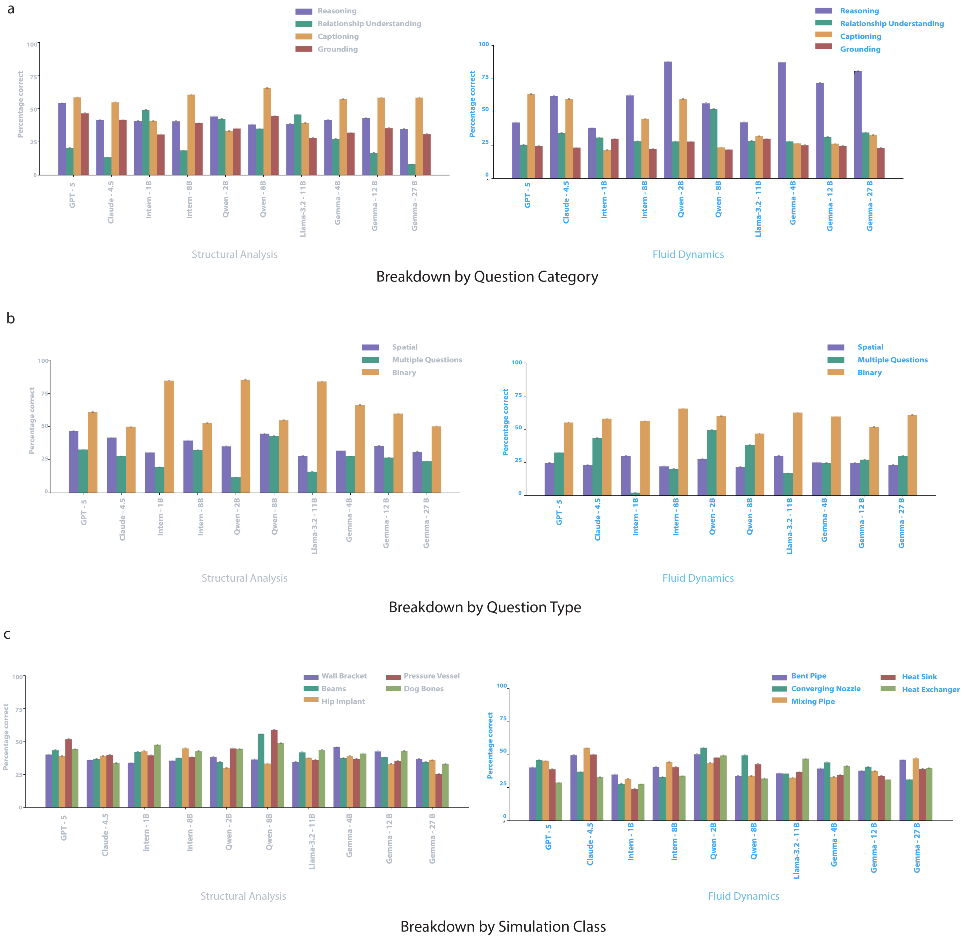
OpenSeeSimE: A Large Benchmark to Assess Vision-Language Model Question Answering Capabilities in Engineering Simulations

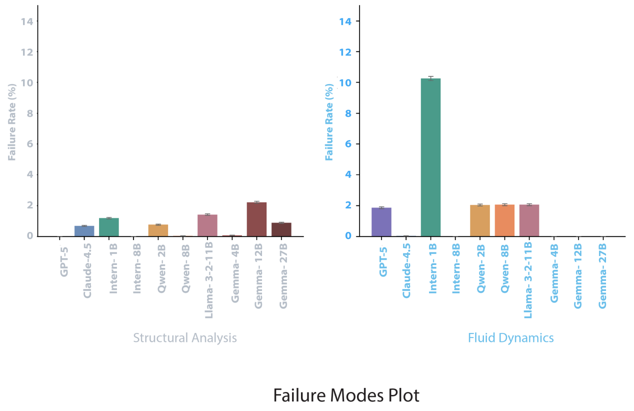
Jessica Ezemba[1*], Jason Pohl[1], Conrad Tucker[1], Christopher McComb[1]

[1*]Department of Mechanical Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, 15213, PA, USA.

*Corresponding author(s). E-mail(s): jezemba@andrew.cmu.edu;
Contributing authors: jpohl@andrew.cmu.edu; conradt@andrew.cmu.edu; ccm@andrew.cmu.edu;



Supplementary Figure 1 Benchmark accuracy by (a) question category for structural analysis and fluid dynamics analysis showing performance across captioning, reasoning, grounding, and relationship understanding tasks. (b) Benchmark accuracy by question type for structural analysis and fluid dynamics analysis comparing binary classification, multiple-choice reasoning, and spatial grounding performance. (c) Benchmark accuracy by simulation class for structural analysis across Wall Bracket, Beams, Hip Implant, Pressure Vessel, and Dog Bone configurations, and fluid dynamics analysis across Bent Pipe, Converging Nozzle, Mixing Pipe, Heat Sink, and Heat Exchanger configurations.



Supplementary Figure 2 Failure rates for structural analysis and fluid dynamics analysis showing model reliability in providing valid responses across all question types

Supplementary Table 1 Flagship model configurations (evaluated on 10% subset for both images and videos). Temperature 0.0 indicates deterministic sampling; higher temperatures follow official model deployment recommendations.

Model	Model Identifier	Max Tokens	Temperature
GPT-5	`gpt-5-2025-08-07`	4096	–[a]
Qwen3-VL-235B	`Qwen3-VL-235B-A22B-Instruct`	4096	0.7[b]
InternVL-3.5-241B	`internvl3.5-241b-a28b`	4096	0.0
Gemini-2.5-Flash	`gemini-2.5-flash`	4096	0.0

[a] GPT-5 uses reasoning effort: minimal, text verbosity: medium

[b] Qwen3-VL-235B uses top-p: 0.8, rate limit: 2s between requests

Supplementary Table 2 Video-specific model configurations (evaluated on video subset only, 32 frames uniformly sampled).

Model	Model Identifier	Frames	Max Tokens	Temp
GPT-5	`gpt-5-2025-08-07`	32	4096	–
Qwen3-VL-8B	`Qwen/Qwen3-VL-8B-Instruct`	32	4096	0.0
InternVL-3.5-8B	`OpenGVLab/InternVL3_5-8B-Instruct`	32	4096	0.0
Gemma-3-12B	`google/gemma-3-12b-it`	32	4096	0.0

Supplementary Table 3 Image-only model configurations (evaluated on complete image dataset). All local models use bfloat16 precision with `device_map="auto"` and `do_sample=False` except where noted.

