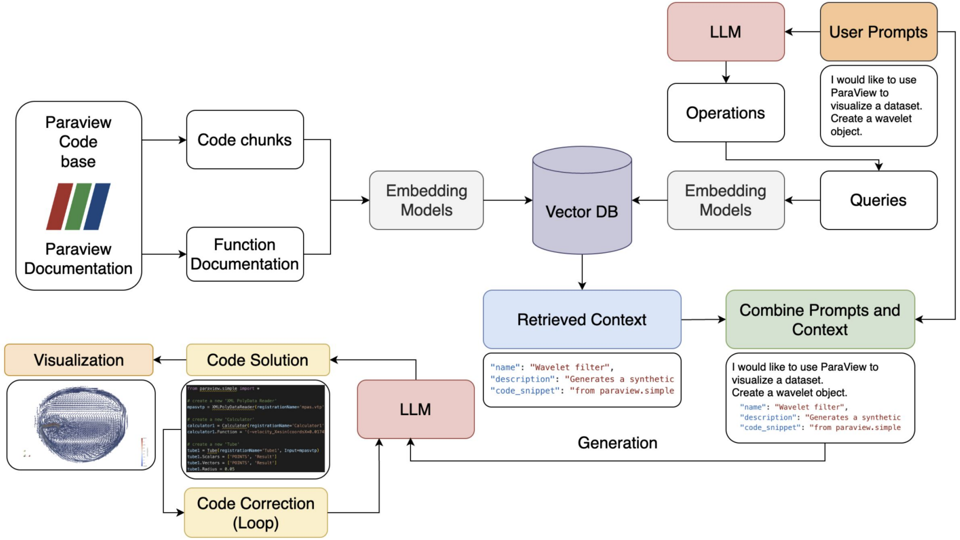
Meeting Agenda

- **Introduction and motivation for the benchmark**
- **A presentation of the current benchmark setup (developed by Kuangshi Ai)**
- **Discussion of the overall goal for the data analysis and visualization benchmark**
    a. As the reference to measure agents advancement, to drive the future innovation in this space
    b. What make it different from existing benchmark (e.g., scientific, multimodal)
- **Discussion of the detailed scope and organization of such a benchmark**
    a. Data Taxonomy
    b. Tasks Taxonomy
- **Potential impacts and perspectives from academia, industry, and labs**
    a. Do academic and industry have different priorities
    b. What would make this benchmark more attractive? (more challenge? More coverage on data/tasks)
- **Logistics for carrying out collaboration for the benchmarks**
    a. Detailed format of the benchmark questions
    b. How to gather, store them, what meta-data should we collect (authors, institution, etc)
    c. Are there tool can be used for this purpose?
    d. Be more inclusive v.s. Contained within a smaller community

