A Benchmark Taxonomy

We list the taxonomy used in SciVisAgentBench, including application domain, data type, complexity level, and visualization operation.

A.1 Application Domain Taxonomy

The benchmark covers the following eight application domains: Astronomy, Biology, Chemistry, Earth System Science, Mathematics, Medical Science, Physics, and Others.

A.2 Data Type Taxonomy

The benchmark spans the following five data types: Scalar field, Vector field, Tensor field, Multivariate, and Time-varying.

A.3 Complexity Level Taxonomy

Each benchmark case is categorized into one of the three complexity levels based on the highest-level procedural scope:

- **Operation**: a single, focused visualization procedure.
- **Task**: multiple coordinated operations toward a specific objective.
- **Workflow**: a multi-stage visualization pipeline or complex analytical process.

In practice, we report only task- and workflow-level entries as benchmark cases; operations serve as the building blocks within these higher-level categories.

A.4 Visualization Operation Taxonomy

A visualization operation describes the primary analytical or rendering actions performed. Each benchmark case is tagged with one or more of the following 15 operations.

A.4.1 Color & Opacity Mapping

This operation assigns visual attributes based on data values. It includes colormap assignment, opacity mapping, texture application, material properties, and contrast adjustments. In molecular visualization, coloring by element, residue, or other attributes falls into this category. The operation defines visual encoding rules rather than geometry.

A.4.2 Data Sampling & Resolution Control

This operation modifies data density or sampling resolution while preserving overall semantics. It includes subsampling, decimation, resampling onto different grids, probing fields at user-specified locations, and generating reduced representations for performance or clarity. Unlike subsetting, this operation controls representation fidelity rather than selecting regions of interest.

A.4.3 Data Smoothing & Filtering

This operation reduces noise or enhances data quality without changing geometry or topology. It includes Laplacian or Gaussian smoothing, noise reduction, interpolation-based gap filling, edge enhancement, feature sharpening, outlier removal, and statistical filtering.

A.4.4 Data Subsetting & Extraction

This operation isolates spatial regions or value-based subsets from a dataset. It includes clipping by plane, box, or sphere; extracting a volume of interest; selecting by point or cell IDs; thresholding by scalar range; isocontouring or isosurfacing to extract constant-value boundaries; and connectivity filtering to isolate connected components. The defining characteristic is reducing the dataset to a subset based on spatial location or attribute value.

A.4.5 Dataset Restructuring

This operation reorganizes datasets at the structural or container level. It includes merging datasets, appending blocks, partitioning datasets, and converting between data formats. The defining characteristic is the modification of the dataset organization rather than geometry or attributes.

A.4.6 Feature Identification & Segmentation

This operation detects, extracts, or labels meaningful structures within data. It includes connected component labeling, object detection, region segmentation, topology-based feature classification, and structure identification. The output is a set of labeled objects or regions.

A.4.7 Field Computation

This operation derives new scalar, vector, or tensor fields from existing attributes. It includes gradient, divergence, curl, vorticity, curvature, arithmetic field operations, vector magnitude computation, statistical summaries, tensor eigendecomposition, interpolation, and distance calculations. Particle tracing and critical point detection are also included.

A.4.8 Geometric & Topological Transformation

This operation modifies spatial structure or connectivity without deriving new field values. It includes translation, rotation, scaling, deformation, warping, triangulation, mesh refinement or coarsening, and boundary extraction. The focus is on the structural transformation of existing data.

A.4.9 Glyph & Marker Placement

This operation places discrete visual symbols at data locations to encode local attributes. It includes oriented vector glyphs, scaled glyphs, tensor glyphs, and seed markers. Molecular representations such as spheres, sticks, cartoons, and surfaces are also included. The defining feature is symbol-based encoding of localized data values.

A.4.10 Plot & Chart Generation

This operation produces 2D statistical visualizations from data attributes. It includes histograms, line plots, bar charts, and other standard statistical graphics. The output is a 2D chart rather than a spatial visualization.

A.4.11 Scientific Insight Derivation

This operation uses visualization or analysis results to answer domain-specific scientific questions. Examples include assessing molecular penetration, verifying anatomical structures, or validating simulation plausibility. The defining characteristic is that the output is knowledge or a decision rather than a visualization artifact.