Model	Model Identifier	Max Tokens	Temp
Qwen3-VL-2B	`Qwen/Qwen3-VL-2B-Instruct`	4096	0.7[a]
Qwen3-VL-8B	`Qwen/Qwen3-VL-8B-Instruct`	4096	0.0
InternVL-3.5-1B	`OpenGVLab/InternVL3_5-1B-Instruct`	4096	0.0
InternVL-3.5-8B	`OpenGVLab/InternVL3_5-8B-Instruct`	4096	0.0
Gemma-3-4B	`google/gemma-3-4b-it`	4096	0.0
Gemma-3-12B	`google/gemma-3-12b-it`	4096	0.0
Gemma-3-27B	`google/gemma-3-27b-it`	4096	0.0
LLaMA-3.2-11B	`meta-llama/Llama-3.2-11B-Vision-Instruct`	4096	0.0

[a] Qwen3-VL-2B uses `do_sample=True` per official guidelines

Supplementary Table 4 Statistical Significance Analysis for Fluid Dynamics Domain. One-tailed binomial test results (alternative: greater) comparing model performance against chance-level accuracy (50% for binary classification, 25% for multiple-choice reasoning and spatial grounding tasks). The test evaluates whether model accuracy significantly exceeds random guessing. The table reports observed accuracies, uncorrected exact p-values, Benjamini-Hochberg (BH) corrected p-values, significance markers, and sample sizes (correct responses/total questions) for all vision-language models across three task categories. Benjamini-Hochberg correction was applied to control the False Discovery Rate across 30 multiple comparisons (10 models × 3 question types). Significance levels: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; ns = not significant. All tests employed $\alpha = 0.05$ with FDR control.

Model	**Task**	**Acc (%)**	***p*-value**	***p*-value (BH)**	**Sig**	**Correct**	**Total**
GPT-5	Binary	55.1	2.11×10^{-47}	4.88×10^{-47}	***	10817	19616
	Multiple Q	32.5	6.30×10^{-123}	2.10×10^{-122}	***	6364	19578
	Spatial	24.5	0.868	1.000	ns	2438	9943
Claude-4.5	Binary	58.0	8.32×10^{-112}	2.50×10^{-111}	***	11377	19617
	Multiple Q	43.3	$< 10^{-300}$	$< 10^{-300}$	***	8486	19581
	Spatial	23.2	1.000	1.000	ns	2308	9943
Intern-1B	Binary	56.0	1.80×10^{-64}	4.91×10^{-64}	***	10992	19617
	Multiple Q	2.2	1.000	1.000	ns	429	19581
	Spatial	29.8	4.68×10^{-28}	9.35×10^{-28}	***	2967	9943
Intern-8B	Binary	65.5	$< 10^{-300}$	$< 10^{-300}$	***	12858	19617
	Multiple Q	20.2	1.000	1.000	ns	3957	19581
	Spatial	22.0	1.000	1.000	ns	2192	9943
Qwen-2B	Binary	59.9	8.59×10^{-170}	3.68×10^{-169}	***	11746	19617
	Multiple Q	49.7	$< 10^{-300}$	$< 10^{-300}$	***	9725	19581
	Spatial	27.7	3.74×10^{-10}	6.61×10^{-10}	***	2755	9943
Qwen-8B	Binary	46.7	1.000	1.000	ns	9161	19617
	Multiple Q	38.3	$< 10^{-300}$	$< 10^{-300}$	***	7497	19581
	Spatial	21.7	1.000	1.000	ns	2160	9943
Llama-3.2-11B	Binary	62.5	1.67×10^{-273}	1.00×10^{-272}	***	12269	19617
	Multiple Q	16.9	1.000	1.000	ns	3305	19581
	Spatial	29.8	4.68×10^{-28}	9.35×10^{-28}	***	2967	9943
Gemma-4B	Binary	59.6	2.33×10^{-160}	8.73×10^{-160}	***	11691	19617
	Multiple Q	24.6	0.924	1.000	ns	4809	19581
	Spatial	24.9	0.557	0.879	ns	2480	9943
Gemma-12B	Binary	51.8	1.99×10^{-7}	3.32×10^{-7}	***	10164	19617
	Multiple Q	27.0	6.83×10^{-11}	1.28×10^{-10}	***	5288	19581
	Spatial	24.4	0.926	1.000	ns	2424	9943
Gemma-27B	Binary	60.8	2.47×10^{-204}	1.24×10^{-203}	***	11935	19617
	Multiple Q	29.9	4.69×10^{-55}	1.17×10^{-54}	***	5859	19581
	Spatial	22.9	1.000	1.000	ns	2281	9943

Supplementary Table 5 Statistical Significance Analysis for Structural Analysis Domain. One-tailed binomial test results (alternative: greater) comparing model performance against chance-level accuracy (50% for binary classification, 25% for multiple-choice reasoning and spatial grounding tasks). The test evaluates whether model accuracy significantly exceeds random guessing. The table reports observed accuracies, uncorrected exact p-values, Benjamini-Hochberg (BH) corrected p-values, significance markers, and sample sizes (correct responses/total questions) for all vision-language models across three task categories. Benjamini-Hochberg correction was applied to control the False Discovery Rate across 30 multiple comparisons (10 models × 3 question types). Significance levels: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; ns = not significant. All tests employed $\alpha = 0.05$ with FDR control.

Model	**Task**	**Acc (%)**	***p*-value**	***p*-value (BH)**	**Sig**	**Correct**	**Total**
GPT-5	Binary	61.1	1.35×10^{-168}	3.68×10^{-168}	***	9458	15488
	Multiple Q	32.7	2.05×10^{-170}	6.16×10^{-170}	***	8446	25811
	Spatial	46.5	$< 10^{-300}$	$< 10^{-300}$	***	4756	10232
Claude-4.5	Binary	49.9	0.583	0.672	ns	7732	15489
	Multiple Q	27.7	2.32×10^{-23}	3.49×10^{-23}	***	7149	25811
	Spatial	41.7	1.37×10^{-297}	5.12×10^{-297}	***	4268	10240
Intern-1B	Binary	84.7	$< 10^{-300}$	$< 10^{-300}$	***	13116	15489
	Multiple Q	19.5	1.000	1.000	ns	5024	25811
	Spatial	30.5	7.79×10^{-37}	1.30×10^{-36}	***	3126	10240
Intern-8B	Binary	52.6	3.76×10^{-11}	5.13×10^{-11}	***	8150	15489
	Multiple Q	32.3	1.19×10^{-152}	2.98×10^{-152}	***	8335	25811
	Spatial	39.4	2.00×10^{-225}	6.67×10^{-225}	***	4036	10240
Qwen-2B	Binary	85.3	$< 10^{-300}$	$< 10^{-300}$	***	13215	15489
	Multiple Q	11.8	1.000	1.000	ns	3051	25811
	Spatial	35.1	3.93×10^{-114}	7.86×10^{-114}	***	3592	10240
Qwen-8B	Binary	54.8	8.61×10^{-33}	1.36×10^{-32}	***	8483	15489
	Multiple Q	42.8	$< 10^{-300}$	$< 10^{-300}$	***	11056	25811
	Spatial	44.6	$< 10^{-300}$	$< 10^{-300}$	***	4568	10240
Llama-3.2-11B	Binary	83.9	$< 10^{-300}$	$< 10^{-300}$	***	12997	15489
	Multiple Q	16.0	1.000	1.000	ns	4134	25811
	Spatial	27.8	6.86×10^{-11}	8.95×10^{-11}	***	2845	10240
Gemma-4B	Binary	66.4	$< 10^{-300}$	$< 10^{-300}$	***	10284	15489
	Multiple Q	27.6	3.71×10^{-22}	5.30×10^{-22}	***	7129	25811
	Spatial	31.9	1.17×10^{-55}	2.19×10^{-55}	***	3266	10240
Gemma-12B	Binary	59.8	6.94×10^{-132}	1.60×10^{-131}	***	9259	15489
	Multiple Q	26.6	2.90×10^{-9}	3.62×10^{-9}	***	6861	25811
	Spatial	35.3	1.17×10^{-119}	2.51×10^{-119}	***	3618	10240
Gemma-27B	Binary	50.2	0.321	0.385	ns	7774	15489
	Multiple Q	23.8	1.000	1.000	ns	6150	25811
	Spatial	30.8	1.16×10^{-39}	2.05×10^{-39}	***	3149	10240

