Thinking with Images: A New Paradigm of Visual-Language Model Reasoning

Presenter: Qunzhong WANG

2025.11.13

Contents

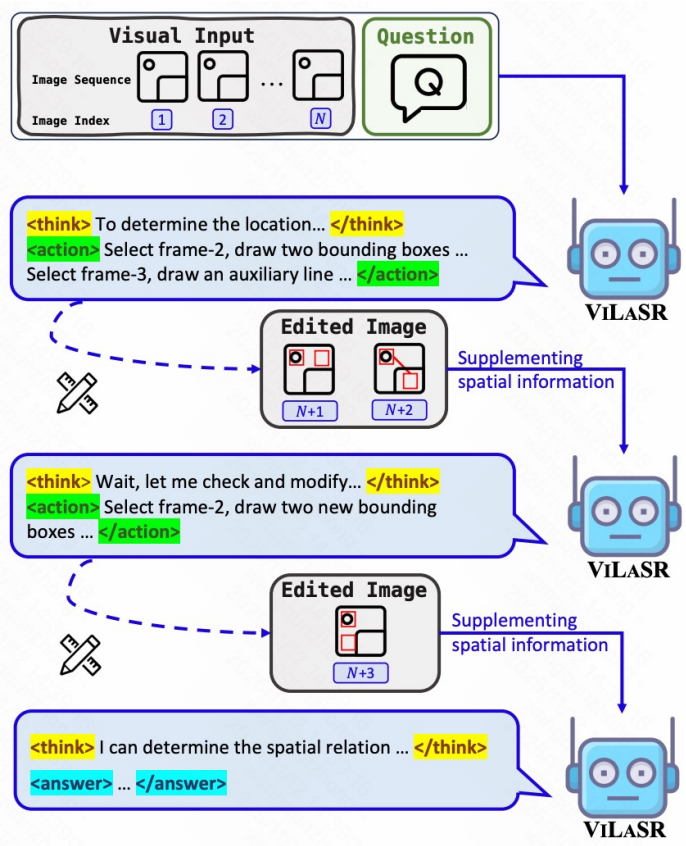
Motivation and Timeline

Think with Image is essentially an ***agentic reasoning approach*** for visual understanding. Through multi-round interactions, the model progressively gathers and refines visual evidence.

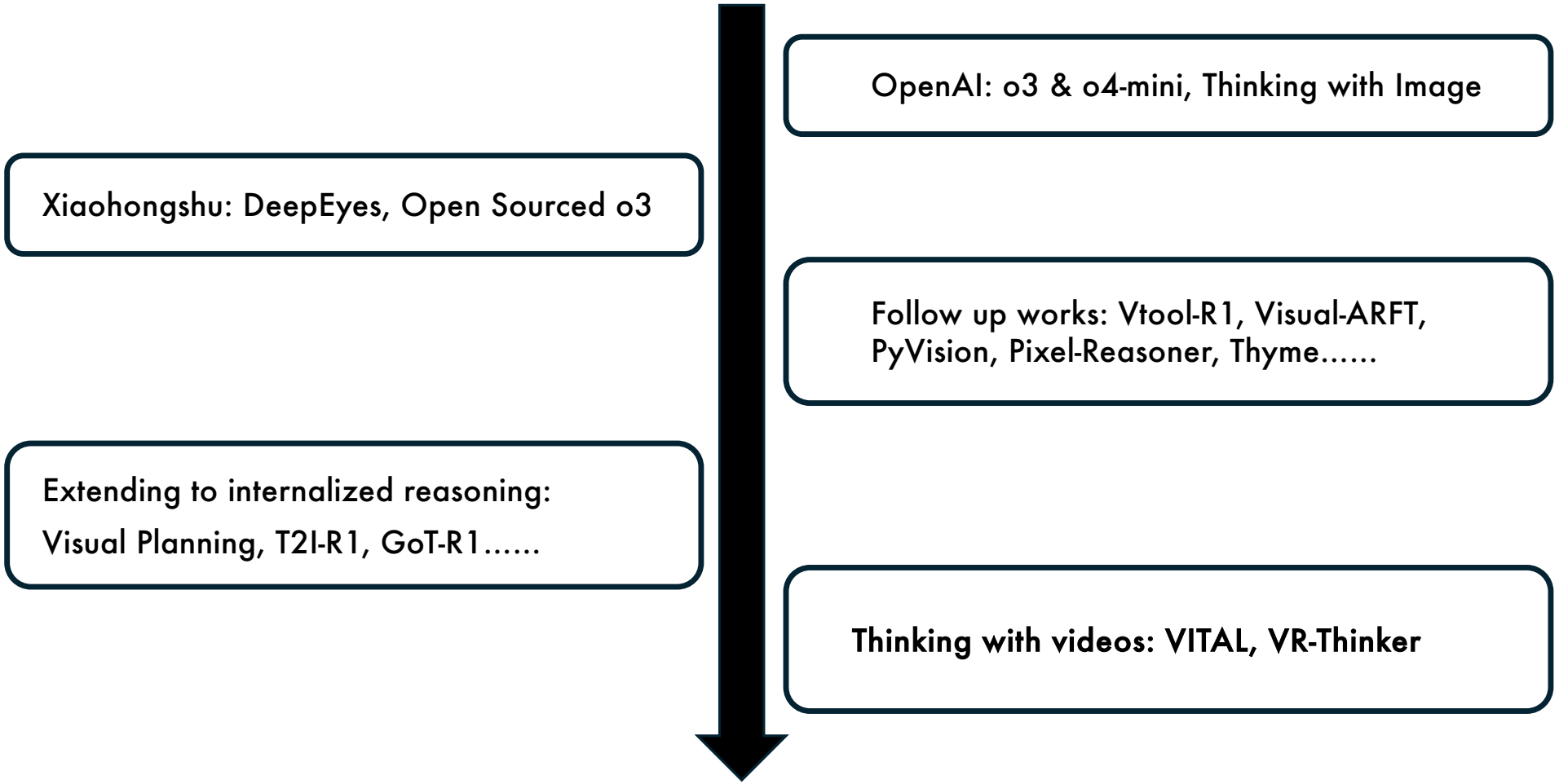
- VLM is based on ViTs which encode an image into embeddings, align them to a language space and then hand them over to LLMs for reasoning.
- This relies on the ***assumption*** that:

  ViT can losslessly align visual information to linguistic representations, which is not true actually.

To raise the model's reasoning ceiling, we do need **dynamically manipulate images**



Motivation and Timeline



GPT o3

OpenAI o3 和 o4-mini是最新的 o 系列视觉推理模型。这是我们的模型第一次能够在思维链中结合图像进行思考，而不仅仅是看到图像。

与我们早期的 OpenAI o1 模型类似，o3 和 o4-mini 在回答问题前进行长时间的思考，并在回答用户问题前调用其余的内部思维链。通过在链中思维结合图像进行思考，o3 和 o4-mini进一步扩展了这一功能。图像思考通过使用工具转换用户上传的图像来实现，除了其他简单的图像处理技术之外，还允许用户对图像进行裁切、放大和旋转。更重要的是，这些功能都是内在的，依赖于单独的专用模型。

从Openai官网摘取

Openai并没有透露过多技术细节，但我们可以看到o3提供了一系列图像处理工具并且支持编程式图像分析。这两点也是后续开源工作的重点。

Thinking with images, Tool invocation

Thinking with Images with **Tool Invocation**, in essence, define a set of tools {T1, T2, ..., Tk}, allowing the model in each reasoning round to freely choose either to output an answer or to invoke any of the tools in the set.

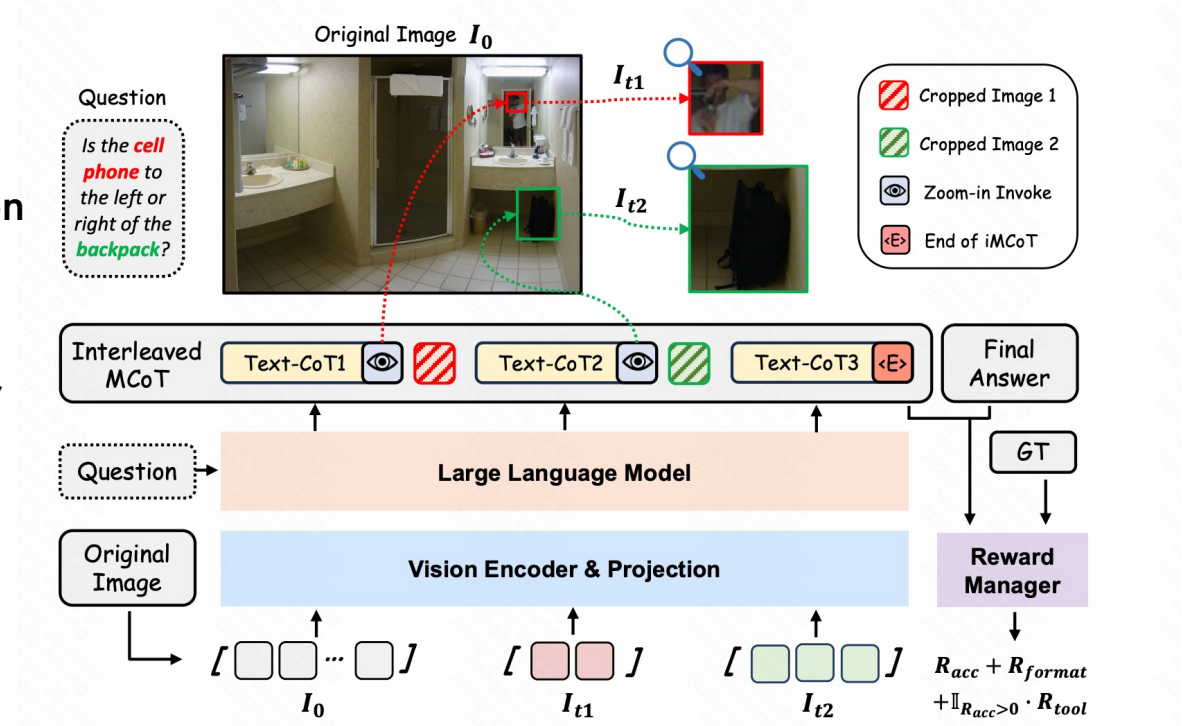
In visual reasoning, the model's choices are relatively stable and consistent, making its execution more reliable and training more manageable.