A.4.12 Surface & Contour Extraction

This operation generates geometric primitives representing boundaries or cross-sections within volumetric data. It includes isosurface generation, contour extraction, and planar slicing. The key distinction is that new geometry is produced to represent feature boundaries or cross-sections.

A.4.13 Temporal Processing

This operation explicitly uses the time dimension in time-varying scalar fields or unsteady vector fields. It includes temporal interpolation, temporal aggregation, pathline/streakline/timeline computation, flow map computation, and timestep navigation.

A.4.14 View & Camera Control

This operation manipulates the viewing context rather than the data. It includes camera rotation, zooming, panning, viewpoint presets, lighting adjustment, switching between 2D and 3D modes, and camera reset. Molecular view centering and orientation adjustments are also included.

A.4.15 Volume Rendering

This operation visualizes volumetric data directly without reducing it to surfaces. It includes ray casting, splatting, and transfer function design for mapping scalar values to color and opacity. Multi-channel rendering and X-ray style transparency are also included. The defining property is that the full volume contributes to the final image.

B Benchmark Cases

Table 1 provides the complete list of 108 benchmark cases in SciVisAgentBench.

C Additional Results

Table 2 reports the overall benchmark performance using GPT-5.2 as the LLM judge. Table 3 further reports the full benchmark performance of Claude Code on the bioimage visualization task suite with different backbones and skill settings.

Table 1: The complete list of all 108 benchmark cases in SciVisAgentBench, including their application domains, data types, complexity levels, and visualization operations. Cases are ordered alphabetically based on case names.

Case Name	Domain	Type	Level	Operations
1CRN_color-by-charge	Biology	Scalar Field	Task	Color & Opacity Mapping; Glyph & Marker Placement; View & Camera Control
1CRN_contacts	Biology	Scalar Field	Task	Field Computation
1CRN_cpk-coloring	Biology; Chemistry	Scalar Field	Task	Color & Opacity Mapping; Glyph & Marker Placement; View & Camera Control
1CRN_distance-angles	Biology	Scalar Field	Task	Field Computation
1CRN_licorice	Biology	Scalar Field	Task	Glyph & Marker Placement; View & Camera Control
1CRN_radius-gyration	Biology	Scalar Field	Task	Field Computation
1CRN_rmsd-rmsf	Biology	Scalar Field	Task	Field Computation
1CRN_select-aromatic	Biology	Scalar Field	Task	Color & Opacity Mapping; Glyph & Marker Placement; View & Camera Control
1CRN_select-carbon	Biology	Scalar Field	Task	Color & Opacity Mapping; Glyph & Marker Placement; View & Camera Control
1CRN_select-oxygen	Biology	Scalar Field	Task	Color & Opacity Mapping; Glyph & Marker Placement; View & Camera Control
ABC	Physics	Vector Field	Workflow	Field Computation; Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control
aneurysm	Medical Science	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
argon-bubble	Chemistry	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping; View & Camera Control
backpack	Others	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
bcell-receptor_camera-ops	Biology	Multivariate	Task	Color & Opacity Mapping; View & Camera Control
bcell-receptor_cell-counting	Biology	Scalar Field	Task	Color & Opacity Mapping; View & Camera Control; Scientific Insight Derivation
bcell-receptor_cell-trace	Biology	Scalar Field	Task	Glyph & Marker Placement; View & Camera Control
bcell-receptor_ingest-labels	Biology	Multivariate	Task	Feature Identification & Segmentation; Scientific Insight Derivation
bcell-receptor_ingest-points	Biology	Multivariate	Task	Glyph & Marker Placement; Scientific Insight Derivation
bcell-receptor_ingest-shapes	Biology	Multivariate	Task	Glyph & Marker Placement; Scientific Insight Derivation
