VIS2025

An Evaluation-Centric Paradigm for Scientific Visualization Agents

From experimental demos to reliable scientific agents through systematic evaluation.

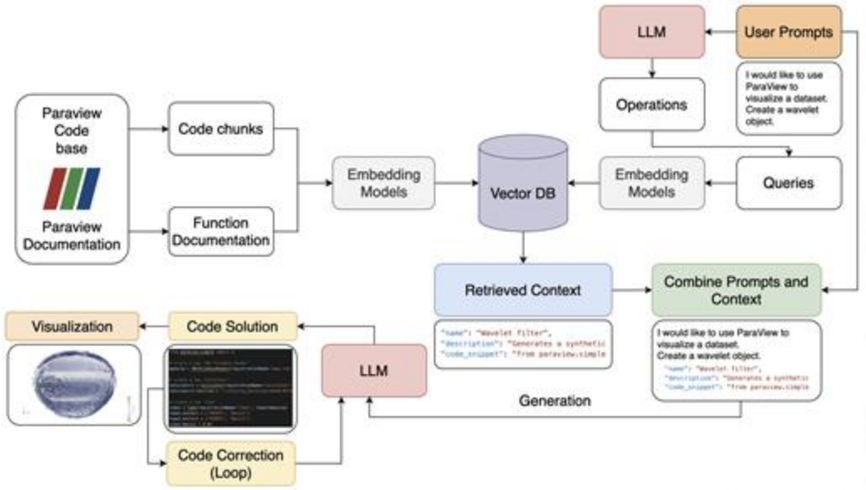
Kuangshi Ai[1], Haichao Miao[2], Zhimin Li[3], Chaoli Wang[1], Shusen Liu[2]

1 University of Notre Dame 2 Lawrence Livermore National Laboratory 3 Vanderbilt University

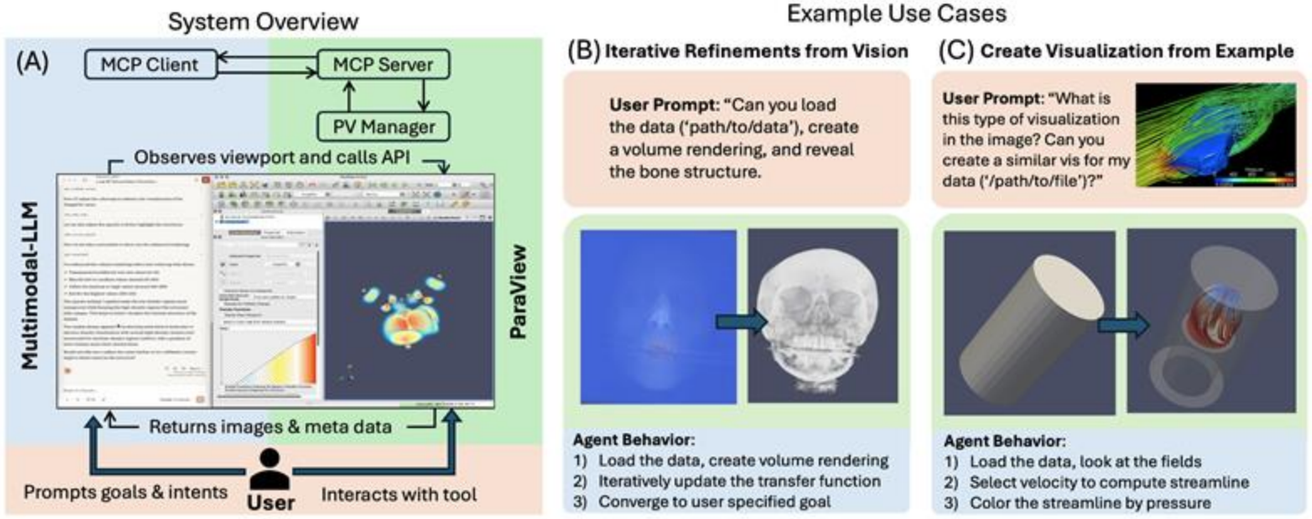
Prepared by LLNL under Contract DE-AC52-07NA27344.

