

Agent-X

Evaluating Deep Multimodal Reasoning in Vision-Centric Agentic Tasks

Team



Tajamul Ashraf*
MBZUAI



Amal Saqib*
MBZUAI



Hanan Ghani
MBZUAI



Jean Lahoud
MBZUAI



Hisham Cholakkal
MBZUAI



Mubarak Shah
UCF



Philip Torr
University of Oxford



Fahad Khan
MBZUAI



Rao Anwer
MBZUAI



Salman Khan
MBZUAI

Introduction

- **Agentic AI Systems**
 - Perceive multimodal inputs (image, video, text)
 - Plan multi-step reasoning
 - Invoke tools
 - Act in dynamic environments
- **Key Challenge:** Deep reasoning across multiple steps in vision-first settings.

Existing Benchmarks

<i>Benchmark</i>	Agentic Tasks	# Tools	Real-world Queries	Multimodal Inputs	Deep Reasoning	Executable Tools	Hybrid Annotation
APIBench [46]	✗	3					✓
APIBank [29]	✗	73				✓	
ToolBench [48]	✗	3451				✓	
MLAgentBench [20]	-	8					
GAIA [38]	466	26	✓	✓	✓	✓	
SWE-Bench [23]	✗	-					
GTA [62]	229	14	✓	✓		✓	
MLE-Bench [10]	✗	-		✓			
m&m's [34]	-	33		✓		✓	✓
RE-Bench [63]	7	-	✓				
ScienceAgent [13]	102	-	✓			✓	
MLGym [43]	13	16	✓	✓	✓		
Agent-X (Ours)	828	14	✓	✓	✓	✓	✓

Comparison of Vision-Centric Agent Benchmarks. Columns show key dimensions including scale, realism, modality, reasoning depth, tool interaction, and annotation quality. Our benchmark Agent-X uniquely supports all criteria with 828 diverse, manually verified agentic tasks.

Existing Benchmarks

- Lack scalability and fail to reflect real-world complexity
- Often include limited, synthetic tasks that don't support deep reasoning or rich multimodal inputs
- Provide explicit tool hints and instructions, reducing the need for independent planning and reasoning

PRIOR BENCHMARKS

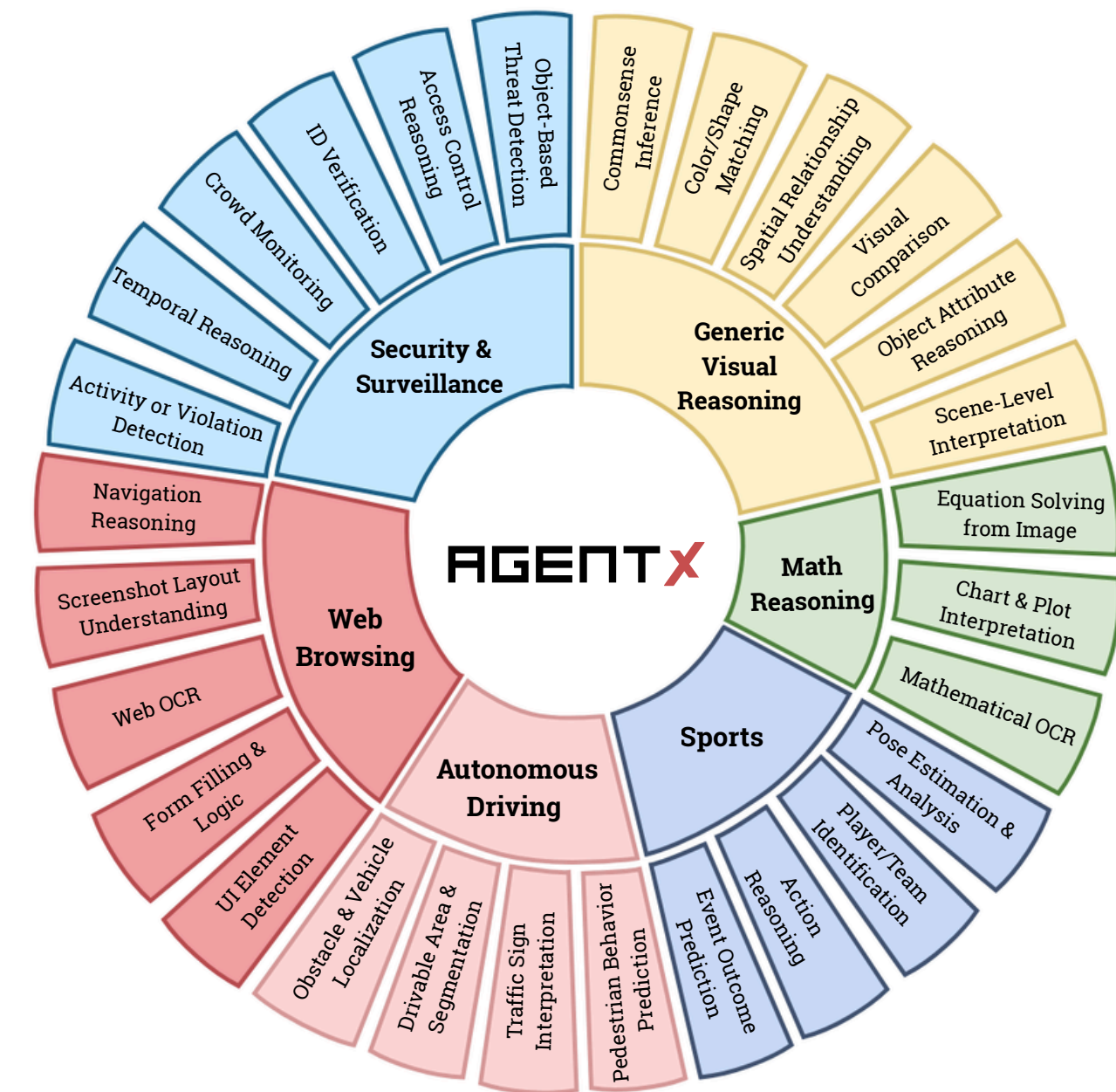
I need an illustration for my children's book. I've imagined a scene where there's a large group of little kids... **After** we have the image, we also need to **identify all the objects, then add labels** to them.

OURS

What store is the scene in the video from and what does the person dressing corresponds to in normal circumstances?