bcell-receptor_multi-channel	Biology	Time-varying; Multivariate	Workflow	Color & Opacity Mapping; Temporal Processing; View & Camera Control; Scientific Insight Derivation
Bernard	Physics	Vector Field	Workflow	Field Computation; Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control
blunt-fin	Physics	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
bonsai_dvr	Biology	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
bonsai_interaction	Biology	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping; View & Camera Control
Boston-teapot	Others	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
brain	Medical Science; Mathematics	Tensor Field	Workflow	Field Computation; Feature Identification & Segmentation; Scientific Insight Derivation
bunny	Others	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
can_color-blocks	Others	Scalar Field	Task	Color & Opacity Mapping; View & Camera Control
can_sub-time-series	Others	Time-varying	Task	Temporal Processing; Volume Rendering; View & Camera Control
can_temporal-animation	Others	Time-varying	Task	Temporal Processing; Color & Opacity Mapping; View & Camera Control
carp_dvr	Biology	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
carp_analysis	Biology	Scalar Field	Workflow	Volume Rendering; Color & Opacity Mapping; View & Camera Control; Scientific Insight Derivation
chameleon_isosurface	Biology	Scalar Field	Task	Surface & Contour Extraction; Color & Opacity Mapping; View & Camera Control
climate	Earth System Science	Vector Field	Workflow	Field Computation; Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control
cortical-neurons_annotation	Biology	Multivariate	Task	Glyph & Marker Placement
cortical-neurons_figure-recreation	Biology	Multivariate	Task	Color & Opacity Mapping; View & Camera Control
cortical-neurons_statistics	Biology	Multivariate	Task	Field Computation; Scientific Insight Derivation
crayfish_streamline	Physics	Vector Field	Workflow	Field Computation; Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control
curved-membrane	Biology	Scalar Field	Workflow	Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control; Scientific Insight Derivation
cylinder	Physics; Mathematics	Scalar Field	Workflow	Field Computation; Feature Identification & Segmentation; Data Smoothing & Filtering; Scientific Insight Derivation
disk-out-ref_line-plot	Others	Scalar Field	Task	Plot & Chart Generation; Color & Opacity Mapping; View & Camera Control
dragon	Mathematics	Scalar Field	Workflow	Field Computation; Feature Identification & Segmentation
engine_dvr	Others	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
engine_interaction	Others	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping; View & Camera Control
foot_dvr	Medical Science	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
foot_analysis	Medical Science	Scalar Field	Workflow	Volume Rendering; Color & Opacity Mapping; View & Camera Control; Scientific Insight Derivation
frog	Biology	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
fuel	Physics	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
hydrogen-atom	Physics	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
Isabel	Earth System Science; Mathematics	Scalar Field	Workflow	Field Computation; Feature Identification & Segmentation; Data Smoothing & Filtering; Scientific Insight Derivation
lobster_dvr	Biology	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
lobster_isosurface	Biology	Scalar Field	Workflow	Surface & Contour Extraction; Color & Opacity Mapping; View & Camera Control; Scientific Insight Derivation
materials	Others	Scalar Field	Task	Data Subsetting & Extraction; View & Camera Control
mhd-magfield_streamribbon	Physics	Vector Field	Workflow	Field Computation; Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control
mhd-turbulence_pathline	Physics	Time-varying; Vector Field	Workflow	Temporal Processing; Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control