LLNL-CFPRES-2012997

Motivation: MLLMs Enabled Increasingly Sophisticated Autonomous Visualization Agents



ChatVis [Peterka et al., 2025]: code-generation agent with RAG, chain-of-thought, and iterative error correction



ParaView-MCP [Liu et al., 2025] & **BioImage-Agent** [Miao et al., 2025]: interactive tool-use agents built on the Model Context Protocol (MCP)

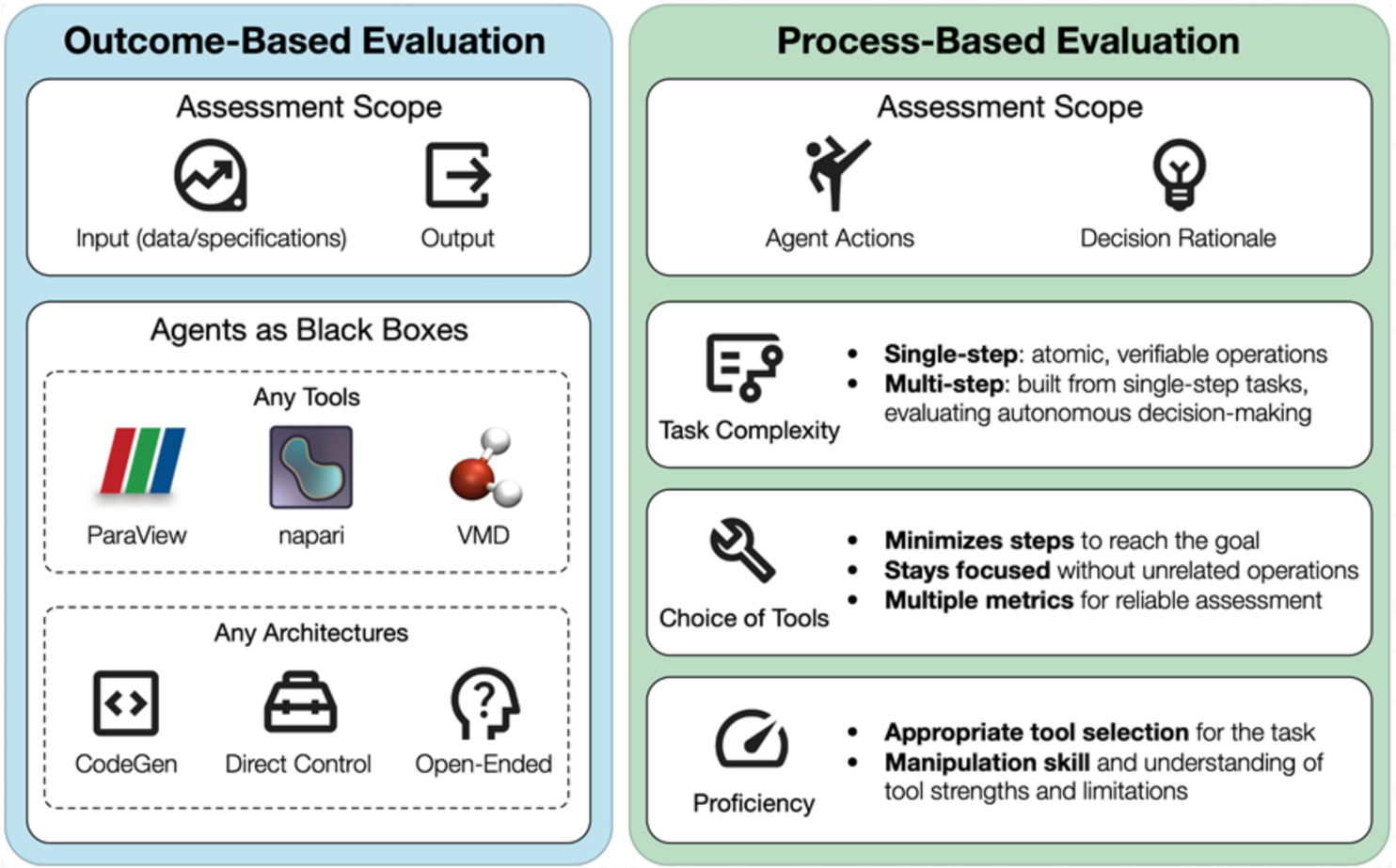
Wednesday, 11:15AM "Visualization for Science" Session

Transforming SciVis Agents from Experimental Tools to Reliable Instruments

- Existing benchmarks mainly focus on simple plotting tasks (VisEval [Cheng et al. 2024], NL2VIS [Wu et al. 2024], AVA [Liu et al. 2024]) or general data science workflows (DA-code [Huang et al. 2024], ScienceAgentBench [Chen et al. 2025])
- SciVis often involves exploratory analysis with emergent insights and complex workflows, making reproducible and measurable evaluation essential.

