GeoR-Bench: Evaluating Geoscience Visual Reasoning

Anonymous authors
Paper under double-blind review

Abstract

Geoscience intelligence is expected to understand, reason about, and predict Earth-system changes to support human decision-making in critical domains such as disaster response, climate adaptation, and environmental protection. Although current research has shown promising progress on specific geoscience tasks, such as remote-sensing interpretation and geographic question answering, existing benchmarks remain largely task-specific and fail to capture open-ended real-world geoscience problems. As a result, it remains unclear how far current AI systems are from achieving genuine geoscience intelligence. To address this gap, we present **GeoR-Bench**, a Benchmark for evaluating Geoscience visual Reasoning through reasoning-informed visual editing tasks. GeoR-Bench contains 440 curated samples spanning 6 geoscience categories and 24 task types, covering Earth-observation imagery and structured scientific representations such as maps and diagrams. We evaluate outputs along three dimensions, including reasoning, consistency, and quality. Results from 21 closed- and open-source multimodal models reveal that geoscience reasoning remains a critical bottleneck. The highest-performing model achieves 42.7% overall strict accuracy, while the best open-source model reaches 10.3%. Notably, the visual consistency and image quality of the outputs frequently surpass their scientific accuracy. Ultimately, these findings indicate that current models generate superficially plausible results but fail to capture underlying Earth-science processes.

1 Introduction

Geoscience intelligence is the capacity of an AI system to understand, reason about, and predict Earth-system changes in ways that can inform human decision-making in domains such as disaster response, climate adaptation, and environmental protection. Recent multimodal models have shown encouraging progress on narrow geoscience tasks, including remote-sensing interpretation and geographic question answering. Yet a central question remains open: how do we *evaluate* geoscience intelligence in a way that reflects the open-ended, visually grounded nature of real-world Earth-system problems? Existing benchmarks remain largely task-specific and measure isolated recognition or captioning abilities, leaving it unclear how far current models are from genuine geoscience intelligence. In practice, geoscience reasoning must operate over intermediate visual artifacts, including satellite images, maps, cross-sections, and scientific diagrams, and produce outputs that are not merely plausible but scientifically valid. A principled evaluation of geoscience intelligence must therefore probe whether a model can convert Earth-system understanding into a correct, visually grounded generation rather than only verify surface-level visual realism or textual recall.

To make geoscience intelligence measurable, we propose to evaluate it through reasoning-informed visual editing: given an input image and a task instruction, a model must infer the relevant Earth-system state and produce an output image that reflects that inference. This setting is challenging because correctness depends on reasoning about process, geometry, scale, topology, and domain conventions at the same time; predicting delta growth requires understanding sediment transport and channel branching, modifying a cyclone image requires respecting circulation structure, and completing a tectonic diagram requires correct labels and an interpretable layout. Pure recognition benchmarks do not force a model to externalize such understanding in a testable visual artifact, while generic generation benchmarks reward plausibility or instruction following without verifying scientific validity. Reasoning-informed visual editing closes this gap: the edited image becomes

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

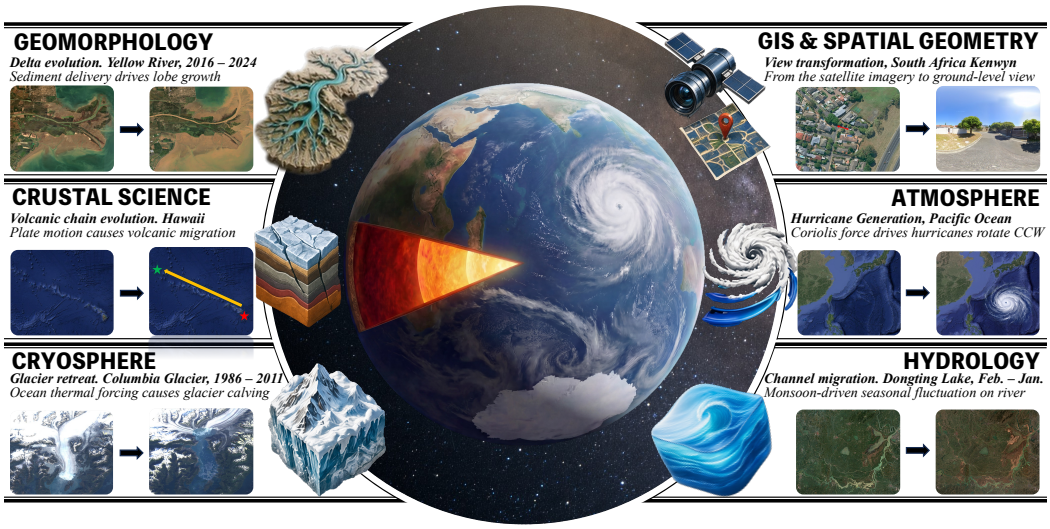


Figure 1: **GeoR-Bench** evaluates geoscience intelligence via reasoning-informed visual editing across six categories: geomorphology, GIS & spatial geometry, crustal science, atmosphere, cryosphere, and hydrology. Each task requires inferring the underlying Earth-system process from an input image and producing a scientifically valid output.

direct evidence of whether the model can turn geoscience reasoning into a grounded visual generation, making editing the measurement interface through which geoscience intelligence becomes observable.