Supplementary Table 6 Practical Significance Analysis for Fluid Dynamics Domain. Cohen's h effect sizes measuring the magnitude of performance differences from chance-level baselines (50% for binary classification, 25% for multiple-choice reasoning and spatial grounding tasks). Effect size categories follow Cohen's conventional benchmarks: Negligible ($|h| < 0.20$), Small ($0.20 \leq |h| < 0.50$), Medium ($0.50 \leq |h| < 0.80$), Large ($|h| \geq 0.80$). The table reports observed accuracies, percentage point differences from chance (Diff), Cohen's h values, effect size categories, and 95% confidence intervals for all vision-language models across three task categories.

Model	Task	Acc	Diff	Cohen's h	Effect	95% CI
GPT-5	Binary	55.1%	+5.1pp	0.103	Negligible	[54.4%, 55.8%]
	Multiple Q	32.5%	+7.5pp	0.166	Negligible	[31.9%, 33.2%]
	Spatial	24.5%	−0.5pp	−0.011	Negligible	[23.7%, 25.4%]
Claude-4.5	Binary	58.0%	+8.0pp	0.161	Negligible	[57.3%, 58.7%]
	Multiple Q	43.3%	+18.3pp	0.390	Small	[42.6%, 44.0%]
	Spatial	23.2%	−1.8pp	−0.042	Negligible	[22.4%, 24.1%]
Intern-1B	Binary	56.0%	+6.0pp	0.121	Negligible	[55.3%, 56.7%]
	Multiple Q	2.2%	−22.8pp	−0.750	Medium	[2.0%, 2.4%]
	Spatial	29.8%	+4.8pp	0.109	Negligible	[28.9%, 30.7%]
Intern-8B	Binary	65.5%	+15.5pp	0.316	Small	[64.9%, 66.2%]
	Multiple Q	20.2%	−4.8pp	−0.115	Negligible	[19.7%, 20.8%]
	Spatial	22.0%	−3.0pp	−0.070	Negligible	[21.2%, 22.9%]
Qwen-2B	Binary	59.9%	+9.9pp	0.199	Negligible	[59.2%, 60.6%]
	Multiple Q	49.7%	+24.7pp	0.517	Medium	[49.0%, 50.4%]
	Spatial	27.7%	+2.7pp	0.061	Negligible	[26.8%, 28.6%]
Qwen-8B	Binary	46.7%	−3.3pp	−0.066	Negligible	[46.0%, 47.4%]
	Multiple Q	38.3%	+13.3pp	0.287	Small	[37.6%, 39.0%]
	Spatial	21.7%	−3.3pp	−0.077	Negligible	[20.9%, 22.5%]
Llama-3.2-11B	Binary	62.5%	+12.5pp	0.254	Small	[61.9%, 63.2%]
	Multiple Q	16.9%	−8.1pp	−0.200	Small	[16.4%, 17.4%]
	Spatial	29.8%	+4.8pp	0.109	Negligible	[28.9%, 30.7%]
Gemma-4B	Binary	59.6%	+9.6pp	0.193	Negligible	[58.9%, 60.3%]
	Multiple Q	24.6%	−0.4pp	−0.010	Negligible	[24.0%, 25.2%]
	Spatial	24.9%	−0.1pp	−0.001	Negligible	[24.1%, 25.8%]
Gemma-12B	Binary	51.8%	+1.8pp	0.036	Negligible	[51.1%, 52.5%]
	Multiple Q	27.0%	+2.0pp	0.046	Negligible	[26.4%, 27.6%]
	Spatial	24.4%	−0.6pp	−0.014	Negligible	[23.5%, 25.2%]
Gemma-27B	Binary	60.8%	+10.8pp	0.219	Small	[60.2%, 61.5%]
	Multiple Q	29.9%	+4.9pp	0.110	Negligible	[29.3%, 30.6%]
	Spatial	22.9%	−2.1pp	−0.048	Negligible	[22.1%, 23.8%]

Supplementary Table 7 Practical Significance Analysis for Structural Analysis Domain. Cohen's h effect sizes measuring the magnitude of performance differences from chance-level baselines (50% for binary classification, 25% for multiple-choice reasoning and spatial grounding tasks). Effect size categories follow Cohen's conventional benchmarks: Negligible ($|h| < 0.20$), Small ($0.20 \leq |h| < 0.50$), Medium ($0.50 \leq |h| < 0.80$), Large ($|h| \geq 0.80$). The table reports observed accuracies, percentage point differences from chance (Diff), Cohen's h values, effect size categories, and 95% confidence intervals for all vision-language models across three task categories.

Model	Task	Acc	Diff	Cohen's h	Effect	95% CI
GPT-5	Binary	61.1%	+11.1pp	0.223	Small	[60.3%, 61.8%]
	Multiple Q	32.7%	+7.7pp	0.171	Negligible	[32.2%, 33.3%]
	Spatial	46.5%	+21.5pp	0.453	Small	[45.5%, 47.4%]
Claude-4.5	Binary	49.9%	−0.1pp	−0.002	Negligible	[49.1%, 50.7%]
	Multiple Q	27.7%	+2.7pp	0.061	Negligible	[27.2%, 28.2%]
	Spatial	41.7%	+16.7pp	0.356	Small	[40.7%, 42.6%]
Intern-1B	Binary	84.7%	+34.7pp	0.766	Medium	[84.1%, 85.2%]
	Multiple Q	19.5%	−5.5pp	−0.133	Negligible	[19.0%, 20.0%]
	Spatial	30.5%	+5.5pp	0.124	Negligible	[29.6%, 31.4%]
Intern-8B	Binary	52.6%	+2.6pp	0.052	Negligible	[51.8%, 53.4%]
	Multiple Q	32.3%	+7.3pp	0.162	Negligible	[31.7%, 32.9%]
	Spatial	39.4%	+14.4pp	0.310	Small	[38.5%, 40.4%]
Qwen-2B	Binary	85.3%	+35.3pp	0.784	Medium	[84.8%, 85.9%]
	Multiple Q	11.8%	−13.2pp	−0.345	Small	[11.4%, 12.2%]
	Spatial	35.1%	+10.1pp	0.221	Small	[34.2%, 36.0%]
Qwen-8B	Binary	54.8%	+4.8pp	0.096	Negligible	[54.0%, 55.6%]
	Multiple Q	42.8%	+17.8pp	0.380	Small	[42.2%, 43.4%]
	Spatial	44.6%	+19.6pp	0.416	Small	[43.6%, 45.6%]
Llama-3.2-11B	Binary	83.9%	+33.9pp	0.745	Medium	[83.3%, 84.5%]
	Multiple Q	16.0%	−9.0pp	−0.224	Small	[15.6%, 16.5%]
	Spatial	27.8%	+2.8pp	0.063	Negligible	[26.9%, 28.7%]
Gemma-4B	Binary	66.4%	+16.4pp	0.334	Small	[65.6%, 67.1%]
	Multiple Q	27.6%	+2.6pp	0.060	Negligible	[27.1%, 28.2%]
	Spatial	31.9%	+6.9pp	0.153	Negligible	[31.0%, 32.8%]
Gemma-12B	Binary	59.8%	+9.8pp	0.197	Negligible	[59.0%, 60.5%]
	Multiple Q	26.6%	+1.6pp	0.036	Negligible	[26.0%, 27.1%]
	Spatial	35.3%	+10.3pp	0.226	Small	[34.4%, 36.3%]
Gemma-27B	Binary	50.2%	+0.2pp	0.004	Negligible	[49.4%, 51.0%]
	Multiple Q	23.8%	−1.2pp	−0.027	Negligible	[23.3%, 24.4%]
	Spatial	30.8%	+5.8pp	0.128	Negligible	[29.9%, 31.7%]