Evaluation must drive design, not follow it.

Evaluation Taxonomy: Outcome vs. Process / Objective vs. Subjective



Objective: scores, metrics, code evals, etc.

Subjective: LLM as a judge

LLNL-CFPRES-2012997

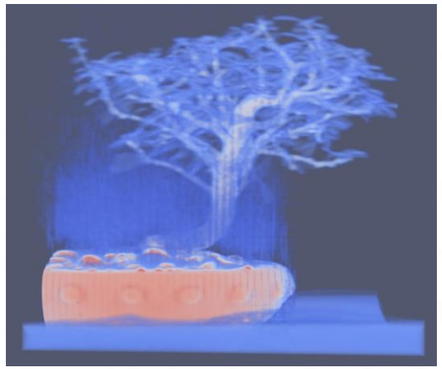
How Effective/Trustworthy is the Evaluation?

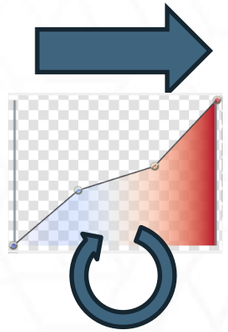
Dimension	Core Question	Example
Accuracy	*Are results/behaviors reliable?*	LLM judge, tool state checks, compare to gold standard scripts
Coverage	*Are real SciVis tasks represented?*	Task complexity, datasets, agent capabilities,
Cost-effectiveness	*Is evaluation feasible?*	Runtime, token & compute cost, human involvement

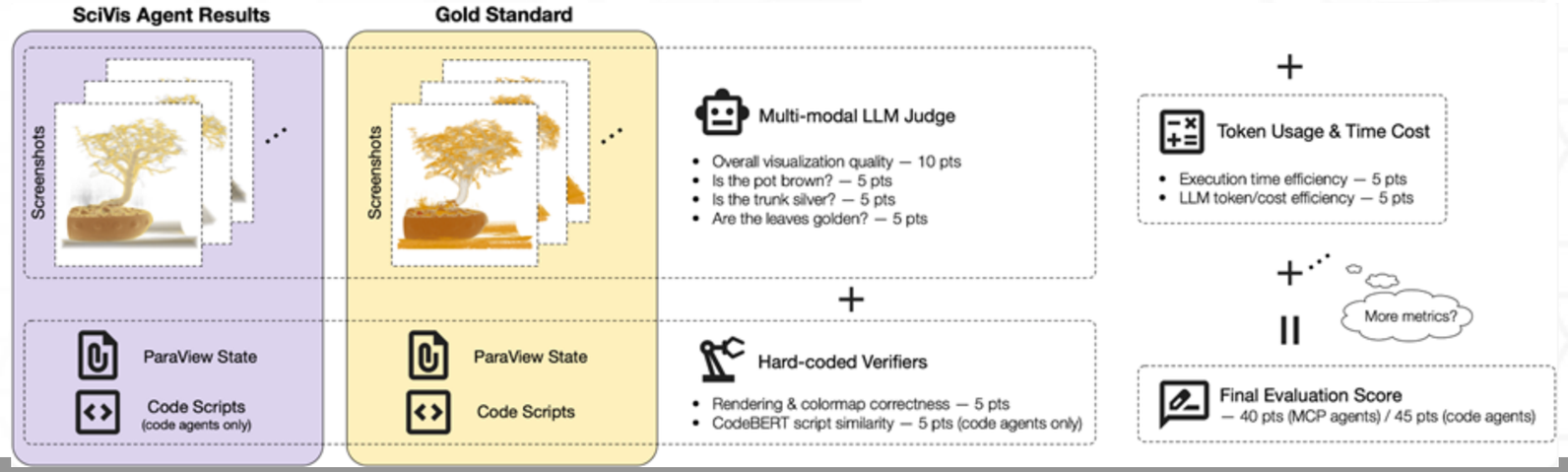
Balancing these three ensures benchmarks are both rigorous and usable.