Case Name	Domain	Type	Level	Operations
mhd-turbulence_pathribbon	Physics	Time-varying; Vector Field	Workflow	Temporal Processing; Glyph & Marker Placement; Data Smoothing & Filtering; Geometric & Topological Transformation; Color & Opacity Mapping; View & Camera Control
mhd-turbulence_streamline	Physics	Vector Field	Workflow	Field Computation; Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control
miranda	Physics	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping; View & Camera Control
ml_dvr	Others	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping; View & Camera Control
ml_isosurface	Others	Scalar Field	Task	Surface & Contour Extraction; Color & Opacity Mapping; View & Camera Control
ml_slice-contour	Others	Scalar Field	Task	Data Subsetting & Extraction; Surface & Contour Extraction; Color & Opacity Mapping; View & Camera Control
molecule	Chemistry; Mathematics	Scalar Field	Workflow	Field Computation; Feature Identification & Segmentation
moons	Mathematics	Scalar Field	Workflow	Field Computation; Feature Identification & Segmentation; Data Smoothing & Filtering
mri-ventricles	Medical Science	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
mri-woman	Medical Science	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
mrt-angio	Medical Science	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
neocortical-layer-1-axons	Biology	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
noisy-terrain	Mathematics	Scalar Field	Workflow	Field Computation; Feature Identification & Segmentation; Data Smoothing & Filtering
nucleon	Physics	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
ocean	Earth System Science; Mathematics	Tensor Field	Workflow	Field Computation; Feature Identification & Segmentation; Scientific Insight Derivation
pancreas	Medical Science	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
points_surf-clip	Others	Scalar Field	Task	Data Subsetting & Extraction; Glyph & Marker Placement; View & Camera Control
QMCPACK	Physics; Mathematics	Scalar Field	Workflow	Field Computation; Feature Identification & Segmentation; Scientific Insight Derivation
ras-raf-membrane	Biology	Scalar Field	Workflow	Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control; Scientific Insight Derivation
Richtmyer	Physics	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping; View & Camera Control
rings_import-gltf	Others	Scalar Field	Task	Dataset Restructuring; View & Camera Control
rotstrat	Physics	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping; View & Camera Control
rti-velocity_glyph	Physics	Vector Field	Workflow	Data Subsetting & Extraction; Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control
rti-velocity_slices	Physics	Vector Field	Workflow	Data Subsetting & Extraction; Color & Opacity Mapping; View & Camera Control
rti-velocity_streakline	Physics	Time-varying; Vector Field	Workflow	Temporal Processing; Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control
shrink-sphere	Others	Scalar Field	Task	Geometric & Topological Transformation; View & Camera Control
silicium	Physics	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
skull	Medical Science	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
solar-plume	Astronomy	Scalar Field	Workflow	Volume Rendering; Color & Opacity Mapping; View & Camera Control
statue-leg	Others	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
stent	Medical Science	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
stream-glyph	Others	Vector Field	Task	Field Computation; Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control
supernova	Astronomy	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
supernova_isosurface	Astronomy	Scalar Field	Workflow	Surface & Contour Extraction; Data Subsetting & Extraction; Color & Opacity Mapping; View & Camera Control
supernova_streamline	Astronomy	Vector Field	Workflow	Field Computation; Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control
tgc-velocity_contour	Physics	Vector Field	Workflow	Surface & Contour Extraction; Field Computation; Color & Opacity Mapping; View & Camera Control
Tangaroa_streamribbon	Physics	Vector Field	Workflow	Field Computation; Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control
tooth	Medical Science	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
tooth_isosurface	Biology	Scalar Field	Task	Surface & Contour Extraction; View & Camera Control
tornado_dvr	Earth System Science	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
tornado_glyph-streamline	Earth System Science	Vector Field	Workflow	Field Computation; Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control
trajectory-inspection	Biology	Time-varying	Workflow	Temporal Processing; Plot & Chart Generation; View & Camera Control; Scientific Insight Derivation
transparent-bg	Others	Scalar Field	Task	View & Camera Control
two-swirls_streamribbon	Physics	Vector Field	Workflow	Field Computation; Glyph & Marker Placement; Color & Opacity Mapping; View & Camera Control
vis-male	Medical Science	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping
vortex	Physics	Scalar Field	Task	Volume Rendering; Color & Opacity Mapping; View & Camera Control
wavelet_camera	Others	Scalar Field	Task	View & Camera Control
wavelet_chart	Others	Scalar Field	Task	Plot & Chart Generation; Volume Rendering; Color & Opacity Mapping; View & Camera Control
wavelet_colormap	Others	Scalar Field	Task	Color & Opacity Mapping; View & Camera Control
wavelet_export-gltf	Others	Scalar Field	Task	Surface & Contour Extraction; Dataset Restructuring; View & Camera Control
wavelet_export-ply	Others	Scalar Field	Task	Surface & Contour Extraction; Dataset Restructuring
wavelet_histogram	Others	Scalar Field	Task	Plot & Chart Generation; Color & Opacity Mapping; View & Camera Control