1 Supplementary Note 1: Automated Ground Truth Extraction Protocols

1.1 Selecting Simulation Examples

Engineering simulation benchmarks traditionally rely on limited, manually curated datasets that may not capture the full diversity of real-world engineering applications. To address this limitation, we developed a systematic approach for generating a comprehensive dataset of approximately 10,000 simulation examples through parametric variation of established simulation models. With 10 questions per simulation instance per domain (20 total per instance), this generates over 200,000 total question-answer pairs across both domains.

Our simulation examples were sourced from publicly available `Ansys` Tutorial files, which provide validated baseline configurations with proper boundary conditions and convergence settings. From the extensive tutorial library, we selected base models using a structured selection framework based on three primary criteria designed to maximize dataset diversity and benchmark coverage.

1. Parametric Variability: Base models were selected based on their capacity for meaningful geometric and boundary condition variations. Each selected simulation contained multiple adjustable parameters that could generate distinct simulation outcomes while maintaining physical validity. This approach captured the range of configurations rather than relying on static, single-configuration examples.
2. Simulation Type Coverage: Models were chosen to represent the full spectrum of simulation categories required by our visual question-answering benchmark. This systematic selection ensured comprehensive coverage of essential engineering phenomena including turbulence modeling and structural failure modes across diverse geometric configurations and loading conditions.
3. Representative Engineering Applications: Selected simulations span diverse engineering domains to ensure our benchmark reflects real-world analysis scenarios that practicing engineers encounter across different industries and applications.

For each base simulation, we implemented parametric design automation using Ansys Python interfaces (`PyMechanical`, `PyFluent`, and `PyGeometry`) and list generation software (`MATLAB`) to systematically vary five critical parameters encompassing geometric dimensions, boundary conditions, and material properties. Parameter ranges were established by expert designers to ensure all generated variations remained within physically meaningful bounds while maximizing solution diversity. Each parameter had 4 values chosen, linearly spaced between 2 ”extreme” cases, generating small, small-medium, large-medium, and large values that created substantial variations in each output instance. These 4 values across 5 parameters generated 1,024 unique simulation instances per base model, with parameter settings generated using systematic looping to create each unique set of conditions.

In the mechanical models, the geometry parameters led to different stress concentrations and loading conditions. The boundary condition parameters (changes in axial or bending load forces) and material property parameters (changes in physical characteristics) produced distinct stress, strain, and displacement effects. In the fluid models, the geometry parameters created unique turbulence regions and flow regimes. The boundary condition parameters (fluid velocity) and material property parameters (viscosity) affected the velocity, pressure, and turbulence results.

1.1.1 Structural Analysis Models:

The *Dog Bone* specimen represents standard tensile testing configurations with stress concentrations at the reduced cross-section, requiring interpretation of von Mises stress distributions and failure prediction across varying geometries and loading conditions. The *Hip Implant* model simulates complex biomedical loading with combined axial and bending stresses, presenting challenging stress visualization patterns around irregular geometries. The *Pressure Vessel* involves internal pressure loading creating circumferential and axial stress fields with material-dependent responses. The *Beams* utilize mechanical loading, requiring analysis of stress patterns and material property variations for different beam profiles. The *Wall Bracket* features complex three-dimensional stress distributions under bending loads with stress concentrations at geometric transitions.

1.1.2 Fluid Dynamics Models:

The *Bent Pipe* generates complex flow patterns and pressure losses with varying turbulence intensities dependent on bend geometry and flow conditions. The *Converging Nozzle* creates acceleration zones with pressure gradients and potential flow separation requiring analysis of velocity vector fields and pressure contours. The *Mixing Pipe* involves multi-stream interactions with complex velocity and pressure patterns at the junction. The *Heat Sink* and *Heat Exchanger* models generate intricate flow patterns around fin geometries with heat transfer effects, creating complex visualization challenges involving velocity vectors and pressure fields that vary with geometric and boundary condition parameters.

1.2 Automated Ground Truth Extraction Infrastructure

The automated ground truth extraction system operates through direct programmatic interfaces to simulation software, bypassing visual interpretation entirely. For fluid dynamics simulations, we employ PyFluent's solver session interface to export field data through Ansys Fluent's Text User Interface (TUI) commands. All fluid simulations utilize three-dimensional representations with Cartesian coordinate systems, extracting velocity components (x-velocity, y-velocity, z-velocity), pressure fields, temperature distributions, and Mach numbers where applicable. Data exports generate ASCII-formatted files containing nodal coordinates and corresponding field values, with file sizes typically ranging from hundreds of kilobytes to several megabytes depending on mesh density.

For structural analysis, PyMechanical provides access to finite element results through Ansys Mechanical's scripting interface. The system extracts von Mises stress tensors, displacement vectors, strain components, and temperature fields at nodal locations. Each extraction preserves spatial coordinate information (X, Y, Z positions) alongside field values, enabling subsequent geometric analysis for symmetry detection and spatial localization tasks. Session management follows a single-instance paradigm where each simulation case file loads once and serves all question extraction procedures sequentially. This approach minimizes computational overhead from repeated file loading operations while maintaining consistency in visualization parameters across questions sharing common data requirements. Fluent sessions initialize with double precision arithmetic and utilize multiple processor cores for parallel data extraction operations.

1.3 Statistical Analysis Procedures

Questions requiring identification of extreme values or aggregate statistics operate directly on extracted field arrays using standard numerical operations. The system loads relevant data files into pandas DataFrame structures, validates data quality through finite value checks (excluding NaN and infinite values), and applies appropriate statistical functions. For maximum and minimum value queries, the system employs NumPy's optimized array operations to identify extrema with computational complexity linear in the number of data points. Relative magnitude assessments, such as determining whether values span one, two, or three orders of magnitude, compute the ratio between maximum and minimum field values. The system applies the following classification scheme: ratios below 10 indicate less than one order of magnitude, ratios between 10 and 100 represent one to two orders, ratios between 100 and 1000 span two to three orders, and ratios exceeding 1000 encompass more than three orders of magnitude. For fields containing negative values, the system employs alternative ratio calculations based on absolute value ranges to ensure meaningful magnitude comparisons.

