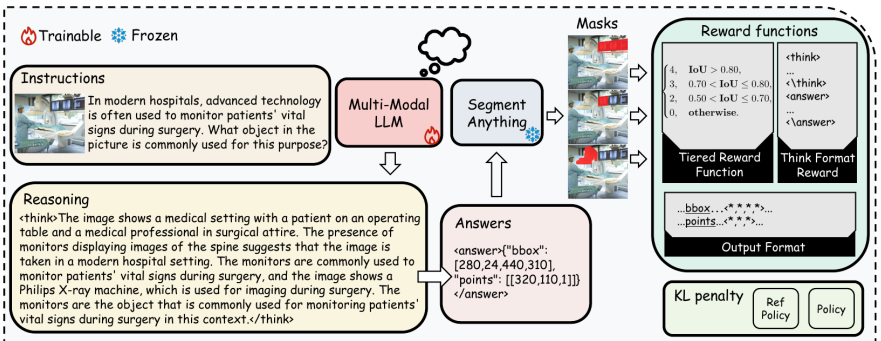

Figure 2: Our framework integrates the Segment Anything Model (SAM) as a reward provider in the reinforcement learning training of a multimodal large model (MLLM). The two models jointly process user-input questions and images to identify target objects and generate masks. Specifically, the MLLM generates the reasoning process and answer, then passes them to SAM. A fine-grained reward based on Intersection over Union (IoU) is calculated to optimize the MLLM.

SAM-R1：训练数据 / Base model / 输出格式

4 Paper 4: SAM-R1

训练数据

- RefCOCOg train 随机采样 3,000
- In-domain：RefCOCOg test
- OOD：RefCOCO testA、RefCOCO+ testA
- 额外：ReasonSeg（zero-shot）

Base

- MLLM：Qwen2.5-VL-7B
- Segmenter：SAM2-Large

输出格式（必须对齐）

- bbox + reference point + textual flag（结构化）
- 再由 SAM 产 mask

SAM-R1：指标

4 Paper 4: SAM-R1

评测指标

- gIoU：per-image IoU 平均
- cIoU：累计 intersection / 累计 union

Table 2: Performance comparison on referring expression benchmarks using cIoU.

Method	refCOCO	refCOCO+	refCOCOg
LAVT [47]	75.8	68.4	62.1
ReLA [19]	76.5	71.0	66.0
LISA-7B [16]	76.5	67.4	68.5
PixelLM-7B [31]	76.5	71.7	70.5
PerceptionGPT-7B [28]	78.6	73.9	71.7
Seg-Zero-7B* [23]	**79.2**	73.9	**73.3**
SAM-R1 (Ours)	**79.2**	**74.7**	73.1

Figure: 在 RefCOCOg/RefCOCO/RefCOCO+ 上的评测结果

SAM-R1：指标

4 Paper 4: SAM-R1

训练 reward（关键）

- 将（bbox+point）交给 SAM 得到 mask，与 GT mask 算 IoU
- 按 IoU 分段给 reward（例如 $> 0.8, 0.7 \sim 0.8, 0.5 \sim 0.7$, else）

为什么重要

把“分割质量”变成可学习信号，避免只在语言空间里对齐。

Table of Contents

5 Cross-paper Comparison

对齐表：数据集/Benchmark → 任务 → 输出 → 判分

5 Cross-paper Comparison

论文	数据集/Benchmark	任务类型 & 输出格式	判分方式/指标
VisualToolBench	VISUALTOOLBENCH	Tool-enabled 单/多轮；输出：最终自然语言 answer（可含工具轨迹）	Rubric（0/1）+ weight 加权；LLM-as-judge 逐条 rubric 判定
V-Thinker	VTBench（1500 QA，3 类任务）	Perception：输出 Python code（画点/坐标）；Interaction：输出 Python code（执行交互）；Reasoning：输出 final answer	Perception/Interaction：执行图 vs 标注图，LMM-as-judge；Reasoning：LLM-as-judge
DeepEyes	V* / HR-Bench / ref-COCO* / POPE / ReasonSeg / Math*	交互式推理（zoom-in）；输出：bbox_2d 工具调用（可多次）+ 最终 answer	多按各 benchmark 官方协议；覆盖高分辨、grounding、hallucination 等
SAM-R1	RefCOCOg / RefCOCO / RefCOCO+ / ReasonSeg	指代分割；输出：bbox + point + flag（结构化）并交给 SAM 产 mask	评测：gIoU/cIoU；训练 reward：IoU 分段（SAM mask vs GT）