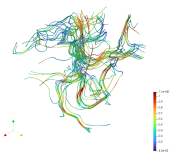
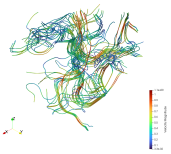
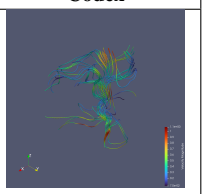
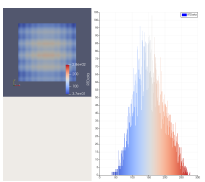
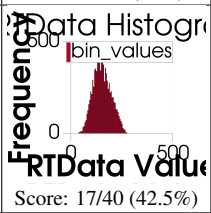
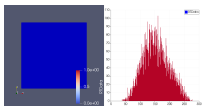
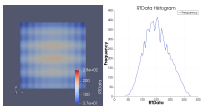
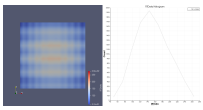
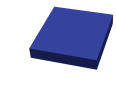
Table 2: Benchmark performance across four task suites of SciVisAgentBench using **GPT-5.2** as the LLM judge for evaluation. Scores and completion rates are reported as mean±std across three repeated trials. Topology visualization results are excluded because they use rule-based evaluation and do not require an LLM judge.

Task Suite	Setting	Overall Score ↑	Completion Rate ↑	pass@1 ↑	pass@2 ↑	pass@3 ↑	pass^1 ↑	pass^2 ↑	pass^3 ↑
ParaView Visualization	ChatVis+GPT-5.2	32.77±0.25	47.92±2.08	0.389	0.569	0.688	0.389	0.208	0.146
	ChatVis+Claude-Sonnet-4.5	39.08±3.50	56.25±4.66	0.479	0.562	0.604	0.479	0.396	0.354
	ParaView-MCP+GPT-5.2	27.47±4.12	51.39±6.70	0.299	0.368	0.417	0.299	0.229	0.208
	ParaView-MCP+Claude-Sonnet-4.5	29.80±9.46	59.72±15.07	0.285	0.389	0.479	0.285	0.181	0.167
	Claude-Code+Claude-Sonnet-4.5	**64.87±0.83**	**98.61±2.41**	**0.778**	0.861	0.896	**0.778**	**0.694**	**0.646**
	Codex+GPT-5.2	63.13±1.34	97.22±2.41	**0.778**	**0.910**	**0.985**	**0.778**	0.646	0.562
Molecular Visualization	GMX-VMD-MCP+GPT-5.2	34.10±15.59	56.41±8.88	0.513	0.667	0.769	0.513	0.359	0.308
	GMX-VMD-MCP+Claude-Sonnet-4.5	58.27±5.46	92.31±7.69	0.821	0.872	0.923	0.821	0.769	0.769
	Claude-Code+Claude-Sonnet-4.5	64.47±8.41	94.87±4.44	**0.923**	**0.974**	**1.000**	**0.923**	**0.872**	**0.846**
	Codex+GPT-5.2	**69.27±4.21**	**97.44±4.44**	0.897	**0.974**	**1.000**	0.897	0.821	0.769
Bioimage Visualization	BioImage-Agent+GPT-5.2	**61.47±14.89**	81.82±0.00	**0.667**	**0.788**	**0.818**	**0.667**	0.545	0.455
	BioImage-Agent+Claude-Sonnet-4.5	56.47±8.45	81.82±0.00	0.606	0.697	0.727	0.606	0.515	0.455
	Claude-Code+Claude-Sonnet-4.5	53.23±7.60	**90.91±9.09**	**0.667**	0.697	0.727	**0.667**	**0.636**	**0.636**
	Codex+GPT-5.2	43.83±7.15	78.79±13.89	0.576	0.727	**0.818**	0.576	0.424	0.364
Object Identification	ParaView-MCP+GPT-5.2	32.43±8.13	60.49±14.02	**0.420**	**0.654**	**0.778**	**0.420**	0.185	0.074
	ParaView-MCP+Claude-Sonnet-4.5	42.77±1.53	**92.59±3.70**	0.259	0.358	0.407	0.259	0.160	0.111
	Claude-Code+Claude-Sonnet-4.5	44.43±2.30	88.89±6.42	0.358	0.556	0.741	0.358	0.160	**0.148**
	Codex+GPT-5.2	**44.50±4.66**	92.59±7.41	**0.420**	0.580	0.630	**0.420**	**0.259**	**0.148**

Table 3: Benchmark performance of Claude Code on the bioimage visualization task suite of SciVisAgentBench using **GPT-5.2** as the LLM judge for evaluation. Scores and completion rates are reported as mean±std across three repeated trials.