1.4 Distribution Analysis Implementation

1.4.1 Structural Uniformity Assessment

Stress distribution uniformity analysis employs coefficient of variation (CV) as the primary metric, defined as the ratio of standard deviation to mean value. The system extracts von Mises stress values across all nodes, computes statistical measures on the resulting distribution, and applies a uniformity threshold of CV ≤ 0.2 (20% coefficient of variation). The system requires a minimum of three data points for meaningful statistical analysis, rejecting datasets below this threshold. Distribution uniformity extends beyond simple variance measures to incorporate spatial considerations. The system validates that extracted stress values span the entire geometric domain rather than representing localized clusters, ensuring that uniformity assessments reflect global distribution characteristics rather than sampling artifacts.

1.4.2 Fluid Stagnation Zone Detection

Dead zone identification in fluid dynamics requires determining regions where flow velocity falls below thresholds indicating effective stagnation. The system applies a velocity magnitude threshold of 1×10^{-6} (one micron per second), representing a value several orders of magnitude below typical flow velocities that effectively indicates numerical zero in the context of engineering simulations. For each node in the extracted velocity field, the system classifies velocities below this threshold as stagnant, computing the fraction of total nodes meeting this criterion. Binary classification as yes/no for dead zone presence depends on whether any significant fraction of the domain exhibits stagnant flow characteristics. The system employs a conservative approach where even small percentages of stagnant nodes (above negligible numerical noise levels) trigger affirmative classification, acknowledging that engineering significance of dead zones relates more to their presence than their spatial extent.

1.5 Symmetry Analysis Protocols

Symmetry detection requires assessing whether field distributions exhibit mirror invariance about specified coordinate planes. The system implements a comprehensive symmetry analysis procedure applicable to both structural deformation patterns and fluid flow fields. For each candidate symmetry plane (X-plane, Y-plane, or Z-plane), the system first determines the plane's spatial location by computing the midpoint of the geometric domain along the relevant axis. It then generates mirrored coordinate sets by reflecting each node's position across this plane. Using scipy's cdist function, the system computes Euclidean distances between original and mirrored coordinate sets, identifying symmetric node pairs where spatial separation falls below a coordinate matching tolerance of 1×10^{-3} (one millimeter). For each identified symmetric pair, the system compares field values through relative difference calculations: $|\text{val}_1 - \text{val}_2| / \max(|\text{val}_1|, |\text{val}_2|, 10^{-10})$, where the denominator's small constant prevents division by zero for near-zero values. The system applies a base symmetry tolerance of 5% for value comparisons, though certain structural analysis questions employ a relaxed 10% tolerance to accommodate numerical solution variability in finite element results. Classification as symmetric requires that at least 90% of identified node pairs exhibit value differences within the specified tolerance for fluid dynamics questions, while structural analysis employs a 60% threshold reflecting the greater solution variability inherent in solid mechanics computations. The system evaluates symmetry about all three coordinate planes independently, classifying overall symmetry based on whether any single plane meets the criteria (questions asking ”Is the pattern symmetric?”) or identifying which specific plane demonstrates the strongest symmetry (questions asking ”What is the axis of symmetry?”).

1.6 Physics-Based Classification Methods

1.6.1 Flow Regime Characterization

Mach number analysis categorizes flow speed relative to the local speed of sound, employing standard aerospace engineering classification criteria. The system first attempts to extract Mach number fields directly from simulation results when available. For simulations lacking explicit Mach data, the system computes Mach numbers from velocity magnitude fields by dividing by the appropriate speed of sound: 343.0 m/s for air at standard conditions (20°C, 1 atmosphere) or 1482.0 m/s for water at 20°C. Flow regime classification operates on maximum Mach numbers rather than domain-averaged values, recognizing that localized supersonic regions may exist within predominantly subsonic flows. The system applies the following thresholds: maximum Mach below 0.8 classifies as subsonic, maximum Mach between 0.8 and 1.2 indicates transonic flow, and maximum Mach exceeding 1.2 designates supersonic conditions.

1.6.2 Flow Direction Analysis

Dominant flow direction determination analyzes mean absolute velocity magnitudes across spatial dimensions. For three-dimensional simulations, the system extracts velocity component fields (x-velocity, y-velocity, z-velocity), computes the mean of absolute values for each component independently, and identifies which component exhibits the largest mean magnitude. This approach correctly handles flows with significant reverse components, where signed mean values would artificially reduce apparent flow strength. Classification as "complex multidirectional" rather than dominant along a single axis employs a tolerance-based criterion. The system computes the mean of all component means and checks whether each individual component mean falls within 5% of this global mean. When all components satisfy this proximity criterion, the flow exhibits insufficient directional bias for classification as dominant along any single axis. Otherwise, the component with maximum mean absolute velocity determines the dominant direction.

1.6.3 Stress Type Classification

Structural analysis questions requiring classification of dominant stress types (bending, shear, axial, or torsion) extract relevant stress tensor components and compare their magnitudes according to solid mechanics principles. The system analyzes stress distributions in critical regions, typically identified as zones exhibiting maximum von Mises stress or maximum deformation magnitude. Classification criteria derive from examining ratios between normal stress components, shear stress components, and their spatial gradients, though specific implementation details vary by geometry and loading conditions.

1.6.4 Deformation Direction Analysis

Significant deformation direction identification follows analogous procedures to flow direction analysis, extracting displacement components (X-displacement, Y-displacement, Z-displacement) and computing mean absolute magnitudes. The system identifies whether deformation primarily occurs along a single coordinate axis or exhibits complex multi-directional character through the same tolerance-based comparison used for fluid flow analysis. An additional classification distinguishes between in-plane and out-of-plane deformation patterns for planar structural geometries, computed through relative magnitude comparisons between displacement components parallel and perpendicular to the structure's primary plane (that is user defined).

1.6.5 Tensile Stress Predominance

Determining whether stresses are predominantly tensile examines the signs of extracted normal stress values. The system counts nodes exhibiting positive (tensile) versus negative (compressive) stress values, classifying the pattern as predominantly tensile when positive values outnumber negative values. This simple criterion suffices for binary classification while avoiding arbitrary threshold definitions for mixed stress states.

1.7 Spatial Localization and Region Labeling

Region-based grounding questions require generating visualizations with labeled locations and determining which region contains specified target features. This process involves three distinct phases: target identification from numerical data, region generation on rendered visualizations, and ground truth determination through spatial proximity calculations.

1.7.1 Visualization Generation

The system generates standardized visualizations through direct control of simulation software rendering parameters. For fluid dynamics, PyFluent's graphics object interface sets contour and vector plot properties, camera positions, and color mapping schemes. Structural analysis employs PyMechanical's result visualization controls to configure stress or displacement contour plots with consistent color schemes. All visualizations render at 1920×1440 pixel resolution to ensure sufficient detail for spatial localization tasks while maintaining consistent aspect ratios across instances. View orientations follow standardized definitions: front, back, left, right, top, and bottom views align camera positions with principal axes, while isometric views employ 45-degree elevation and azimuth angles. The system saves rendered images as PNG files with lossless compression, preserving color fidelity essential for subsequent region labeling operations.