- DeepEyes
- Pixel Reasoner
- Vtool-R1
- VAT
- Chain-of-Focus
- VisTA……

DeepEyes

DeepEyes, the first open-source replication of o3

- Define the core and the only tool: Zoom-In. With format {"bbox_2d": [x1, x2, y1, y2]}"
- Define such fashion of multi-turn reasoning Interleaved Multi-modal Chain-of-Thought (iMCoT)



训练模式：

end-to-end reinforcement learning without cold-start/sft

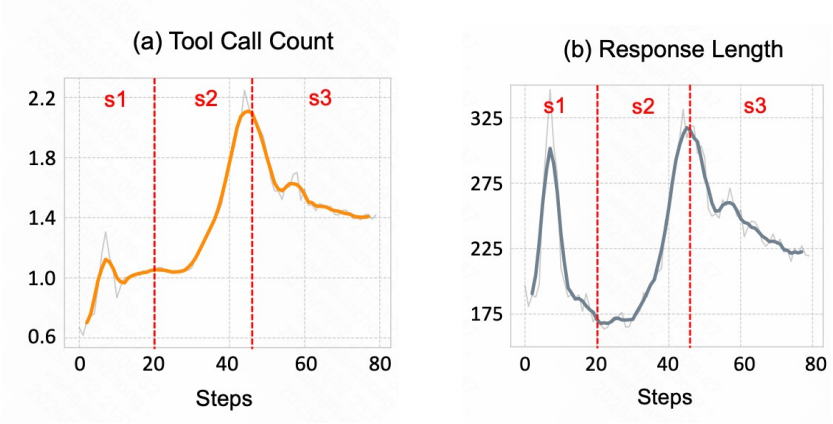
奖励设计：

R_tool other than R_acc & R_fmt **只有当最终答案正确时（R_acc > 0），并且在推理过程中至少使用了一次工具**，模型才会获得这个额外的奖励，鼓励有效调用