Agent-X

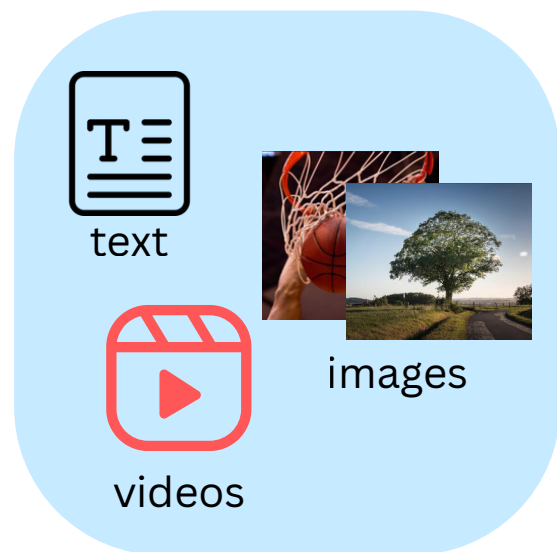
- Targets deep reasoning in vision-centric, real-world tasks
- Covers **6 multimodal environments** involving images, multi-images, and videos
- Provides **828** richly annotated tasks, surpassing prior benchmarks
- Evaluates **12+ LMMs** using fine-grained metrics for reasoning, tool use, and coherence