Setting	Overall Score ↑	Completion Rate ↑	pass@1 ↑	pass@2 ↑	pass@3 ↑	pass^1 ↑	pass^2 ↑	pass^3 ↑
Claude-Code+Sonnet-4.5 (without skill)	53.23±7.60	90.91±9.09	0.667	0.697	0.727	0.667	0.636	0.636
Claude-Code+Sonnet-4.5 (with skill)	59.50±3.54	**96.97±5.25**	0.727	0.848	0.909	0.727	0.606	0.545
Claude-Code+Opus-4.6 (without skill)	57.57±5.16	93.94±5.25	0.697	0.758	0.818	0.697	0.636	0.636
Claude-Code+Opus-4.6 (with skill)	**70.90±3.31**	93.94±5.25	**0.848**	**0.879**	**0.909**	**0.848**	**0.818**	**0.818**

Table 4: Representative qualitative outcomes for ParaView visualization tasks. Each row compares the expert-defined ground truth with outputs from ParaView-MCP, ChatVis, Claude Code, and Codex. Scores below the agent outputs are vision-based scores from LLM judges, computed by summing the goal-level scores reported in Table 5. These scores are only part of the final benchmark scores.

Case Name	Ground Truth	ParaView-MCP	ChatVis	Claude Code	Codex
`mhd-turbulence_streamline`		Score: 8/40 (20%)	Score: 0/40 (0%)	 Score: 35/40 (87.5%)	 Score: 24/40 (60%)
`render-histogram`		 Score: 17/40 (42.5%)	 Score: 25/40 (62.5%)	 Score: 29/40 (72.5%)	 Score: 28/40 (70%)
`lobster`		 Score: 0/30 (0%)	Score: 6/30 (20%)	Score: 19/30 (63.3%)	Score: 17/30 (56.7%)

D Vision-Based Outcome Evaluation Examples with LLM Judges

We provide three representative qualitative examples to make the benchmark scores more interpretable and to address concerns that aggregate numerical metrics alone may obscure scientific failure modes. For each case, we compare the expert-defined ground-truth visualization with outputs from different agents and report the vision-based score assigned by the LLM judge, including its goal-level rubric breakdown. These

Table 5: Task descriptions, goal-level evaluation criteria, and per-goal vision-based rubric scores for the examples in Table 4. Each case lists the benchmark instruction, the scientific or visualization property assessed by each goal, and the corresponding scores for ParaView-MCP, ChatVis, Claude Code, and Codex. The total row sums all vision-based goal scores and explains the score shown below each image in Table 4. This total is only one component of the final benchmark score.

Case Name & Task Description	Goal	Criterion	ParaView-MCP	ChatVis	Claude Code	Codex
`mhd-turbulence_streamline` Load the 128^3 VTI MHD turbulence velocity field; generate 3D streamlines from a z-axis line source at x=64, y=64 with 50 seed points; color streamlines by velocity magnitude using the Turbo colormap; apply a Tube filter with radius 0.3; add a "Velocity Magnitude" color bar; use a white background and the specified isometric camera; and save the rendered visualization and ParaView state or script files.	G1	**Overall Visualization Goal:** Does the result match the ground-truth streamline visualization of the MHD turbulence velocity field?	2/10	0/10	9/10	6/10
	G2	**Streamline Patterns:** Do the streamlines show similar flow patterns and structures as the ground truth?	2/10	0/10	9/10	6/10
	G3	**Streamline Coverage:** Is the spatial distribution and density of streamlines similar to the ground truth?	1/10	0/10	9/10	5/10
	G4	**Color Mapping:** Is the color distribution along streamlines visually similar to the ground truth?	3/10	0/10	8/10	7/10
	Total		**8/40**	**0/40**	**35/40**	**24/40**
`render-histogram` Create a wavelet object; render it as an RTData-colored surface with a visible color bar; rescale colors to the data range using the Cool-to-Warm colormap; split the view horizontally; create an RTData histogram on the right using the same colormap; and save one screenshot plus ParaView state or script files.	G1	**Wavelet Visualization:** Is the wavelet object properly rendered with RTData coloring and a visible color bar?	6/10	4/10	9/10	9/10
	G2	**Split View Layout:** Is the view correctly split with the wavelet visualization on the left and histogram on the right?	4/10	10/10	10/10	10/10
	G3	**Histogram Generation:** Is the histogram properly generated from RTData, showing the data distribution?	5/10	8/10	6/10	4/10
	G4	**Color Map Consistency:** Are both the wavelet visualization and histogram using the same Cool-to-Warm color map?	2/10	3/10	4/10	5/10
	Total		**17/40**	**25/40**	**29/40**	**28/40**
`lobster` Load the raw Lobster CT dataset with the correct uint8 type, little-endian byte order, $301 \times 324 \times 56$ extent, and $1 \times 1 \times 1.4$ spacing; extract a specimen-boundary isosurface; color it red-orange; answer the walking-leg multiple-choice question; use a white background, optimal view, no color bar or axes, and the specified camera; and save the visualization, text answer, and ParaView state or script files.	G1	**Overall Goal:** Does the visualization clearly show the structure and details of the lobster?	0/10	3/10	5/10	4/10
	G2	**Boundary Clarity:** Are the surface details and boundaries of the lobster well-defined?	0/10	3/10	6/10	6/10
	G3	**Correct Color:** Does the color of the lobster mimic a real lobster, i.e., red-orange?	0/10	0/10	8/10	7/10
	Total		**0/30**	**6/30**	**19/30**	**17/30**