1.7.2 Legend and Text Detection

Before placing region labels, the system must identify areas to avoid to prevent obscuring critical information or overlapping with existing annotations. We employ `EasyOCR` with English language models to detect text regions within generated visualizations, applying a confidence threshold of 0.3 to filter spurious detections. Detected text regions receive padding of 30 pixels on all sides to ensure labels maintain readable separation. Legend detection specifically identifies color bars and their associated numerical labels through pattern matching on scientific notation text. The system searches for text strings matching the regular expression pattern `-?\d+\.\d*e[+-]?\d+`, representing floating-point numbers in exponential format commonly used for engineering field values. When multiple scientific notation strings appear vertically or horizontally aligned within 30 pixels, the system groups them as belonging to the same legend bar. Valid legend groups require at least two numerical labels to avoid false positives from isolated exponential notation. Once identified, the system estimates the spatial extent of each legend by computing bounding boxes around detected text groups, extending 100 pixels to the left (the typical colorbar width) and 40 pixels above and below the text cluster. An additional safety margin applies a 20×20 pixel dilation kernel to create buffer zones around all detected legend areas, ensuring robust separation between labels and legends even when initial detection boundaries prove imprecise. Additional avoidance regions include image borders (30 pixels from edges), very light areas (RGB values exceeding 240 on 0-255 scale, indicating white background), and very dark areas (RGB values below 20, indicating black background or unlabeled regions). For structural visualizations, the system also detects and avoids axis indicators—small colored arrows or text typically rendered in pure red, green, or blue that denote coordinate system orientation. These indicators occupy areas between 50 and 2000 pixels, with 40-pixel padding applied around each detected indicator.

1.7.3 Color Gradient Analysis

Engineering visualization standards employ rainbow color gradients mapping from red (maximum values) through yellow and green to blue (minimum values). The system generates a reference gradient containing 20 discrete color steps spanning this spectrum through RGB interpolation. For fluid dynamics, the gradient represents flow field magnitudes; for structural analysis, it represents stress or displacement magnitudes. To classify any pixel in the visualization as belonging to the simulation color scheme versus background or annotation elements, the system computes Euclidean distances in RGB space between the pixel's color and all reference gradient colors. Colors falling within a tolerance of 80 Euclidean distance units (on a 0-255 RGB scale) from any reference gradient color classify as simulation colors; colors exceeding this threshold classify as background or annotation elements. This tolerance accommodates rendering antialiasing and color interpolation artifacts while maintaining sufficient specificity to distinguish simulation data from interface elements.

1.7.4 Region Point Selection

The system selects four points (A, B, C, D) for region labeling through a constrained random sampling procedure that ensures spatial distribution, simulation color association, and sufficient mutual separation. Starting from the set of all pixels classified as simulation colors and not falling within legend, text, or border avoidance masks, the system randomly shuffles candidate positions and iteratively selects points meeting the following criteria:

- The pixel color must fall within the simulation color tolerance (80-unit Euclidean distance from the gradient)
- Spatial separation from all previously selected points must exceed 50 pixels initially
- If fewer than four points satisfy the initial constraint after exhaustive search, the system relaxes the spatial separation requirement to 20 pixels and repeats selection

This approach balances the competing objectives of spatial distribution (ensuring labels span the visualization domain rather than clustering) and color diversity (ensuring labels correspond to meaningful field value ranges rather than uniform regions). The relaxed spatial constraint accommodates visualizations where simulation colors occupy relatively small portions of the image domain due to large legends or extensive background areas. For each selected point, the system determines its color gradient level by identifying which of the 20 reference gradient colors exhibits minimum Euclidean distance in RGB space. Higher gradient indices correspond to colors closer to red (representing higher field values), while lower indices correspond to colors closer to blue (representing lower field values). This mapping enables subsequent label assignment based on relative color intensities.

1.7.5 Label Assignment and Ground Truth Determination

For questions asking about maximum value locations, the system assigns labels such that at least one of the four selected points exhibits a local maximum color gradient level among the four points. The point with the highest gradient level receives the ground truth label, while remaining points receive alternative labels in descending order of their gradient levels. This scheme ensures that selecting the "reddest" region among the four labeled options yields correct answers, but critically, it does not guarantee that any labeled region corresponds to the global maximum across the entire visualization domain. Conversely, for questions asking about minimum value locations, label assignment proceeds in reverse order, with the point exhibiting the lowest gradient level receiving the ground truth label. This bidirectional assignment strategy prevents models from learning simple heuristics such as "always choose the reddest region" or "always choose the bluest region" across different question types. The label assignment approach deliberately introduces variability in absolute color intensities of correct answers across different instances and viewing orientations. In isometric views where maximum stress concentrations may appear edge-on or obscured, the labeled point nearest to the numerical maximum location may not exhibit the deepest red coloring in the visualization. Similarly, certain viewing angles may render minimum value regions larger or smaller depending on three-dimensional geometry. This variability ensures that successful localization depends on spatial reasoning about field distributions rather than simple color intensity comparisons among labeled regions.

1.7.6 Validation and Consistency Checks

After generating labeled visualizations and determining ground truth, the system performs consistency validation by verifying that numerical target locations extracted from simulation data correspond spatially to assigned ground truth regions. For maximum value questions, the system computes the Euclidean distance between the coordinates exhibiting maximum field magnitude and the pixel coordinates of each labeled region, confirming that the minimum distance corresponds to the region designated as ground truth. When validation fails—typically due to extreme viewing angles rendering target locations outside the visible domain or due to numerical precision issues in coordinate transformations between three-dimensional simulation space and two-dimensional image space—the system flags the instance for manual review or regenerates the visualization with alternative camera parameters. Across the complete benchmark dataset, validation failure rates remain below 2%, occurring primarily in cases where maximum values concentrate at geometric features (corners, edges) that project to image boundaries in certain viewing orientations.

1.8 OCR and Image Processing Parameters

All image processing operations employ standardized parameters derived from extensive testing across diverse simulation visualizations. The `EasyOCR` reader initializes with English language models and processes images at their native 1920×1440 resolution without downsampling. Text detection employs a confidence threshold of 0.3, representing a balance between capturing legitimate text elements (which typically exhibit confidence scores above 0.5) and avoiding false positives from visual artifacts or simulation features that superficially resemble text. Color distance calculations throughout the system employ Euclidean metrics in RGB space: $d = \sqrt{(R_1 - R_2)^2 + (G_1 - G_2)^2 + (B_1 - B_2)^2}$. While perceptually uniform color spaces such as CIELAB might provide more accurate color similarity measures, RGB Euclidean distance suffices for the relatively saturated rainbow gradients employed in engineering visualizations and avoids computational overhead from color space transformations. Grayscale detection identifies pixels where maximum channel differences fall below 30 units on the 0-255 scale: $\max(|R-G|, |G-B|, |B-R|) < 30$. This criterion successfully distinguishes achromatic background elements, text, and annotations from chromatic simulation data across diverse visualization styles while accommodating subtle color casts that may arise from rendering antialiasing.

1.9 Computational Efficiency and Scalability

The automated extraction system processes complete question sets (10 questions spanning multiple visualization orientations and field variables) for a single simulation case in approximately 5-15 minutes on standard workstation hardware, depending on mesh density and complexity of required analyses. This represents a 100-fold improvement in throughput compared to manual expert annotation while eliminating subjective variability inherent in human interpretation of complex visualizations. Session reuse constitutes the primary efficiency optimization, avoiding repeated file loading and solver initialization overhead. Secondary optimizations include vectorized array operations through NumPy for statistical calculations and batched visualization generation for questions sharing common rendering parameters. The system's architecture supports straightforward parallelization across multiple simulation cases, enabling scalable dataset generation limited only by available computational resources rather than human expert availability. All threshold values, tolerance parameters, and computational procedures remain consistent across the entire benchmark dataset, ensuring that ground truth quality depends on implementation fidelity rather than annotator expertise.