Distribution of tasks across the six vision-centric environments

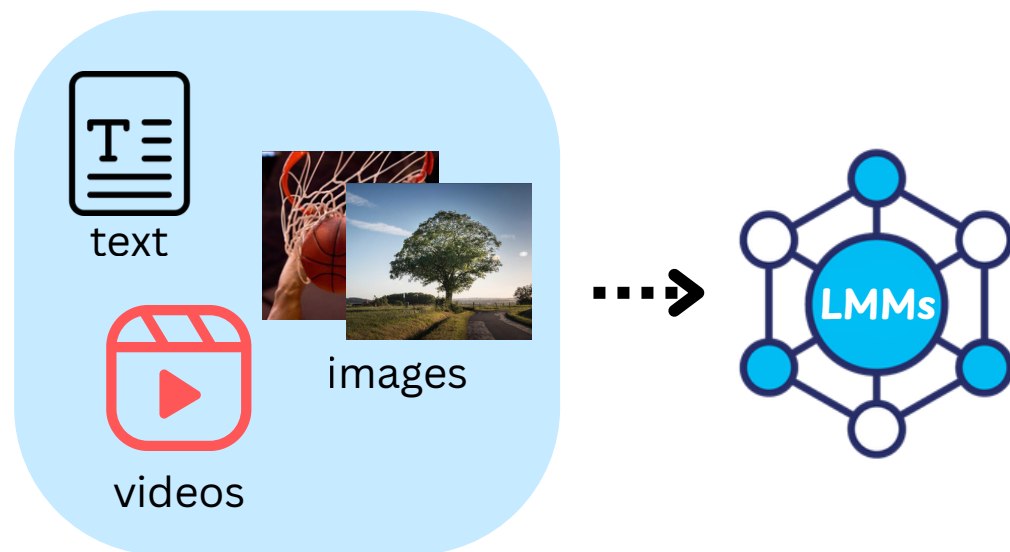
Dataset Curation Pipeline

Multimodal data

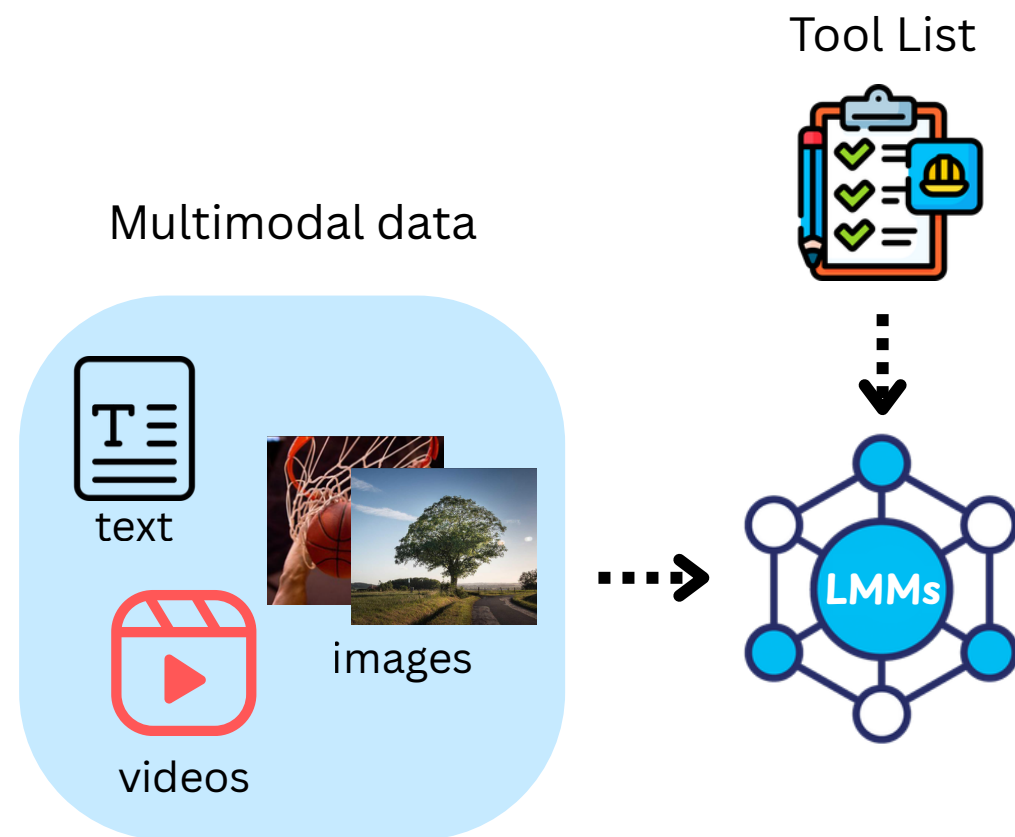


Dataset Curation Pipeline

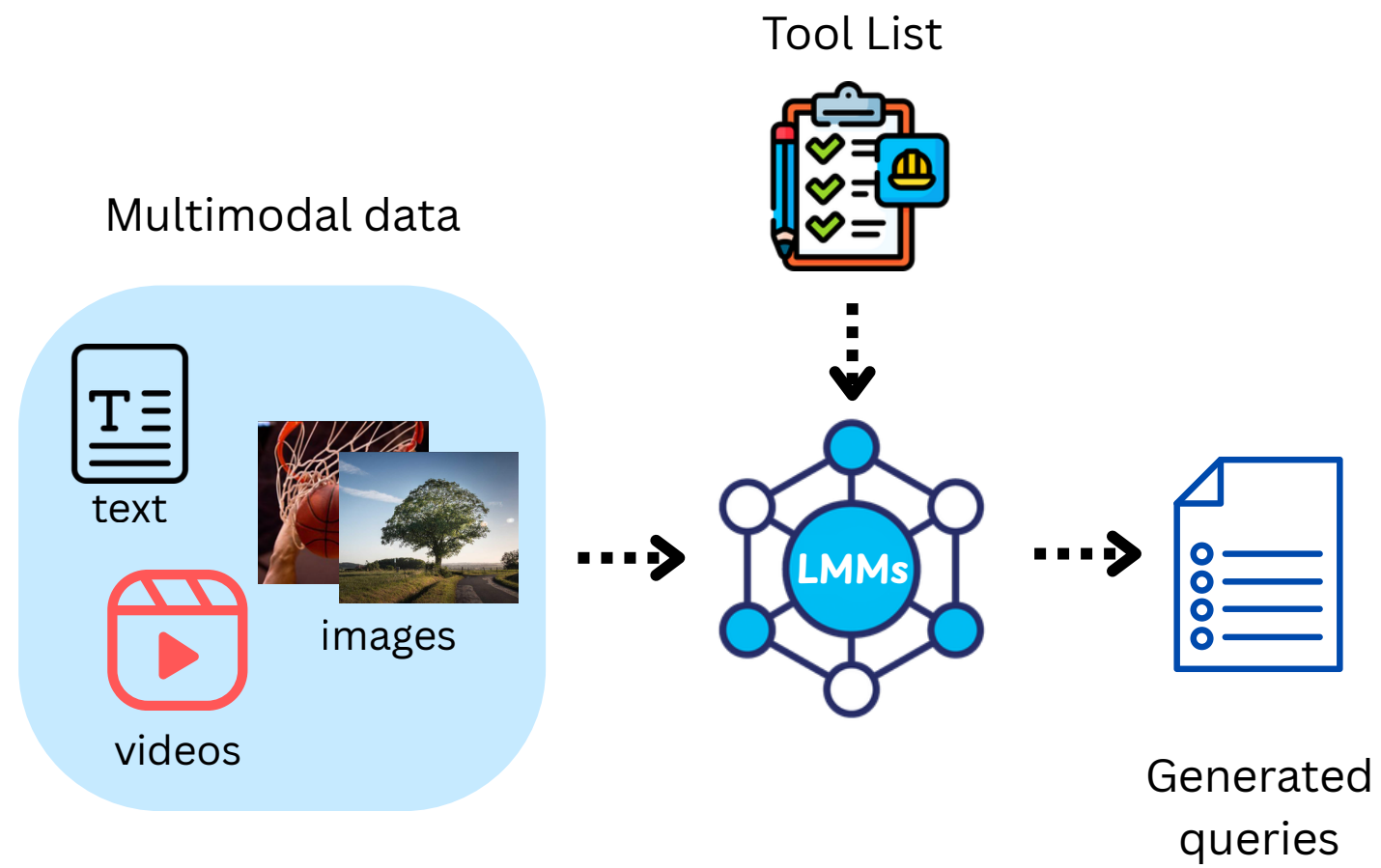
Multimodal data



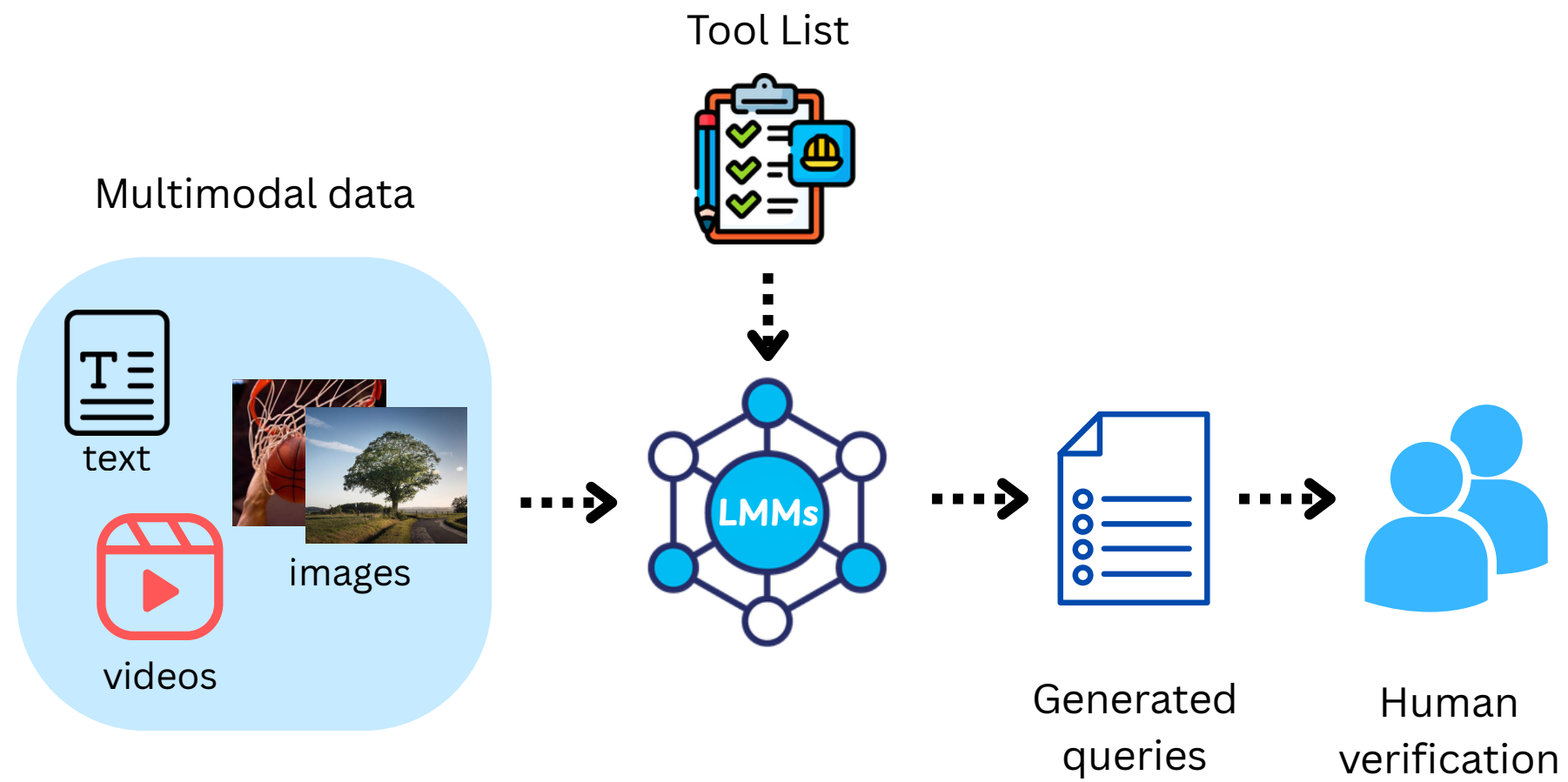
Dataset Curation Pipeline



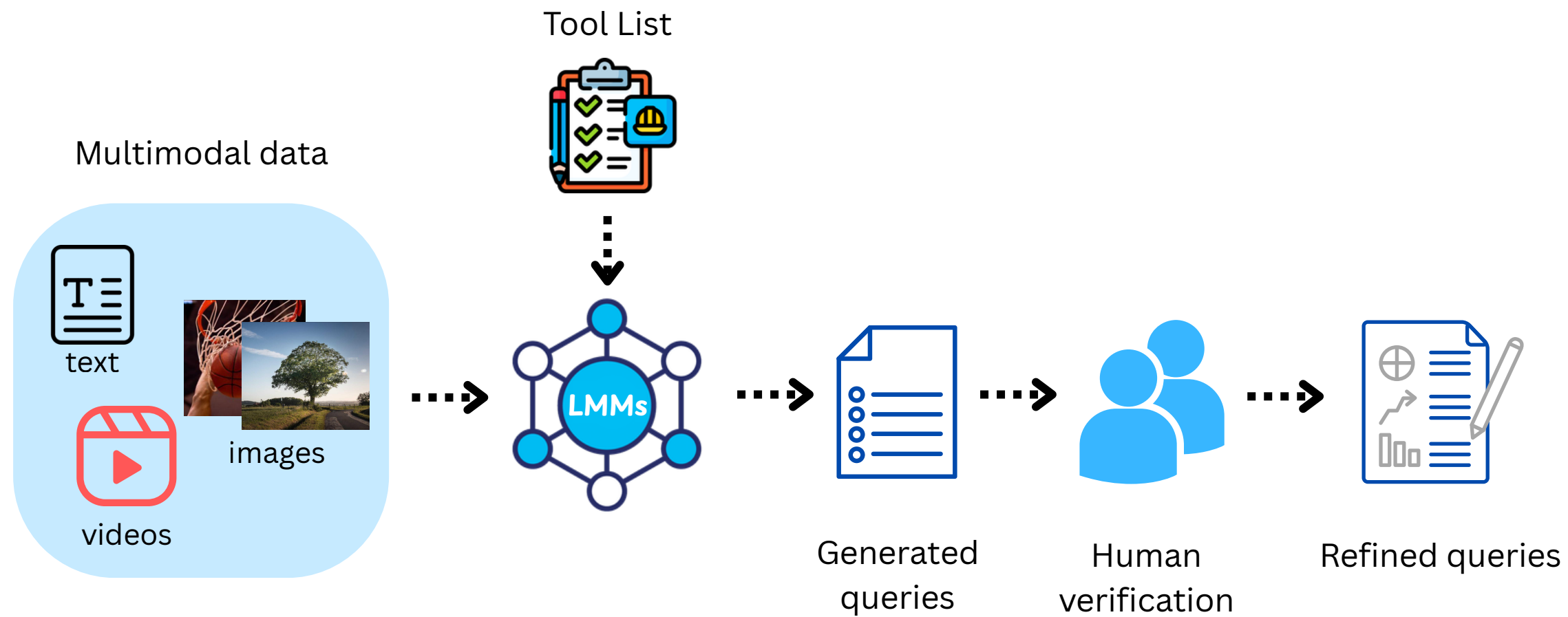
Dataset Curation Pipeline



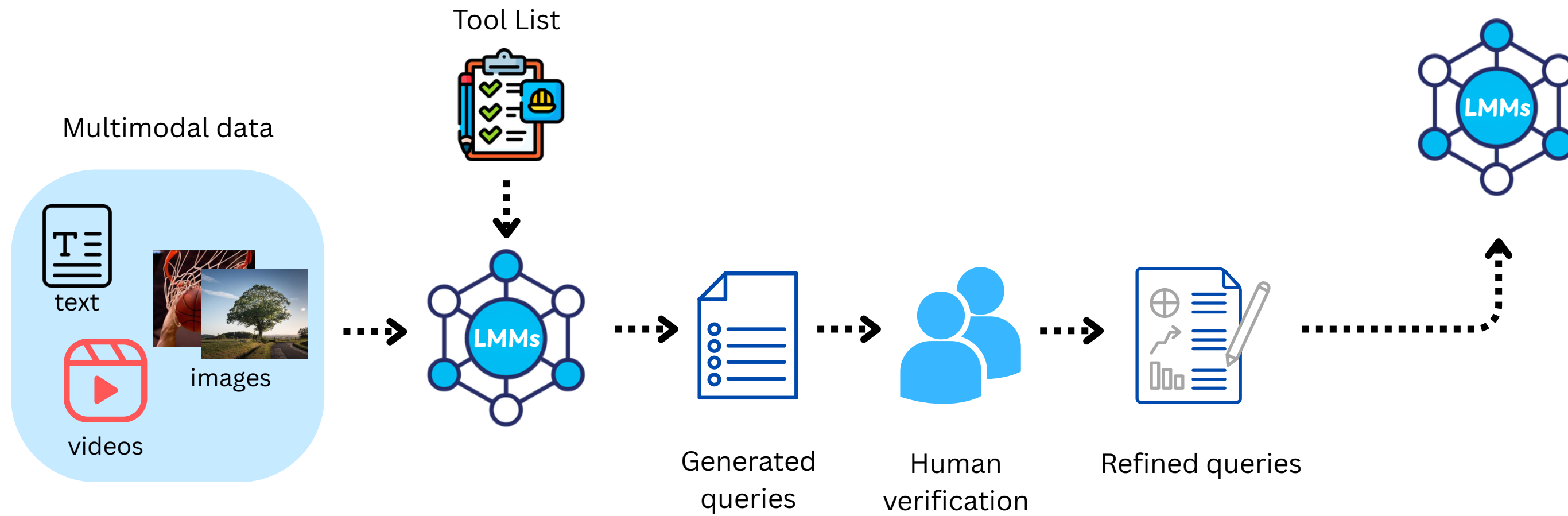
Dataset Curation Pipeline



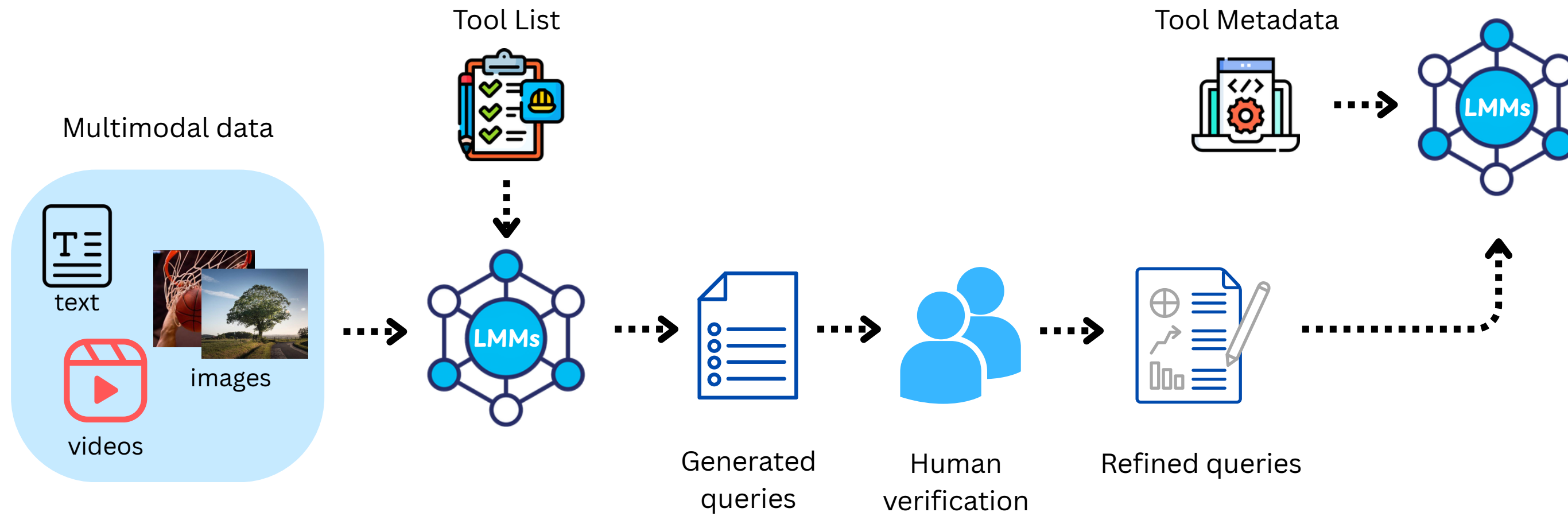
Dataset Curation Pipeline



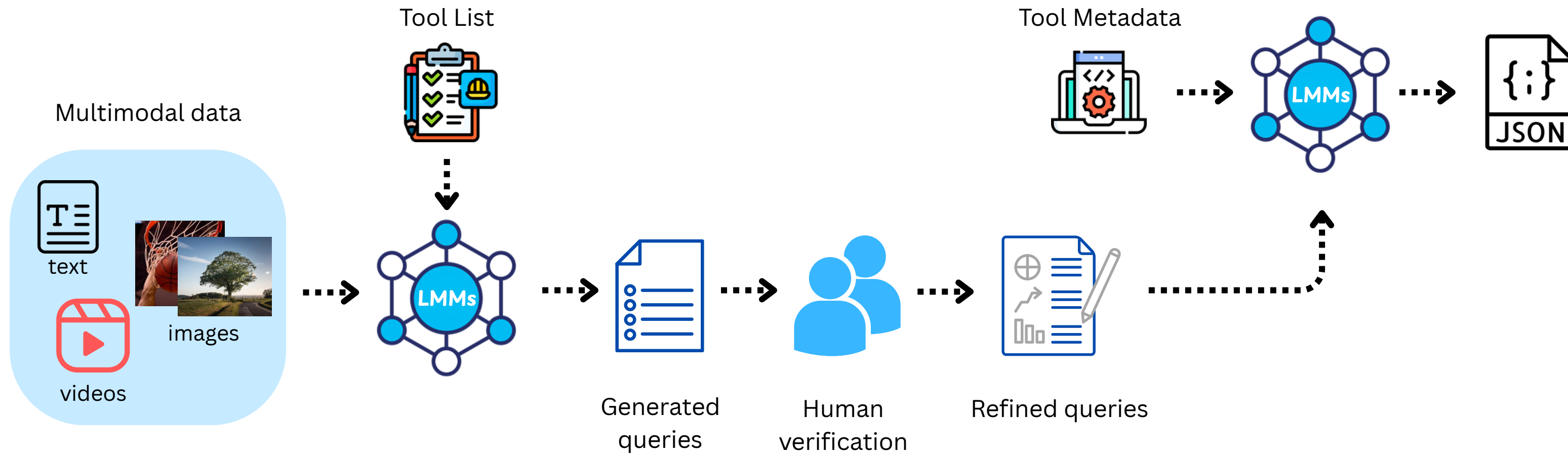
Dataset Curation Pipeline



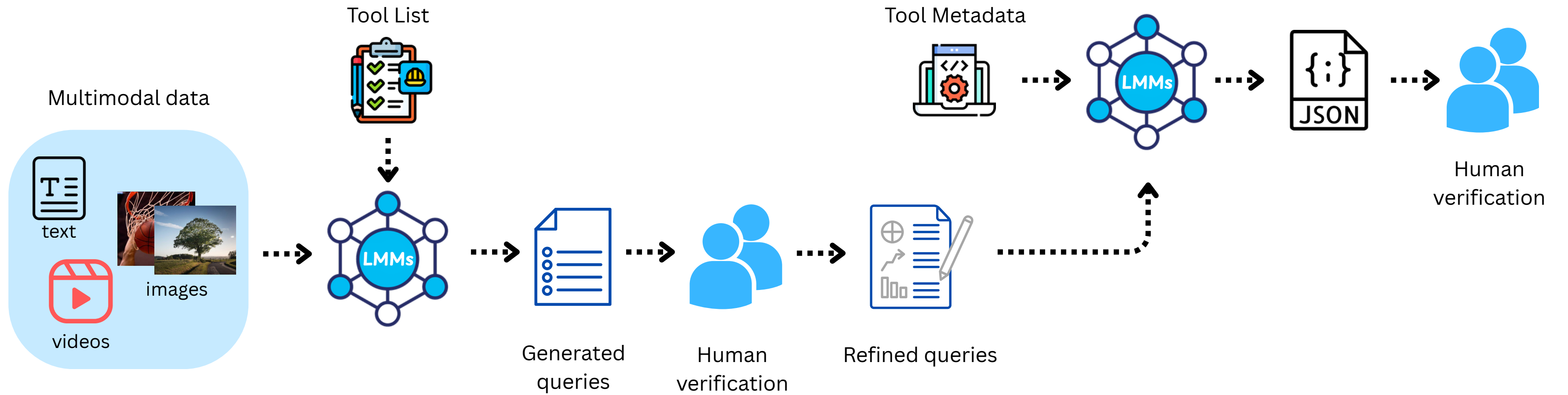
Dataset Curation Pipeline