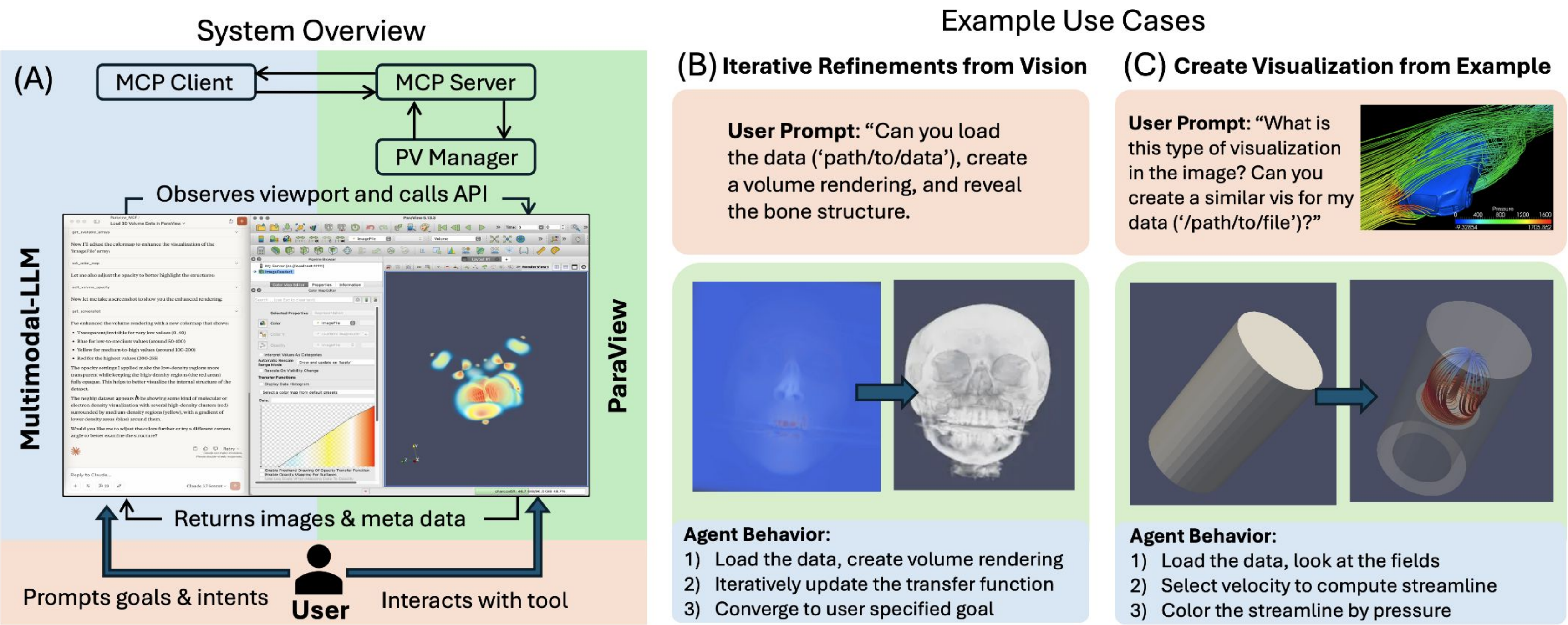
Motivation

Recent multimodal LLMs (MLLMs) enable autonomous agents that turn natural language into complex scientific visualizations (SciVis)



- ChatVis [Peterka et al., 2025]: code-generation agent with retrieval-augmented generation (RAG), chain-of-thought (CoT), and iterative error correction
- More works from this year's vis: e.g., VizGenie

Motivation



- ParaView-MCP [Liu et al., 2025] & BioImage-Agent [Miao et al., 2025]: interactive tool-use agents built on the model context protocol (MCP), **Promptfoo-based eval setup** work with anthropic as part of LLNL-anthropic pilot program

Motivation

- Reliability and robustness of the agent under varying scenarios are often questionable
- Existing benchmarks mainly focus on simple plotting tasks (VisEval [Cheng et al. 2024], NL2VIS [Wu et al. 2024], MatplotAgent [Yang et al. 2025] or general data science workflows (DA-code [Huang et al. 2024], ScienceAgentBench [Chen et al. 2025]) that focus on code generation
- SciVis often involves more involved analysis and complex workflows, making reproducible and measurable evaluation essential
- We argue that **evaluation must drive design**, allow SciVis agents to transition from experimental tools into reliable scientific instruments
    - Integral part of agent design process
    - Forward looking eval needed, i.e., evals the current model / agent can not solve