数据筛选：

面向工具使用的数据选择策略，筛选出那些通过单次观察会答错，但如果给定了正确的局部放大图就能答对的样本

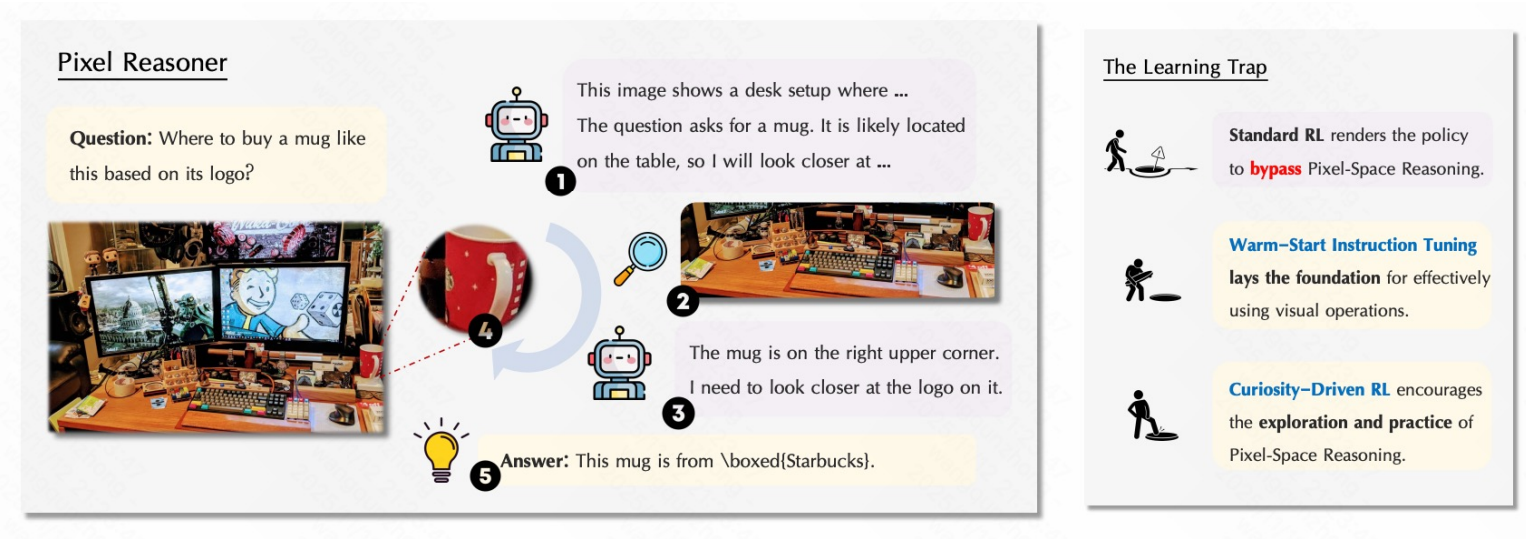
DeepEyes



Model	E2E	Param Size	V^* Bench [41] Attr	V^* Bench [41] Spatial	V^* Bench [41] Overall	HR-Bench 4K [59] FSP	HR-Bench 4K [59] FCP	HR-Bench 4K [59] Overall	HR-Bench 8K [59] FSP	HR-Bench 8K [59] FCP	HR-Bench 8K [59] 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

- Archive Huge gain on High-Resolution Benchmarks(Where target objects referred to in the questions are often quite small in these images)
- Tool call count experience going up and down
- Resp Length experience going up and down twice

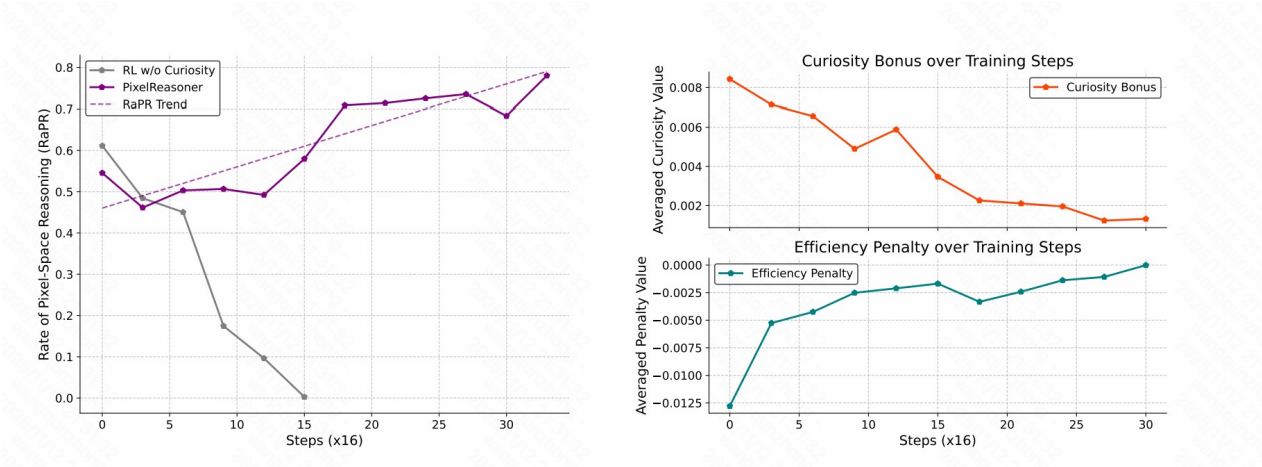
Pixel Reasoner



Pixel Reasoner first unify 2 kinds of task: image & Video

- Define 2 kinds of tools: Zoom-In and Select Frames for image and videos respectively
- Define such fashion of multi-turn reasoning Pixel-Space Reasoning

Pixel Reasoner



训练模式:

1. Warm-Start Instruction Tuning to learn the reasoning format
2. Curiosity-Driven Reinforcement Learning to incentivize Pixel-Space Reasoning

特殊发现:

learning trap: model downgrade to text reasoning due to stronger text CoT ability

奖励设计:

R_curiosity & R_penalty to keep the rate of pixel-space reasoning exceed a threshold H, while the number of visual operations not exceed a bound N

数据构建:

Distillation from GPT4o while inserting erroneous reasoning segments to boost self-correction

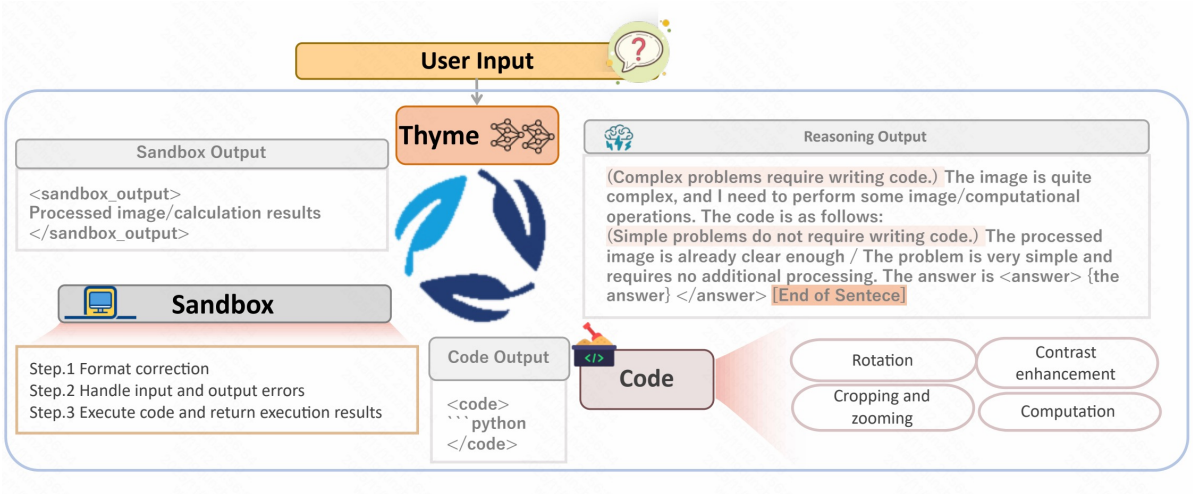
Thinking with images, Programmatic manipulation