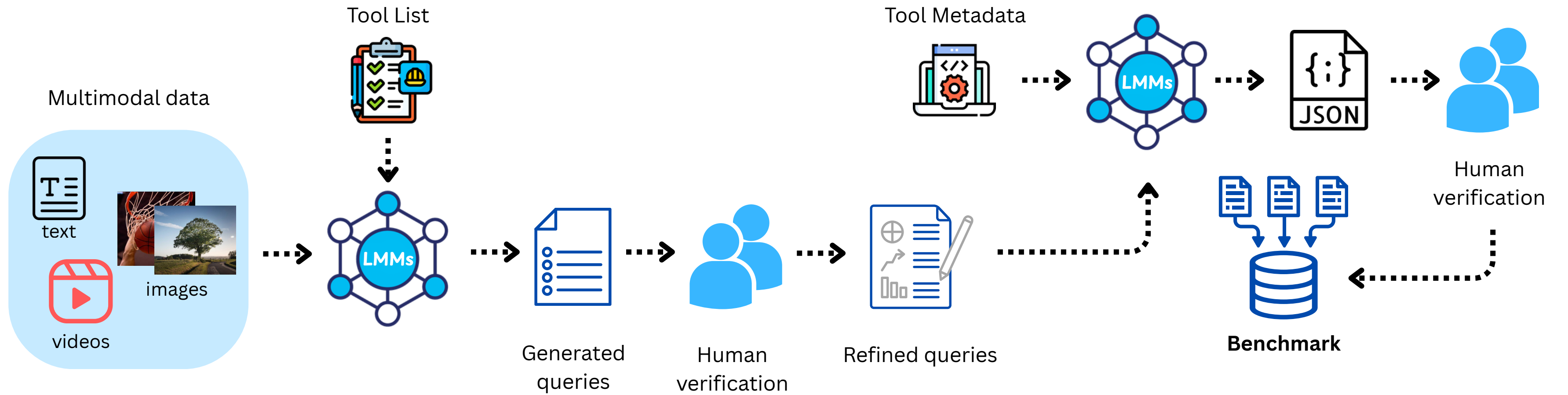
Dataset Curation Pipeline



Dataset Curation Pipeline



Dataset Curation Pipeline



Dataset Curation

Step 1: Multimodal Data & Toolset Definition

- **Goal:** Define visual inputs and executable tool library
- **Visual Inputs (V):** Images and videos from public datasets
- **Tool Library (T^c):** 14 executable tools
 - Perception, Visual Operation, Math, Artistic
- Each future task must be solvable using a subset $T \subseteq T^c$

Dataset Curation

Step 2: Candidate Query Generation (LMM)

- **Goal:** Generate complex, tool-requiring queries
- **Input to LMM:** Visual input V , Full toolset description T^c
- **Process:**
 - Generate 3 candidate queries per visual