2 Supplementary Note 2: Complete Evaluation Specifications and Protocols

2.0.1 System Prompt

The following system prompt was used across all models without modification:

```
"You are a visual question answering assistant. You MUST follow
     this exact format:\n\n"
"FORMAT REQUIREMENTS:\n"
"Line 1: Copy the EXACT answer text from the provided options (
    word-for-word, including all symbols)\n"
"Line 2: One brief explanation sentence (10-15 words)\n\n"
"CRITICAL RULES:\n"
"1. The first line MUST be an EXACT COPY of one option - do not
     paraphrase or summarize\n"
"2. Copy ALL words, punctuation, and mathematical symbols
    exactly as shown in the option\n"
"3. Do NOT add phrases like ’The answer is’ or explanatory text
     on line 1\n"
"4. Do NOT shorten or reword long options - copy them
    completely\n\n"
"EXAMPLE 1 (Simple):\n"
"Question: Is the sky blue?\n"
"Options: Yes, No\n"
"CORRECT:\n"
"Yes\n"
"The clear atmosphere scatters blue wavelengths effectively.\n\
    n"
"EXAMPLE 2 (Complex option with symbols):\n"
"Question: What is the range?\n"
"Options: Less than 10x min, More than 1000x min\n"
"CORRECT:\n"
"More than 1000x min\n"
"The values span from 7 billion to 1.6 trillion.\n\n"
"INCORRECT:\n"
"More than three orders of magnitude\n"
"(This paraphrases instead of copying the exact option)\n\n"
"Remember: Line 1 = EXACT COPY of option. Line 2 = explanation
    ."
```

2.0.2 User Prompt Template

For each question instance, the following template format was used:

```
prompt += "Instructions:\n"
prompt += "1. First line: Provide ONLY your answer exactly as
    it appears in the options above (e.g., ’A’, ’Yes’, ’X axis’,
     etc.). Do NOT add any other text on this line.\n"
prompt += "2. Second line onwards: Provide a brief summary (1-2
     sentences) explaining your reasoning.\n\n"
prompt += "Answer:"
```

For video inputs, no prompt modifications were applied beyond the standard template. Both image and video modalities received identical prompting to enable direct performance comparison.

2.0.3 Rationale

The two-line structured output format addresses two critical evaluation requirements: (1) enabling automated answer extraction through simple line-based parsing without requiring complex natural language interpretation of model responses, and (2) requiring models to provide reasoning justification for post-hoc error analysis. Pilot testing revealed that models frequently paraphrased answer options or embedded answers within explanatory text when using free-form prompts, creating ambiguity in correctness determination. The strict format requirements with explicit positive and negative examples eliminate this source of evaluation error while maintaining consistency across diverse model architectures and deployment methods.

2.1 Model Configurations

Supplementary tables 1, 2, and 3 present complete configuration parameters for all evaluated models. All parameters remained fixed across the entire evaluation to ensure reproducibility.

Temperature 0.0 configurations enforce deterministic sampling for reproducibility, while non-zero temperatures (Qwen models) follow official deployment guidelines specifying optimal operating points for visual reasoning tasks.

2.2 Video Processing Specifications

2.2.1 Source Video Characteristics

Original simulation videos were generated with domain-specific parameters:

1. **Structural Analysis:** 200 frames at 29 frames/second (7 seconds duration). Maximum deformation occurs at frame 100 (temporal midpoint), after which the simulation reverses to initial state.
2. **Fluid Dynamics:** 200 frames at 40 frames/second (5 seconds duration). Frames represent pathlines showing steady-state flow solution.

All videos rendered at 1920×1440 pixel resolution with H.264 compression, matching static image resolution to ensure consistent visual detail across media types.

2.2.2 Frame Extraction Strategy

Video frame extraction employed middle-frame-centered uniform sampling: for videos with N total frames requiring K extracted frames, the system first selected the middle frame at position $\lfloor N/2 \rfloor$, then sampled $(K-1)/2$ frames before and after this midpoint at uniform intervals. This strategy ensures that structural analysis videos always include the maximum deformation state (which occurs at the temporal midpoint) in the frame set provided to models.

Extracted frames maintained 1920×1440 resolution and saved as PNG with lossless compression before model input.

2.3 Reproducibility Protocols

2.3.1 Dataset Access

The benchmark dataset is available through HuggingFace for both structural analysis (https://huggingface.co/datasets/cmudrc/OpenSeeSimE-Structural) and fluid dynamics (https://huggingface.co/datasets/cmudrc/OpenSeeSimE-Fluid).

2.3.2 Random Seed Configuration

All stochastic components (Python random module, NumPy random number generator, PyTorch CUDA random number generator) initialized with seed value 42 before evaluation. For models employing non-deterministic sampling (Qwen-235B at temperature 0.7), complete response logs can be requested for exact replication.

2.3.3 Software Environment

Critical dependency versions: Python 3.10, PyTorch 2.1.0 (CUDA 12.8), HuggingFace Transformers 4.36.0, HuggingFace Datasets 2.16.0, OpenCV 4.8.1, OpenAI Python SDK 1.6.1, Anthropic Python SDK 0.8.1, Google Generative AI 0.3.2. Hardware: 2 X NVIDIA 5090 32GB GPUs with single-GPU inference for models $\leq$8B parameters and dual-GPU tensor parallelism for larger models.

2.3.4 Code Availability

Complete evaluation code, configuration files, and documentation are available at `https://github.com/cmudrc/OpenSeeSimE-Full` under MIT License. The repository includes shared utilities for prompt construction and response parsing, checkpoint management infrastructure, and setup instructions.

3 Supplementary Note 3: Complete Question Specifications and Failure Analysis

During experiments we observed that VLMs would produce a variety of noncompliant responses. We categorize these responses into three primary types: explicit refusals (models claim insufficient information despite adequate visual evidence), contradictory reasoning (models generate conflicting analyses without resolution), and purely descriptive responses (models describe observations without completing reasoning to answer).

3.0.1 Explicit Refusals

In fluid dynamics evaluation, models occasionally refuse to answer despite adequate visual information. The most extreme case occurs in InternVL-1B which displayed systematic refusals. The dominant refusal phrase “not directly comparable to the sound speed in water without additional context” appears in these failures, despite images containing sufficient information (velocity values) to perform straightforward Mach number calculations. Additional refusals citing “not specified in the image” occur even when velocity magnitudes are explicitly displayed. This represents a calibration failure where the model refuses to make reasonable inferences from available visual data. In structural analysis, refusal rates remain negligible across all models, suggesting this failure mode is task-specific.