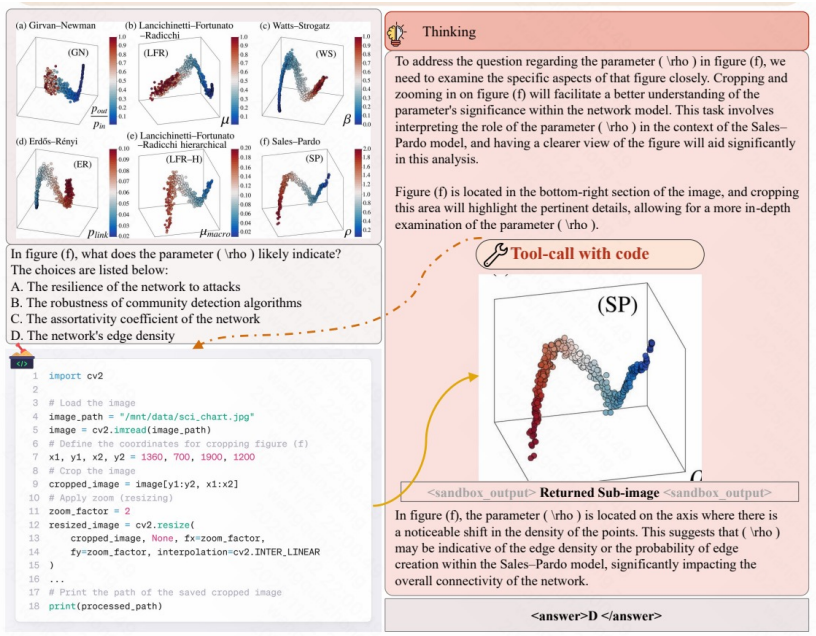
Thinking with Images, Programmatic Manipulation, in essence, allows the model in each reasoning round to freely choose either to output an answer or to generate a piece of code that performs operations on the image.

This approach imposes higher demands on the model's code generation ability, accuracy, and robustness.

- Thyme
- MathCoder-VL
- Visual-ARFT
- VRAG-RL
- CoSyn……

Thyme



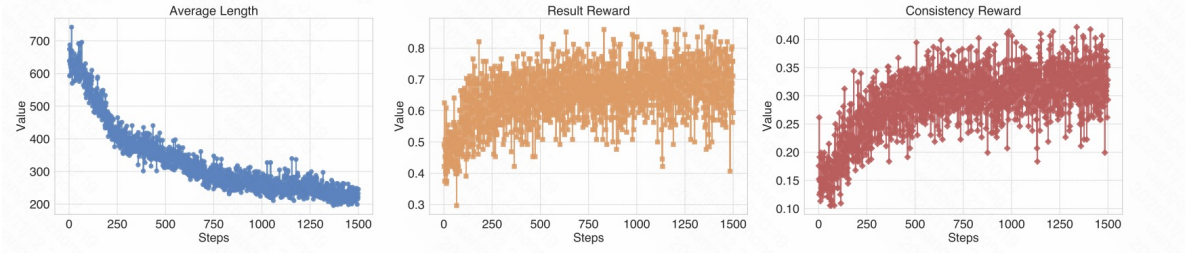


Thyme ask the model to generate executable code to manipulate the input image

- Code is executed in the Sandbox environment, where common errors are handled to reduce model's burden
- "超越生成图像思维(Thinking with Generated Images), 超越裁剪思维(Thinking with Cropping) "

Thyme

Training Data	Benchmark *Split*	Hallucination *Overall*	MME-Realworld-Lite *Perception*	 *Reasoning*	V* *Overall*	HRBench 8K *Overall*	MathVista *mini*	RealWorld QA *Overall*	Avg
			Baselines						
-	Qwen 2.5 VL 7B	48.29	48.75	37.73	76.40	65.50	68.20	68.20	59.01
Thyme-SFT Data	Thyme-SFT	53.74	51.75	45.40	79.58	65.12	68.70	69.80	62.01
			Reward Design						
Thyme-RL Data	Outcome+Format	58.20	53.70	45.60	80.10	70.75	67.40	69.41	63.59
	+ Consistency	56.66	58.25	49.06	81.76	72.25	69.70	72.06	65.68
	+ Process Reward	55.63	52.95	44.60	80.10	72.25	67.50	67.32	62.91
	+ Code Reward	52.18	56.80	49.86	82.72	70.87	69.10	70.28	64.54