An Evaluation-Centric Paradigm for Scientific Visualization Agents

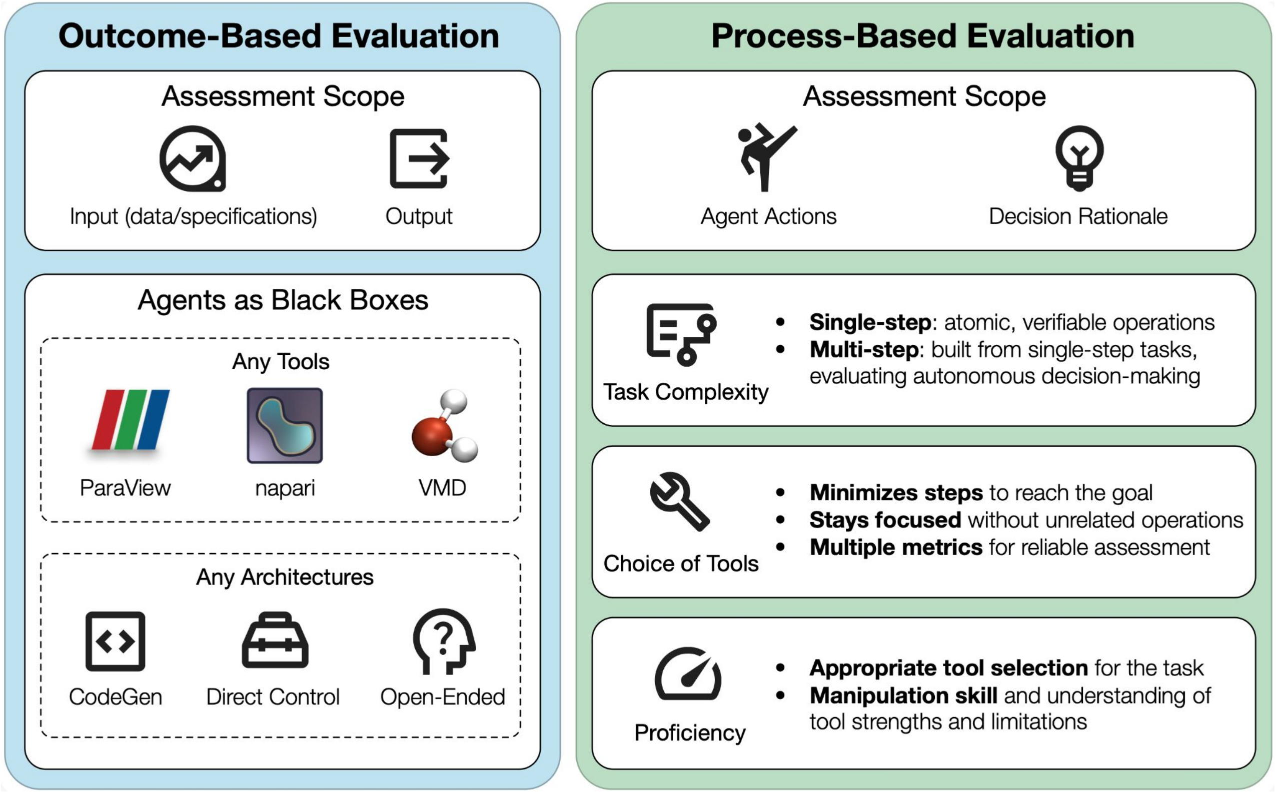
Kuangshi Ai[1], Haichao Miao[2], Zhimin Li[3], Chaoli Wang[1], Shusen Liu[2]

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

UNIVERSITY OF NOTRE DAME
Lawrence Livermore National Laboratory
VANDERBILT UNIVERSITY

VIS 2025
VIS x GenAI

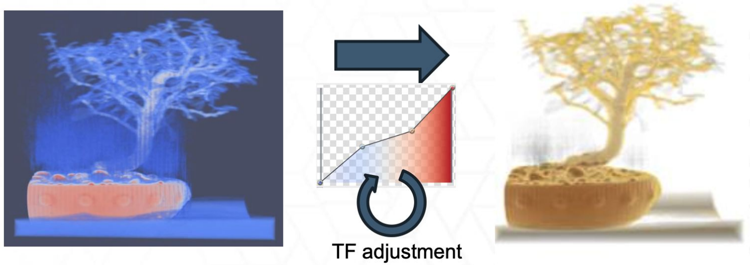
Evaluation Taxonomy: Outcome vs. Process / Objective vs. Subjective