scores are not the final benchmark scores reported in the main result tables. The final benchmark score also incorporates deterministic components, such as output-compliance checks, token usage, execution time, multiple-choice answers, and other case-specific validators, as applicable. These examples show that differences in the vision-based component are not merely low-level image-similarity artifacts: they often correspond to concrete visualization or scientific errors, such as missing flow structures, incorrect chart construction, weak boundary extraction, or mismatched visual encodings.

Claude Sonnet 4.5 is the primary agent model for ParaView-MCP, ChatVis, and Claude Code, whereas Codex uses GPT-5.2. GPT-5.2 evaluates all outcomes shown in this appendix. In Tables 4 and 5, each vision-based score is the sum of the per-goal rubric scores for that case; each goal is scored on a 0-10 scale, and the denominator is the maximum possible score for all applicable goals. The full set of all 108 cases and outcomes from all tested agents is available on the project page.

E Further Analysis on Coding Agents with Skills

Coding agents often fail due to frictions in tool setup and use. In the main benchmark test, we provided no additional tool instructions to obtain an unbiased baseline. However, such frictions can be mitigated by using agent skills [6]: lightweight instruction documents with working code snippets and tool-specific guidance that help agents focus on the task rather than environment exploration.

We conduct a focused follow-up study on bioimage visualization tasks, in which coding agents underperformed purpose-built MCP-based agents. To preserve benchmark integrity, the skill was developed without exposing the agent to benchmark tasks. Instead, we asked the agent to inspect the napari MCP agent, distill its capabilities into a reusable napari skill, and test the skill in the provided environment. The resulting skill captured practical implementation details, version-specific constraints, and common failure-avoidance patterns.

For a direct comparison, we evaluated Claude Code with and without the skill using Sonnet-4.5 and Opus-4.6. GPT-5.2 was used as the judge to avoid using Opus-4.6 both as the task agent and evaluator. As shown in Table 6, skills improve performance for both models and reduce token consumption by roughly half. Opus-4.6 consistently achieves stronger results with fewer tokens, suggesting that stronger models can solve the task more efficiently with fewer interaction steps. Full bioimage results are provided in Table 3.

The results show that Claude Code with napari skill achieved a score of 70.9 compared to 57.6 without skill, a 23% improvement, while simultaneously reducing token consumption by approximately 50%. This suggests that the most effective architecture for SciVis agents may center on flexible reasoning engines paired with modular, domain-specific knowledge that provides tool-use patterns without constraining exploratory capabilities. A broader follow-up study, SciVisAgentSkills [4], further extends this analysis to reusable skills across multiple SciVis tools and task suites.

Table 6: Claude Code results on the bioimage visualization task suite of SciVisAgentBench with different backbone models and skill settings. **GPT-5.2** is used as the LLM judge for evaluation, as Opus-4.6 is used as the task agent. Values are mean±std across three repeated trials.

Setting	Overall Score ↑	Completion Rate ↑	Input Tokens ↓	Output Tokens ↓
Claude-Code+Sonnet-4.5 (without skill)	53.23±7.60	90.91±9.09	8.60M±0.17M	125.66K±2.43K
Claude-Code+Sonnet-4.5 (with skill)	59.50±3.54	**96.97±5.25**	4.31M±3.83M	53.14K±41.02K
Claude-Code+Opus-4.6 (without skill)	57.57±5.16	93.94±5.25	6.67M±1.17M	97.98K±16.93K
Claude-Code+Opus-4.6 (with skill)	**70.90±3.31**	93.94±5.25	**2.67M±0.05M**	**44.81K±0.69K**