训练模式:

- Thyme SFT + Thyme RL双阶段
- SFT: 1. 沙箱内容屏蔽, 遮蔽外界响应对应的logp 2. 只训练最后一轮, 避免模型学习先犯错再改正
- RL: GRPO-ATS：**代码生成**将采样**温度（temperature）设为 0.0，文本推理**：将**温度设为 1.0**
- RL: 奖励函数设计：R_acc, R_fmt, R_consistency. 用**一致性奖励**评估最终答案是否与推理过程逻辑一致
  - Final Reward = Result Reward × (1 + 0.5 × Consistency Reward) + 0.5 × Formatting Reward

数据构建

- SFT:从400万的原始数据，通过流水线模式构建SFT数据集, 包括无需Code数据，真实世界Thyme数据，手动构建Thyme数据，多轮对话纠错数据
- RL:一部分来自开源数据集。另一部分是**手动收集和标注的1万张高分辨率、高难度复杂图像**

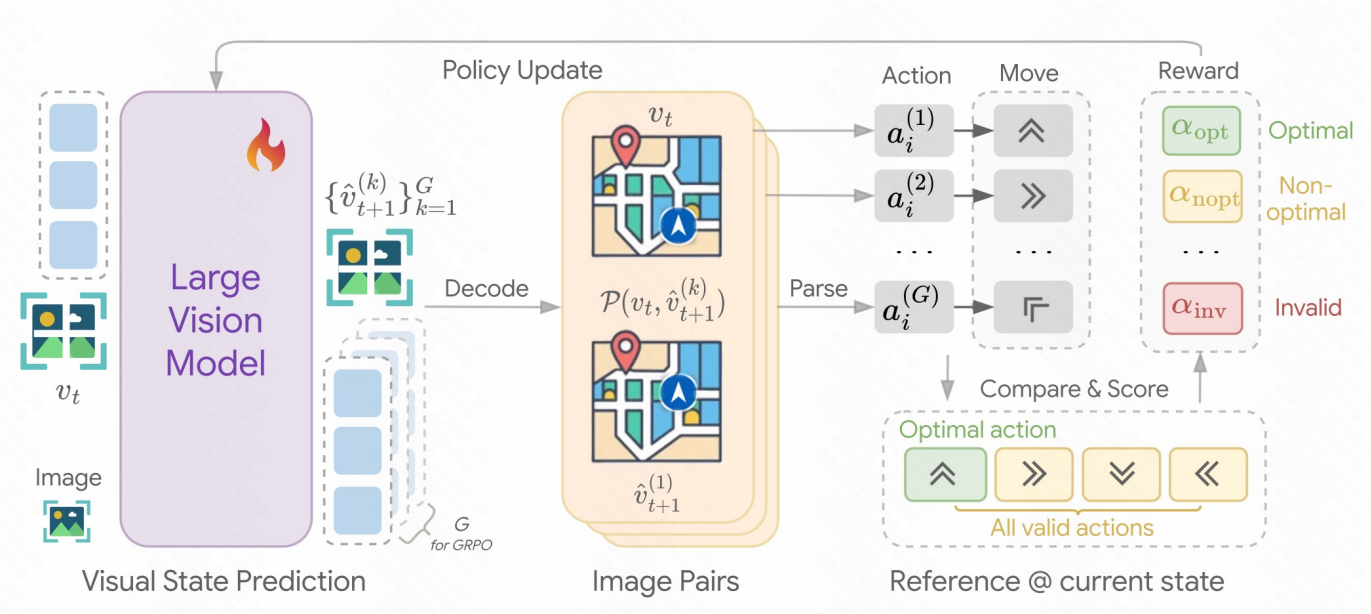
Thinking with images, internalized reasoning

Thinking with Images: Internal Reasoning (Generative Thinking), in essence, involves choosing a hybrid model capable of autoregressively generating both text and images, or a purely image-autoregressive model.

It no longer relies on any external tools or execution environments; instead, it leverages a unified internal architecture to directly generate new visual information—such as images or feature maps—as intermediate steps in the reasoning process.

- VPRL
- Chameleon
- Emu2
- CoT-VLA……

Visual Planning



Visual Planning is based on LVM-7B, rather than any VLM, which is a novel modeling approach which enables learning a Large Vision Model without making use of any linguistic data, by defining a common format, visual sentences.