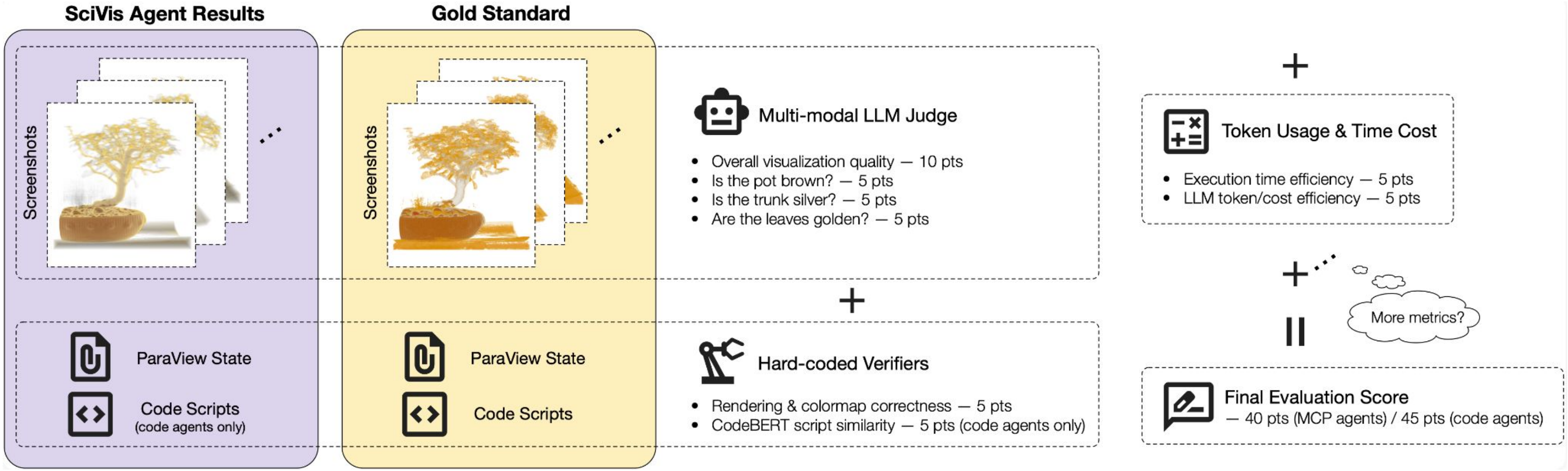


Objective: Hard-coded verifiers (e.g., software state-checks)

Subjective: LLM as a judge

An Illustrative Example: SciVis Agent Evaluation





The evaluation combines: (1) a multi-modal LLM judge for visualization quality, (2) hard-coded verifiers for correctness of visualization primitives and techniques, and (3) token usage and execution time for system performance

An Example of SciVis Agent Evaluation

- For **outcome quality**, we employ instruction-tuned multi-modal LLM judges to check whether the outcome meets the intended goals
- For **process verification**, we use case-specific hard-coded verifiers that inspect the visualization engine's internal state to verify the correct use of visualization primitives (e.g., isosurfaces) and techniques (e.g., volume rendering)
- For **code-generating agents**, additional checks compare the generated scripts to gold-standard references
- For **system efficiency**, we track runtime, token usage, and monetary cost for each run

An Example of SciVis Agent Evaluation

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

- Compared with code generation, MCP delivered higher stability and better visual quality, but with greater latency
- Significant difference across models (Claude, LLaMA, and Qwen)
- For MCP-based approach, smaller language models (SLMs) achieve comparable results (lower latency, reduced costs)

Working-in-progress benchmark

SciVisAgentBench: A Comprehensive Evaluation Framework for Scientific Data Analysis and Visualization Agents

A prototype for now, check our GitHub repo:

https://github.com/KuangshiAi/SciVisAgentBench

Dataset Coverage

See our repo: https://github.com/KuangshiAi/SciVisAgentBench-tasks

ParaView-based tasks (“main/” in our repo, 12 datasets)

- 3D scientific visualization
- Medical imaging, computational simulations, and molecular structures

Volume visualization tasks (“sci_volume_data/”, 37 datasets)

- Additional volumetric data scenarios

ChatVis benchmark (“chatvis_bench/”)

- 20 official ChatVis test cases were transformed to be compatible with our evaluation

Bioimage analysis tasks (“napari_mcp_evals/”)

- Biological imaging workflows via napari

Implementation & Technical Architecture

Evaluated SciVis agents: ParaView-MCP, ChatVis, bioimage-agent

Dual evaluation methodology:

- LLM-as-a-judge (for output text and visualization results): Semantic assessment of task completion and quality based on given metrics
- Quantitative image metrics: PSNR, SSIM, LPIPS across multiple viewpoints

Infrastructure features:

- YAML-based test cases (***promptfoo***-compatible)
- MCP communication logging and automated screenshot capture
- Multi-rubric support (text-based v.s. vision-based evaluation)
- Dataset anonymization tools for blind assessment

Meeting Agenda

- **Introduction**
- **A presentation of the current benchmark setup**
- **Discussion of the overall goal for the data analysis and visualization benchmark**
  a. As the reference to measure agents advancement, to drive the future innovation in this space
  b. What make it different from existing benchmark (e.g., scientific, multimodal)
- **Discussion of the detailed scope and organization of such a benchmark**
  a. Data Taxonomy
  b. Tasks Taxonomy
- **Potential impact and perspectives from academia, industry, and labs**
  a. Do academic and industry have different priorities
  b. What would make this benchmark more attractive? (more challenge? More coverage on data/tasks)
- **Logistics for carrying out collaboration for the benchmarks**
  a. Detailed format of the benchmark questions
  b. How to gather, store them, what meta-data we should collect (authors, institution, etc)
  c. Are there tool can be used for this purpose?
  d. Be more inclusive v.s. Contained within a smaller community

Design Goals / Philosophy

- Mainly focus on outcome-based evaluation, and each benchmark entry contains:
  - Dataset
  - task description
  - ground truth (tricky with visualization tasks as traditionally benchmark ideally need multiple-choice and short-answer questions suitable for automated grading)
  - evaluation criteria
  - Meta info (taxonomy axis)

- Supports multiple ground truth types: text answers, rendered images (for all), spatial info (voxel coordinate), code scripts

- Designed for broad community collaboration across different domains and expertise levels

- Forward looking: what do we want to achieve, rather than what is possible today

Evaluation Effectiveness

- **Accuracy**: the reliability of individual evaluation results
  - MLLMs as judges can be unreliable
  - Effect of recall vs. actually derive answer from the visualization process
  - More effective/reliable eval are for process-based eval: e.g., automated verification against the visualization engine's internal states, compare generated code with reference scripts

- **Coverage**: how much of the potential real-world usage scenarios are covered by the benchmark
  - A full range of SciVis tasks (e.g., volume rendering, isosurface extraction) and datasets (e.g., simulation, biomedical, flow)
  - Varying task complexity (from simple parameter adjustment to complex multi-step pipelines)

- **Cost-effectiveness**: strike a balance between the amount of computational and human efforts and achieving good accuracy and coverage
  - Defining ground truth for exploratory SciVis tasks is challenging
  - Running evaluations demands substantial computation resources

Meeting Agenda

- **Introduction**
- **A presentation of the current benchmark setup**
- **Discussion of the overall goal for the data analysis and visualization benchmark**
    a. As the reference to measure agents advancement, to drive the future innovation in this space
    b. What make it different from existing benchmark (e.g., scientific, multimodal)
- **Discussion of the detailed scope and organization of such a benchmark**
    a. Data Taxonomy
    b. Tasks Taxonomy
- **Potential impact and perspectives from academia, industry, and labs**
    a. Do academic and industry have different priorities
    b. What would make this benchmark more attractive? (more challenge? More coverage on data/tasks)
- **Logistics for carrying out collaboration for the benchmarks**
    a. Detailed format of the benchmark questions
    b. How to gather, store them, what meta-data should we collect (authors, institution, etc)
    c. Are there tool can be used for this purpose?
    d. Be more inclusive v.s. Contained within a smaller community

Task Taxonomy: Data Dimension

Data sources

- CT / MRI / medical scans / n-dim microscopy images
- Physical or fluid simulations
- Molecular Dynamics simulations

Data types

- Scalar: temperature, density, pressure
- Vector: velocity, magnetic field
- Tensor: stress, diffusion, orientation