 - Cover factual, interpretive, generative types
 - Avoid trivial or perception-only questions
- **Output:** 1,021 visuals \rightarrow 3,063 candidate queries
- Queries must not explicitly mention tools or reasoning steps.

Dataset Curation

Step 3: Human Query Refinement

- **Goal:** Select and refine high-quality queries.
- Annotators select best query among 3 candidates
- Edit for clarity and realism
- **Remove:**
 - Tool-hinting queries
 - Trivially answerable questions
 - Ambiguous or underspecified prompts
- Filtered to 828 validated tasks
- **Quality Criteria:** Requires multi-step reasoning, Requires tool usage, Realistic scenario

Dataset Curation

Step 4: Tool-Augmented Reasoning Trace Generation

- **Goal:** Construct structured reasoning traces
- **Process:**
 - Provide (V_i, Q_i, T^c) to LMM
 - LMM generates:
 - Tool sequence
 - JSON-style arguments
 - Intermediate outputs
 - Final answer

Dataset Curation

Step 5: Human Trace Verification & Correction

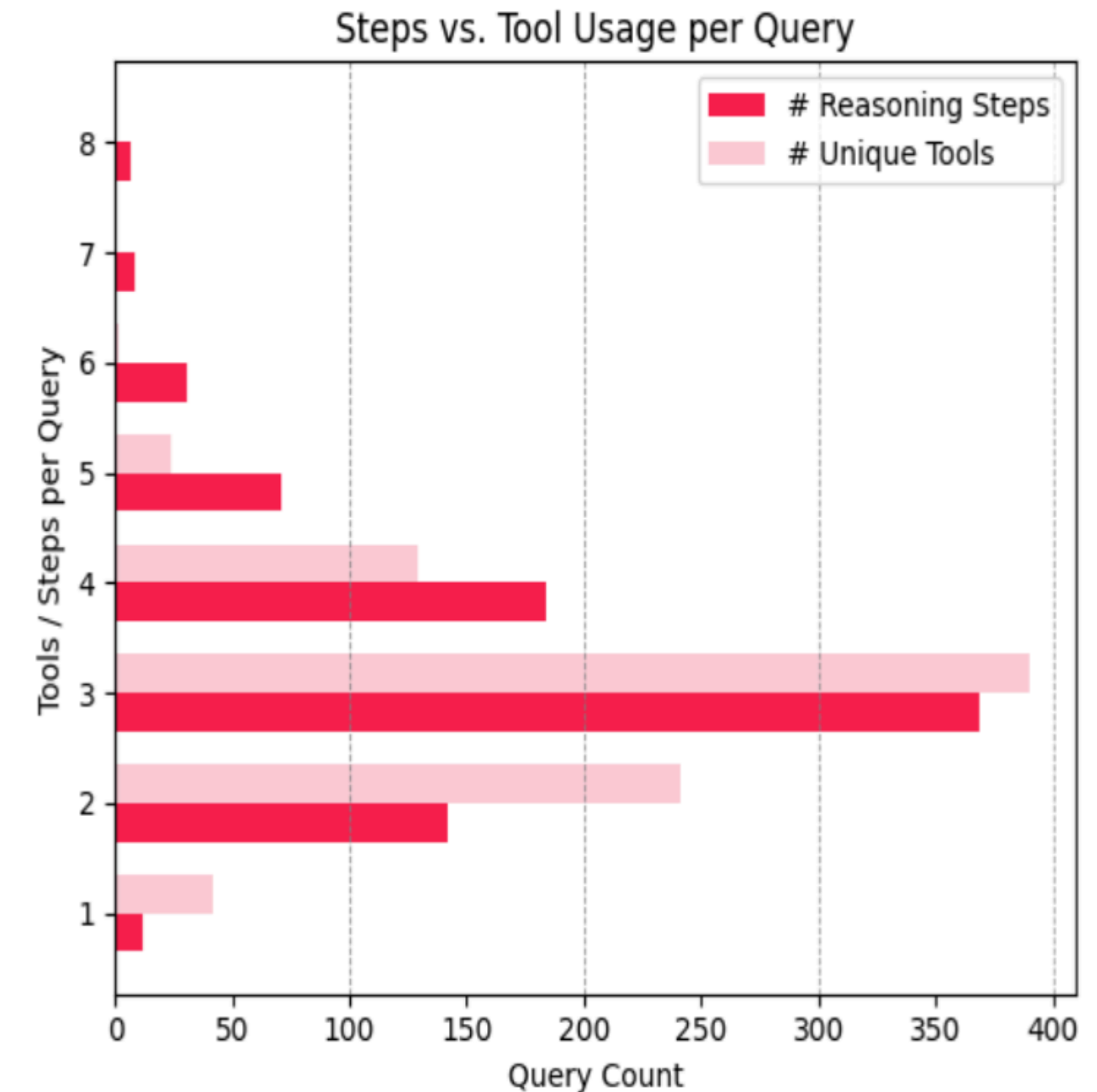
- **Goal:** Ensure logical and executable correctness
- **Annotators verify:**
 - Correct tool selection
 - Proper argument formatting
 - Logical step ordering
 - Alignment between outputs and final answer
 - Evidence support (URLs/screenshots for web queries)
- **Corrections include:**
 - Replacing incorrect tools
 - Fixing malformed JSON
 - Removing unsupported reasoning
 - Tasks that cannot be reliably solved are discarded