Visual Planning

Model	Input	Output	FROZENLAKE		MAZE		MINIBEHAVIOR		AVG.	
			EM (%)	PR (%)	EM (%)	PR (%)	EM (%)	PR (%)	EM (%)	PR (%)
Closed-Source Model										
Gemini 2.0 Flash										
- Direct	A+🖼	A	21.2	47.6	8.3	31.4	0.7	29.8	10.1	36.3
- CoT	A+🖼	A	27.6	52.5	6.9	29.8	4.0	31.2	12.8	37.8
Gemini 2.5 Pro (*think*)	A+🖼	A	72.0	85.0	21.5	35.5	37.6	59.9	43.7	60.1
Open-Source Model										
Qwen 2.5-VL-Instruct-7B										
- Direct	A+🖼	A	1.2	15.0	0.6	14.5	0.3	9.8	0.7	13.1
- CoT	A+🖼	A	8.2	29.1	2.3	15.2	0.5	14.7	3.7	19.7
- SFT†	A+🖼	A	68.6	84.4	60.9	70.3	31.3	56.1	53.6	69.9
LVM-7B										
- VPFT† (ours)	🖼	🖼	75.4	79.5	59.0	64.0	33.8	52.2	56.1	65.2
- **VPRL**† (ours)	🖼	🖼	**91.6**	**93.2**	**74.5**	**77.6**	**75.8**	**83.8**	**80.6**	**84.9**

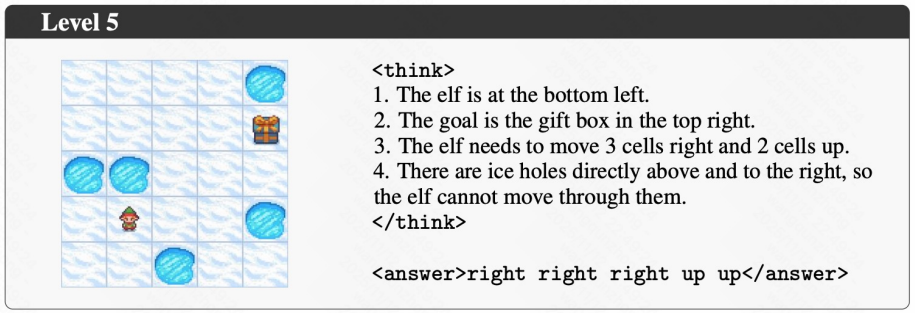
训练阶段

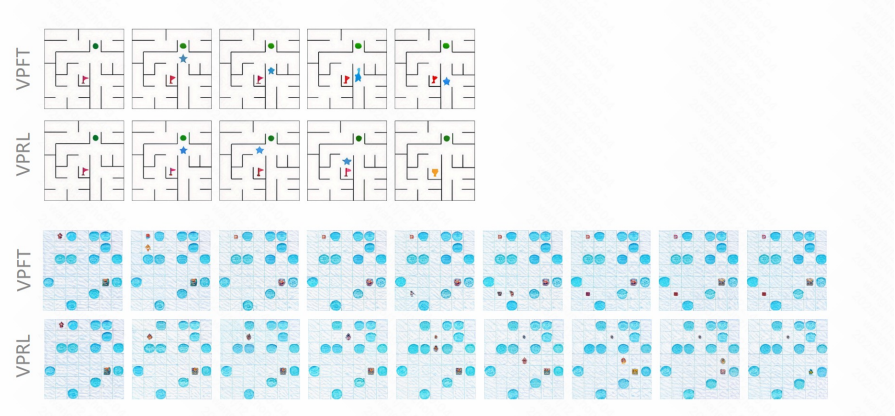
- Policy Initialization 实现热身与探索
- RL for Visual Planning 学习制定**朝向最终目标的有效规划**
  - 对于一个给定的当前状态图像，让第一阶段产出的LVM模型生成一组候选的下一步图像

奖励设置

- 对每一个生成的候选图像进行打分
  - **动态解释器 (Dynamics Interpreter)**：分析图像的变化，判断这个变化代表了什么**动作**（如“left”）,以及这个动作是否**有效**（如是否撞墙）
  - **进展评估器 (Progress Estimator)**：判断新的状态是否比当前状态**更接近目标**
- 分配**奖励:**有效且更接近目标 **+1,** 有效但未更接近目标 **0 无效动作 (**违反规则**) -5 (**重罚**)**

Visual Planning





模型训练完成后，解决任务的流程如下：

1.**输入**：给定一个来自未知任务的**初始状态图像**（例如，一个新的迷宫起点图）。

2.**自回归生成**

1. 将初始图像输入到训练好的VPRL模型中。
2. 模型生成最有可能的**第一个规划步骤图像** $\hat{v}1$。
3. 将 (初始图像, $\hat{v}1$) 作为新的输入，模型生成第二个规划步骤图像 $\hat{v}2$。
4. 直到模型生成一个完整的图像轨迹。

Thinking with Videos

Thinking with Videos, in essence, applies techniques similar to *Thinking with Images*—such as tool invocation and programmatic manipulation—to process videos and update visual evidence.