3.0.2 Contradictory/Conflicting Information

Models frequently generate internally inconsistent analyses, particularly in spatial reasoning tasks. In structural analysis, models exhibit contradictory reasoning in 44.7% of None responses, describing spatial features or stress distributions without mapping these observations to required answer choices. Representative examples include statements like "The color red in the color bar and the 'Max' label indicate the highest value, which corresponds to the maximum displacement" or "clear axis of symmetry along the X-axis, as indicated by the symmetrical pattern," but failing to conclude which labeled point (A/B/C/D) corresponds to these observations. In fluid dynamics, contradictory reasoning manifests as factually incorrect assessments that contradict the correct answer. The most prominent pattern occurs in Llama-3.2-11B, where 38.3% of its fluid failures contain the phrase "greater than the speed of sound in water," without specifying transonic or supersonic. These contradictions indicate reasoning failures where models generate contradictory answers rather than merely failing to format answers correctly.

3.0.3 Purely Descriptive Responses

Several models are solely in observation mode, providing detailed descriptions without reasoning to conclusions. This pattern appears predominantly in structural analysis, affecting 60% of None responses in smaller models. Representative responses include "The image shows a 3D stress distribution with a clear axis of symmetry" or "color-coded map representing total deformation, indicating a gradual change across the structure" without identifying requested locations or classifying deformation types. Models provide accurate visual observations but fail to complete the reasoning chain to categorical answers. This failure mode is notably less prevalent in fluid dynamics tasks, suggesting particular difficulty in bridging visual observations to spatial categorical answers in structural mechanics contexts.

3.0.4 Model-Specific Patterns

InternVL-1B demonstrates the most severe and systematic failures, with 5,042 None responses in fluid dynamics (55.9% of all fluid failures across models) driven primarily by explicit refusals. This substantially exceeds other models' failure rates and represents a fundamental limitation in the model's ability to perform basic inferential reasoning from visual data. In structural analysis, the same model shows 84.8% purely descriptive failures, indicating consistent difficulty in completing reasoning chains across both domains. Larger models (GPT-5, Claude-Sonnet-4-5) show minimal non-extraction failures, with most None responses attributable to API infrastructure issues rather than cognitive limitations. Mid-size models (Gemma-12B, Llama-3.2-11B) demonstrate intermediate failure rates with more varied patterns including incomplete analyses and contradictory answers.

3.1 Complete Question Specifications

This section provides the comprehensive question sets employed in OpenSeeSimE. All questions were designed to assess engineering visualization interpretation across two primary analysis domains: structural mechanics and computational fluid dynamics.

3.1.1 Structural Analysis Questions

The structural analysis question set comprises ten questions spanning symmetry detection, stress classification, deformation characterization, and spatial localization tasks.

1. Is the deformation pattern symmetric across any axis?
   - Question Category: Relationship Understanding
   - Question Type: Binary
   - Options: Yes, No
2. Are the stresses predominantly tensile in nature?
   - Question Category: Reasoning
   - Question Type: Binary
   - Options: Yes, No
3. Is the [stress/strain/temperature/pressure] distribution pattern uniform?
   - Question Category: Captioning
   - Question Type: Binary
   - Options: Yes, No
4. The significant deformation in the model is primarily:
   - Question Category: Relationship Understanding
   - Question Type: Multiple Question
   - Options:

   (a) Y axis
   (b) X axis
   (c) Z axis
   (d) Complex multi-directional
5. What is the axis of symmetry?
   - Question Category: Relationship Understanding
   - Question Type: Multiple Question
   - Options:

   (a) X

(b) Y
(c) Z
(d) None/Multiple

6. The dominant stress type in the critical region is:

   - Question Category: Reasoning
   - Question Type: Multiple Question
   - Options:

   (a) Bending dominant
   (b) Shear dominant
   (c) Axial dominant
   (d) Torsion dominant

7. Where is the maximum [displacement/stress/strain/temperature] located in the model?

   - Question Category: Grounding
   - Question Type: Spatial
   - Options: A, B, C, D

8. Where is the minimum [displacement/stress/strain/temperature] located in the model?

   - Question Category: Grounding
   - Question Type: Spatial
   - Options: A, B, C, D

9. What is the approximate range (maximum-minimum) of values in the displayed contour plot?

   - Question Category: Captioning
   - Question Type: Multiple Question
   - Options:

   (a) Less than one order of magnitude (max < 10× min)
   (b) One to two orders of magnitude (10× min < max < 100× min)
   (c) Two to three orders of magnitude (100× min < max < 1000× min)
   (d) More than three orders of magnitude (max > 1000× min)

10. There is primarily what type of deformation?

   - Question Category: Captioning
   - Question Type: Multiple Question
   - Options:

   (a) In-plane deformation
   (b) Out-of-plane deformation
   (c) Complex multi-directional deformation
   (d) No significant deformation

3.1.2 Fluid Analysis Questions

The computational fluid dynamics question set comprises ten questions evaluating flow field interpretation, symmetry detection, flow regime classification, and spatial pattern recognition. These questions assess comprehension of velocity, pressure, and temperature distributions in fluid simulations.

1. Are there dead zones (stagnant flow areas) visible in the simulation?
   - Question Category: Reasoning
   - Question Type: Binary
   - Options: Yes, No
2. Is the flow field symmetric across any axis?
   - Question Category: Captioning
   - Question Type: Binary
   - Options: Yes, No
3. The flow is only in one direction
   - Question Category: Captioning
   - Question Type: Binary
   - Options: Yes, No
4. Are there regions of low velocity and low pressure?
   - Question Category: Relationship Understanding
   - Question Type: Binary
   - Options: Yes, No
5. What is the axis of symmetry?
   - Question Category: Relationship Understanding
   - Question Type: Multiple Question
   - Options:

   (a) X
   (b) Y
   (c) Z
   (d) None/Multiple
6. How would you characterize the flow speed relative to sound speed?
   - Question Category: Reasoning
   - Question Type: Multiple Question
   - Options:

   (a) Subsonic (Mach < 0.8)
   (b) Transonic ($0.8 <$ Mach < 1.2)
   (c) Supersonic (Mach > 1.2)
   (d) N/A

7. Where is the maximum [velocity/pressure/temperature] located in the flow field?
   - Question Category: Grounding
   - Question Type: Spatial
   - Options: A, B, C, D
8. Where is the minimum [velocity/pressure/temperature] located in the flow field?
   - Question Category: Grounding
   - Question Type: Spatial
   - Options: A, B, C, D
9. What is the approximate range (maximum-minimum) of values in the displayed contour plot?
   - Question Category: Captioning
   - Question Type: Multiple Question
   - Options:

   (A) Less than one order of magnitude (max $<$ 10× min)
   (B) One to two orders of magnitude (10× min $<$ max $<$ 100× min)
   (C) Two to three orders of magnitude (100× min $<$ max $<$ 1000× min)
   (D) More than three orders of magnitude (max $>$ 1000× min)
10. Flow is dominantly along which axis (If axis is not visible, select between axial and radial as X and Y respectively)?
    - Question Category: Relationship Understanding
    - Question Type: Multiple Question
    - Options:

    (a) X
    (b) Y
    (c) Z
    (d) Complex Multidirectional

The complete OpenSeeSimE dataset including all simulation instances, question-answer pairs, visualizations, and metadata is publicly available for both structural analysis (https://huggingface.co/datasets/cmudrc/OpenSeeSimE-Structural) and fluid analysis (https://huggingface.co/datasets/cmudrc/OpenSeeSimE-Fluid).