Illustrative Example: SciVis Agent Evaluation

Goal: A potted tree with a brown pot, silver branches, and golden leaves.







Illustrative Example: Results

We evaluated ChatVis and ParaView-MCP, both powered by GPT-5, GPT-4.1, and GPT-4o.
Each experiment was run 10 times. SR denotes success rate, and Score reflects the best evaluation result per setting.

agent	model	I/O tokens	avg cost	time (s)	SR	score
MCP-based	GPT-5	220 ± 0 / 838 ± 203	\$0.0087	301.7 ± 32.3	10/10	27/40
ChatVis	GPT-5	2430 ± 847 / 2994 ± 956	\$0.0330	158.9 ± 29.9	10/10	25/45
MCP-based	GPT-4.1	220 ± 0 / 1460 ± 210	\$0.0121	49.3 ± 8.0	10/10	21/40
ChatVis	GPT-4.1	638 ± 555 / 1217 ± 530	\$0.0110	24.0 ± 5.7	10/10	23/45
MCP-based	GPT-4o	220 ± 0 / 908 ± 109	\$0.0239	41.7 ± 14.2	10/10	23/40
ChatVis	GPT-4o	1945 ± 753 / 1909 ± 672	\$0.0240	38.4 ± 9.4	7/10	24/45

- MCP delivered higher stability and better visual quality, but with greater latency.
- ChatVis achieved faster runs with lower quality and higher token cost.

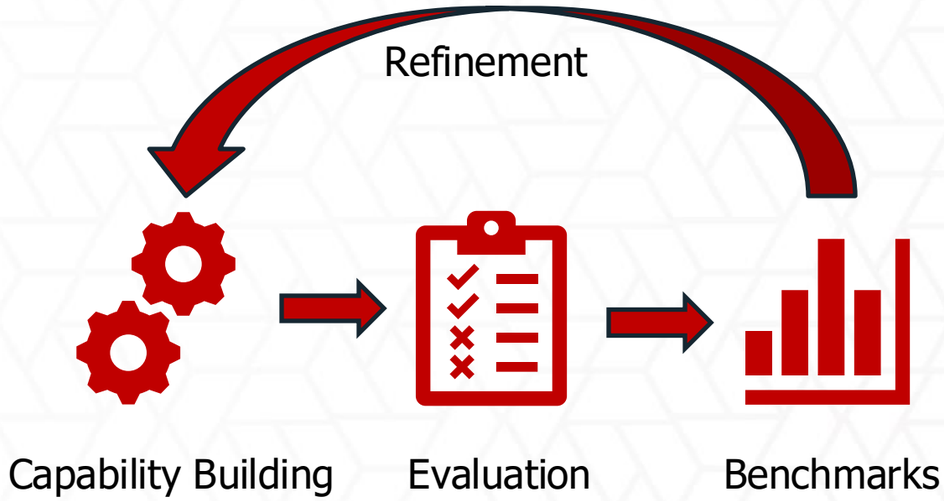
LLNL-CFPRES-2012997

Illustrative Example: Insights

- Significant difference across models (Claude, LLaMA, and Qwen)
- For MCP-based approach, smaller LLMs achieve comparable results (lower latency, reduced costs)
- Combine objective (hard-coded verifiers) with subjective evaluations (LLM as a judge)

Conclusion & Collaboration

- Reliable SciVis agents require standardized, reproducible evaluation across tools and models.
- Building such benchmarks demands **community collaboration** among visualization, AI, and domain experts.
- Enables self-refining agents that learn from evaluation results for continuous improvement → turns benchmarking into **a development accelerator.**



Join us in shaping an open, evaluation-centric framework for the next generation of agentic SciVis!

Thank you for your attention!

Kuangshi Ai[1], Haichao Miao[2], Zhimin Li[3], Chaoli Wang[1], Shusen Liu[2]

1 University of Notre Dame 2 Lawrence Livermore National Laboratory 3 Vanderbilt University

VIS 2025

https://github.com/KuangshiAi/SciVisAgentBench/

Funding:

LDRD NSF

LLNL-CFPRES-2012997