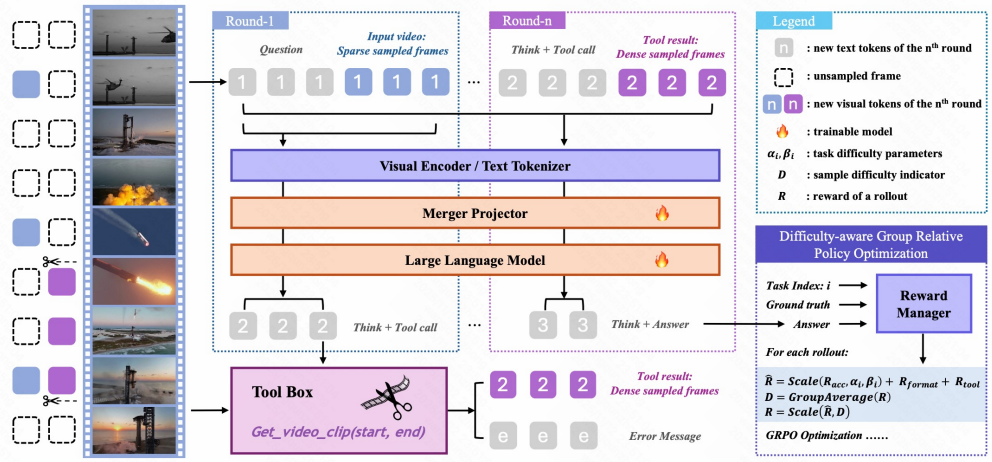
This approach helps alleviate issues where models tend to "fill in" details that do not actually exist in the video or mistakenly recall the temporal order of events.

- VITAL
- VR-Thinker……

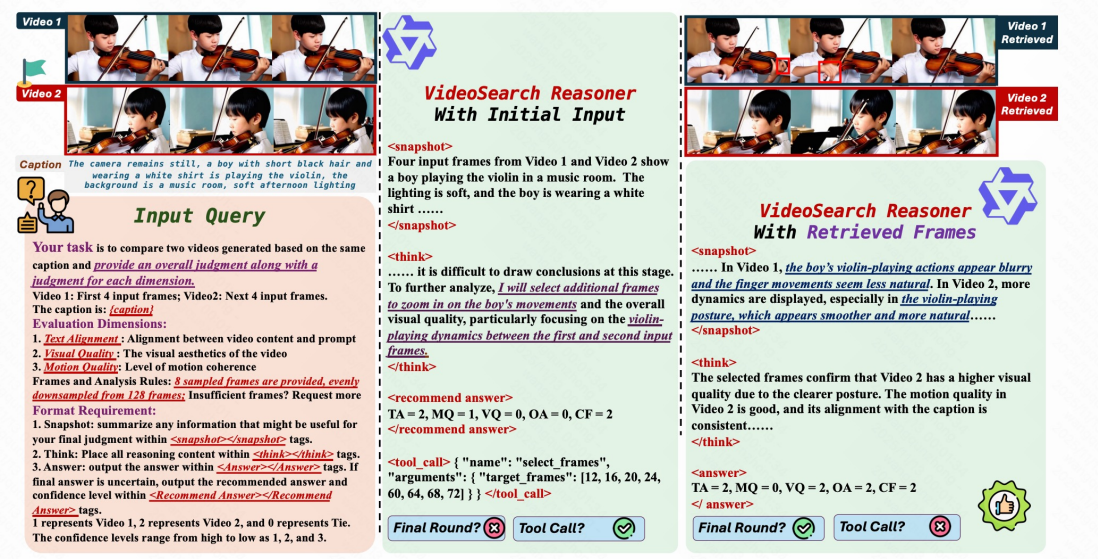
Video Intelligence via Tool-Augmented Learning

- VITAL 超越纯文本推理，引入了**Visual Toolbox**
- 模型对某个时间段记忆模糊或需要更多细节，可以主动调用工具：“视频裁剪工具”，获取该时间段内更密集的视频帧。
- 这些新获取的视觉信息会被整合到下一步的思考中，形成一个**文本思考 → 调用工具 → 观察新视觉证据 → 继续文本思考 → 得出答案**的闭环

- 数据集覆盖了**时间定位 (Temporal Grounding)**、**视频问答 (VQA)** 和 **带定位的问答 (Grounded VQA)** 三种任务
- **MTVR-CoT-72k**：用于监督微调（SFT），让模型初步学会多模态思维链的格式和基本推理能力。
- **MTVR-RL-110k**：用于强化学习（RL），通过奖励机制进一步优化模型的决策和推理能力。



VR-Thinker



- 超越了传统奖励模型（RM）将视觉信息一次性、静态输入的方式
- 模型在推理中若发现初始帧信息不足或细节模糊，可以主动调用视觉推理工具，**select_frame**，来**主动获取或回顾**视频中的任意帧。
- 通过滑动窗口记忆（**Window Memory**）机制，模型只保留最近的视觉信息，从而在有限的上下文长度内实现多轮次的帧选择，有效解决了长视频处理的难题
  - 文本思考 → 调用工具获取新帧 → 观察新证据并用**<Snapshot>**压缩信息 → 基于新证据继续思考 → 得出更精准判断