Dataset Curation

Step 6: Final Assembly & Dataset Statistics

Statistic	Value
Total Tasks	828
Answers w/ Image	222
Answers w/ Text	606
Raw queries / Refined queries	1021 / 828
Image inputs	716
Video inputs	112
Total tool calls	2807
Average steps per task	3.4
Unique tools used	14
Examples w/ 2 / 3 / 4 / 5 tools	239 / 389 / 130 / 24
Agentic Environments	6
Total Human Annotators	5
Verification time per annotator	~50 hrs

Key data statistics



Overall frequency of tool usage and number of steps

Evaluation Metrics

Metric	Notation	Description
Step-by-Step Mode		
Grounding Score	G_s	<i>Correct reference to objects, regions, or attributes in the input.</i>
Tool Precision	T_p	<i>Accuracy of selecting the correct tool at each reasoning step.</i>
Tool Accuracy	T_a	<i>Correct use of tools with appropriate inputs and outputs.</i>
Deep Reasoning Mode		
Faithfulness	F_{acc}	<i>Logical consistency across the reasoning process.</i>
Context Score	C_s	<i>Effective use of multimodal and commonsense context.</i>
Factual Precision	F_p	<i>Correctness of factual information without hallucination.</i>
Semantic Accuracy	S_{acc}	<i>Coverage of all semantically necessary elements.</i>
Outcome Mode		
Goal Accuracy	G_{acc}	<i>Final answer accuracy for factual and interpretive queries.</i>
Goal Accuracy w/ImgGen	G_{acc}^*	<i>Final answer accuracy for generative queries.</i>
Toolset Accuracy	T_{acc}	<i>F1 score for overall correct tool selection and use.</i>

Evaluation Metrics categorized by Mode. This table outlines the full suite of metrics used in Agent-VBench benchmark, organized by Step-by-Step, Deep Reasoning, and Outcome modes.

Evaluation Setup

- 14 LMMs (Open + Closed) evaluated
- **Judges & Validation:** GPT-4o (closed-source judge), Qwen-14B / 15B (open-source judge), Human validation subset
 - **Inter-judge agreement:** Pearson > 0.9 (Goal & Tool Accuracy), Cohen's $\kappa > 0.8$, Stable model rankings
- **Reproducibility:** 3 independent runs per model, Max variance $< 0.4\%$, Deterministic rankings
- **Ablation: Does Agent-X Measure Reasoning?**

Setting	Goal Accuracy
Perception-only (no reasoning trace)	0.33
Full reasoning chain (Agent-X)	0.43

+10% gain confirms deep reasoning contribution

GPT-4o as a judge

Model	STEP-BY-STEP			DEEP REASONING				OUTCOME		
	G_s	T_p	T_{acc}	F_{acc}	C_s	F_p	S_{acc}	G_{acc}	G_a^*	T_{acc}^s
<i>Open-source*</i>										
Phi-4-VL-Instruct [1]	0.13	0.21	0.24	0.61	0.19	0.47	0.40	0.11	0.26	0.42
InternVL2.5-8B [12]	0.45	0.31	0.47	0.68	0.47	0.52	0.60	0.28	0.55	0.58
Gemma-3-4B [57]	0.26	0.30	0.78	0.61	0.54	0.38	0.54	0.27	0.67	0.60
InternVL3-8B [81]	0.46	0.34	0.54	0.68	0.45	<u>0.70</u>	0.40	0.20	0.59	0.62
VideoLLaMA3-7B [75]	0.45	0.28	0.46	0.65	0.46	<u>0.62</u>	0.54	0.28	0.54	0.54
Qwen2.5-VL-7B [6]	<u>0.54</u>	<u>0.43</u>	0.63	<u>0.75</u>	0.57	0.56	0.67	0.36	0.65	<u>0.67</u>
<i>Closed-source*</i>										
Gemini-1.5-Pro [56]	0.43	0.23	<u>0.84</u>	0.62	0.45	0.53	0.62	0.04	0.56	0.48
Gemini-2.5-Pro [55]	0.40	0.36	0.81	0.72	0.48	0.64	<u>0.73</u>	0.40	0.56	0.62
GPT-4o [21]	0.60	0.47	0.72	0.81	0.57	0.79	0.59	<u>0.37</u>	0.70	0.68
OpenAI-o4-mini [45]	0.42	0.32	0.89	0.71	<u>0.51</u>	0.60	0.80	0.45	<u>0.67</u>	0.63

Qwen-14B as a judge

Model	STEP-BY-STEP			DEEP REASONING				OUTCOME		
	G_s	T_p	T_{acc}	F_{acc}	C_s	F_p	S_{acc}	G_{acc}	G_a^*	T_{acc}^s
<i>Open-source</i>										
Phi-4-VL-Instruct [1]	0.27	0.11	0.32	0.54	0.39	0.59	0.46	0.16	0.35	0.39
InternVL2.5-8B [12]	0.38	0.16	0.49	0.63	0.51	0.61	0.55	0.29	0.53	0.53
Gemma-3-4B [57]	0.50	0.24	0.67	0.74	0.66	0.59	0.74	0.30	<u>0.68</u>	<u>0.68</u>
InternVL3-8B [81]	0.41	0.16	0.51	0.71	0.61	0.60	0.69	0.23	0.51	0.62
VideoLLaMA3-7B [75]	0.39	0.15	0.40	0.68	0.56	0.60	0.68	0.27	0.53	0.56
Qwen2.5-VL-7B [8]	0.51	<u>0.27</u>	0.63	0.77	0.66	0.64	0.77	<u>0.37</u>	0.62	0.67
<i>Closed-source</i>										
Gemini-1.5-Pro [56]	<u>0.57</u>	<u>0.36</u>	0.80	0.82	0.73	0.76	0.63	0.05	0.77	<u>0.71</u>
Gemini-2.5-Pro [55]	0.63	0.40	<u>0.84</u>	<u>0.86</u>	<u>0.76</u>	0.80	<u>0.83</u>	<u>0.50</u>	<u>0.74</u>	0.72
GPT-4o [21]	0.46	0.27	0.63	0.72	0.59	0.75	0.69	0.44	0.48	0.56
OpenAI-o4-mini [45]	0.63	0.35	0.86	0.89	0.78	<u>0.79</u>	0.88	0.53	0.64	0.69