Task Taxonomy: Task Dimension

Visualization-specific tasks

- Scalar → isosurface / volume rendering / TDA /
- Vector → arrows, streamlines, critical points, glyphs
- Tensor → glyphs etc. material science, TDI, eigenvalue, etc

Interaction & manipulation (tools)

- Zoom / clip / viewpoint
- Guided tour / temporal / spatial exploration
- Visualization-based QA (understand the plot)

Task Taxonomy: Feature & Complexity Dimension

Vision-based tasks (or should this belong to more general tasks)

- Count sub-objects
- Measure size, length, orientation
- Predict future feature behavior (time-varying)

Complexity levels (how to define what is easy vs. hard)

- 🟢 Easy (e.g., single action)
- 🟡 Medium (e.g., workflow, a sequence of actions)
- 🔴 Hard (e.g., derive specific scientific insight from data)

Other Potential Axes?

Task Taxonomy: Other considerations

[Visualization vs. Analysis]
[vtk filter]
[Presentation, slice, position, viewpoint]

[separate tools with tasks]

[evaluation for your grader]

Task Taxonomy: Other considerations

What not include in the benchmark:

- Process based (often tie to tools): Interaction focused
- Extremely large (TB level)
- Label / criteria not clear, Multiple solutions are corrects

Meeting Agenda

- **Introduction**
- **A presentation of the current benchmark setup**
- **Discussion of the overall goal for the data analysis and visualization benchmark**
    a. As the reference to measure agents advancement, to drive the future innovation in this space
    b. What make it different from existing benchmark (e.g., scientific, multimodal)
- **Discussion of the detailed scope and organization of such a benchmark**
    a. Data Taxonomy
    b. Tasks Taxonomy
- **Potential impacts and perspectives from academia, industry, and labs**
    a. Do academic and industry have different priorities
    b. What would make this benchmark more attractive? (more challenge? More coverage on data/tasks)
- **Logistics for carrying out collaboration for the benchmarks**
    a. Detailed format of the benchmark questions
    b. How to gather, store them, what meta-data we should collect (authors, institution, etc)
    c. Are there tool can be used for this purpose?
    d. Be more inclusive v.s. Contained within a smaller community

Meeting Agenda

- **Introduction**
- **A presentation of the current benchmark setup**
- **Discussion of the overall goal for the data analysis and visualization benchmark**
  a. As the reference to measure agents advancement, to drive the future innovation in this space
  b. What make it different from existing benchmark (e.g., scientific, multimodal)
- **Discussion of the detailed scope and organization of such a benchmark**
  a. Data Taxonomy
  b. Tasks Taxonomy
- **Potential impacts and perspectives from academia, industry, and labs**
  a. Do academic and industry have different priorities
  b. What would make this benchmark more attractive? (more challenge? More coverage on data/tasks)
- **Logistics for carrying out collaboration for the benchmarks**
  a. Detailed format of the benchmark questions
  b. How to gather, store them, what meta-data we should collect (authors, institution, etc)
  c. Are there tool can be used for this purpose?
  d. Be more inclusive v.s. Contained within a smaller community

Collaboration Opportunities

What's implemented:

- Complete outcome-based evaluation pipeline with dual assessment methods
- Standardized test case format
- Test cases about VolVis and FlowVis

What we want the community to help with:

1. Dataset contribution

- Submit new scientific datasets with task descriptions
- Provide reference visualizations and ground truths

Collaboration Opportunities

What we want the community to help with:

2. Task taxonomy design & expansion (higher priority)

- Propose new tasks for underrepresented scientific domains (like topological analysis, feature extraction)
- Design evaluation scenarios according to ground truths

3. Evaluation methodology (higher priority)

- Refine comparison criteria for complex spatial information (e.g., vortex detection)
- Develop new metrics for domain-specific assessment
- **Evaluate the grader, calibrate the evaluation tool**

Collaboration Opportunities

What we want the community to help with:

4. Target venues

- IEEE VIS 2026: March 31, 2026
- ICML 2026: Jan 29, 2026
- NeurIPS (Dataset & Benchmark Track) 2026: likely May 15, 2026