Quantative Error Analysis

Error Type	GPT-4o [21]	Gemini-1.5-Pro [55]	InternVL3-8B [81]
<i>Planning Errors:</i>			
No action, no response.	157 (17.6%)	3 (0.2%)	172 (12.8%)
No action, the whole response is a model thought.	0	2 (0.1%)	0
<i>Formatting Errors:</i>			
Invalid JSON format in argument specification.	235 (26.4%)	755 (44.5%)	454 (33.8%)
Multiple tool calls in a single step.	118 (13.2%)	172 (10.1%)	126 (9.4%)
Final answer generation without adhering to the format.	60 (6.7%)	174 (10.3%)	220 (16.4%)
<i>Reasoning Errors:</i>			
Misinterpreting visual content (e.g., wrong object recognition)	165 (18.5%)	581 (34.3%)	189 (14.1%)
Incorrect spatial reasoning (e.g., wrong relative position)	156 (17.5%)	8 (0.5%)	181 (13.5%)
Total Errors	891(100%)	1695 (100%)	1342(100%)

Summary of common planning, formatting, and reasoning errors on Agent-X benchmark across GPT-4o, Gemini-1.5-Pro, and InternVL3-8B.

Qualitative Error Analysis

Prompt

You are an intelligent multi-modal agent. You are provided with: - A **text query <query>** - An **image or video <files>** - A set of tools to assist with your reasoning with meta data of tools given as follows: {meta_data}. Your objective is to answer the query based on the given visual content by choosing and using the most appropriate tools. You must reason step-by-step. Each reasoning step should include: "reasoning_step_format": [{"task": "Describe the sub-task being performed."}, {"tool_used": "Specify the tool selected and justify its choice."}, {"tool_output": "Provide the tool's output."}, {"thought": ".....", "Ensure each step is self-contained and clearly explained."}, {"....."final_output": {"final_answer": "Provide a clear and concise answer based on all previous steps", "justification": "Provide a justification for the final answer."}]

GPT-4o

Input Files: AgentX_488.mp4

Query: "How many women are there in the video and what is the color of their clothes?"

Error Response

reasoning_steps": [{"step": 1, "task": "Count the number of women in the video.", "tool": "ObjectCounter", "tool_output": "2", "thought": "The ObjectCounter tool detected 2 women in the video frames."}, {"step": 2, "task": "Identify the color of their clothes.", "tool_used": "SceneDescriber", "tool_output": "One woman is wearing red and another is wearing dark clothing.", "thought": "The description tool provided clear information on the clothing colors of the women."}], "final_answer": {"value": "There are **2 women** in the video. One is wearing red, and the other is wearing dark clothing.", "justification": "Using object counting and scene description tools, we accurately identified the number of women and described the colors of their clothes."}]

- **Incorrect count**
- **Shallow reasoning**
- **Wrong tool invocation (no iterative tool calls)**

✓ **The correct response format**
reasoning_steps": [{"step": 1, "task": "Determine how many women are in each frame.", "tool": "ObjectCounter", "tool_output": "", "thought": "I identified the presence of women in each video frame to establish how many unique women appear across the video."}, {"step": 2...}, {"step": 3...}, {"step": 4...}], "final_answer": {"value": "1 woman in the video wearing red clothing.", "justification": "Each frame consistently showed one woman wearing red clothing throughout the video, indicating that it is likely the same individual in every frame."}]

VideoLLaMA3-7B

Input Files: AgentX_215.1.jpg, AgentX_215.2.jpg

Query: "How much do those bottles cost?"

Error Response

"final_answer": "\$50.394",
"reasoning_steps": [{"step": 1, "task": "Find the price of Hennessy.", "tool_used": "Table Search Tool", "tool_output": "Hennessy is priced at \$50.394", "thought": "The table provides a list of products and their prices, so we can directly look up the price of Hennessy from the table."}, {"step": 2, "task": "Provide the answer.", "tool_used": "", "tool_output": "\$50.394", "thought": "The answer is directly available from the table search result."}]

- **Invalid JSON format**
- **Missing justification**
- **Incorrect final answer**
- **Hallucinated tool ("Table Search Tool")**
- **Empty tool entry**
- **No step-by-step reasoning**
- **Redundant steps**

✓ **The correct response format**
"reasoning_steps": [{"step": 1, "task": "Describe the objects in the first image to identify the type of bottles.", "tool": "SceneDescriber", "tool_output": "The image shows two Hennessy cognac bottles.", "thought": "Determining the type of bottles is crucial for identifying the correct price from the price list."}, {"step": 2 ...}], "final_answer": {"value": "\$100.788", "justification": "The bottles are identified as Hennessy, with each costing \$50.394. Thus, two bottles cost \$100.788."}]

Qualitative comparison of GPT-4o and VideoLLaMA3-7B on visual reasoning tasks from Agent-X. GPT-4o hallucinates tool use and produces wrong justifications, while VideoLLaMA3-7B lacks temporal reasoning and fails to align observations across frames. Both models struggle with tool grounding and structured reasoning.

Analysis Summary

- **Real-world tool-use remains challenging:** No model exceeded **50%** Goal Accuracy; even top models struggle with grounding and consistent execution.
- **Strong reasoning = better outcomes:** Models with high faithfulness, precision, and semantic accuracy performed better on final tasks.
- **Tool invocation is a major bottleneck:** Many models fail due to argument formatting issues and incorrect tool chaining.
- **Common errors:** Include shallow or hallucinated reasoning, invalid formats (esp. JSON), visual misinterpretation, and poor temporal grounding in videos.

Conclusion

- Agent-X is a comprehensive benchmark for evaluating reasoning and tool use in real-world, vision-centric tasks.
- It spans diverse multimodal environments and uses executable tools to simulate authentic agent behavior.
- Evaluation of 14 models reveals that even top-performing LMMs struggle with deep reasoning and tool execution.
- Agent-X offers actionable insights and a foundation for advancing robust, multimodal agentic systems.

🌟 Thank you!