However, existing benchmarks do not directly evaluate this capability. General image-editing benchmarks emphasize natural-image manipulation, instruction following, and broad visual reasoning Ye et al. (2025); Xu et al. (2026). More reasoning-aware suites such as KRIS-Bench and RISEBench move beyond low-level editing quality, but they still focus primarily on natural scenes and generic structural reasoning rather than scientific correctness (Wu et al., 2025b; Zhao et al., 2026). Scientific and multimodal knowledge benchmarks such as MMMU evaluate expert-level understanding (Yue et al., 2024), while scientific illustration benchmarks test technical drawing ability (Chang et al., 2025). These works are important, but they rarely require a model to express domain understanding through a visually grounded geoscience generation. Therefore, in this work, we address this gap by evaluating whether multimodal models can convert geoscience understanding into a correct edited or generated image.

GeoR-Bench is built around this principle and contains 440 samples spanning 6 categories and 24 task types, with 4 task types per category. The benchmark covers geomorphology, hydrology, atmosphere and ocean dynamics, cryosphere science, GIS and spatial geometry, and crustal science. Its tasks span both real-world Earth-observation imagery and formal scientific representations, including river-delta evolution, wetland fluctuation, cyclone generation under Coriolis constraints, snowline prediction, cross-view geospatial reconstruction, earthquake epicenter triangulation, and geological labeling. This breadth is deliberate: geoscience visual reasoning must operate across natural imagery, spatial products, and structured scientific diagrams rather than within a single visual modality. Figure 1 illustrates the benchmark scope.

To evaluate model outputs, we use a three-part protocol to judge the generated images: **reasoning**, **consistency**, and **quality**. Reasoning evaluates whether the generated output reflects the intended physical, spatial, or process-based transformation by jointly comparing the input image, ground-truth output, model prediction, and task-specific rubric. Consistency evaluates whether the generated output preserves the required visual relationship between the input image and target output. Quality evaluates whether the generated image is visually usable as an artifact. We also report *strict accuracy*, which counts a sample as correct only when all three dimensions are satisfied simultaneously. These metrics support claims about benchmark-defined geoscience visual reasoning ability. Figure 4 provides a complementary diagnostic by replacing implicit task descriptions with explicit procedural instructions while holding the input and target fixed.

Our contributions are as follows:

- We articulate a comprehensive evaluation principle for geoscience intelligence: it should be tested through reasoning-informed visual editing tasks that require models to convert Earth-system understanding into scientifically correct visual outputs.
- We introduce GeoR-Bench, a benchmark of 440 curated samples across 6 categories and 24 task types. Each sample follows a standardized format, each task has a dedicated judging rubric, and all evaluative claims are clearly documented.
- We provide experiments and analysis showing that 21 multimodal models remain substantially weaker at geoscience reasoning than at preserving visual appearance or producing readable outputs, yielding evidence about the present limits of multimodal geoscience visual reasoning.

2 Related Work

2.1 Large Language Models and Multimodal Benchmarks

Recent progress in multimodal models has produced a new generation of evaluation benchmarks for image editing, instruction-following generation, and visually grounded reasoning. General-purpose editing benchmarks such as ImgEdit assess instruction alignment, local content preservation, and visual plausibility across diverse natural scenes (Ye et al., 2025), but strong performance there does not imply fidelity to deeper causal or domain constraints. Reasoning-oriented benchmarks including KRIS-Bench and RISEBench show that visually attractive edits often violate temporal, spatial, causal, or logical requirements (Wu et al., 2025b; Zhao et al., 2026), while WorldGenBench tests whether text-to-image systems convert implicit world knowledge into faithful outputs (Zhang et al., 2025), and broader knowledge benchmarks such as MMMU and SrdBench highlight persistent difficulty with expert knowledge and technical visual conventions (Yue et al., 2024; Chang et al., 2025). Yet these benchmarks center on natural scenes or generic knowledge, leaving untested the geoscience-specific combination of Earth-system process, spatial geometry, scale consistency, and disciplinary convention: an output may look plausible while misrepresenting channel branching, cyclone structure, or crustal labels. GeoR-Bench fills this gap by evaluating whether large models can convert domain understanding into scientifically correct visual outputs.

2.2 Geoscience Benchmarks and Evaluation

Within remote sensing and Earth science, existing benchmarks have expanded multimodal evaluation but emphasize different capabilities. One line of work targets visual understanding of Earth-observation imagery: VRSBench benchmarks remote-sensing vision-language models on captioning, grounding, and question answering (Li et al.), while multimodal datasets such as MDAS combine optical, hyperspectral, SAR, DSM, and GIS sources to support richer perception tasks (Hu et al., 2023). A second line moves toward generation and scientific reasoning: MMM-RS extends remote-sensing benchmarking to text-to-image generation across scenes, modalities, weather, and resolutions (Luo et al., 2024), and MSEarth and GRADE curate scientific figures and enriched reasoning content from the literature for multimodal scientific understanding (Zhao et al., 2025; Liu et al., 2026). However, none of these directly tests whether a model can start from an *input* geoscience image, infer the required Earth-system change, and produce an *output* image that remains scientifically correct under task-specific constraints. GeoR-Bench is therefore neither a pure remote-sensing understanding benchmark nor a pure text-to-image generation benchmark. It covers both real-world Earth-observation imagery and formal representations such as maps, cross-sections, and diagrams, and complements prior work by targeting a capability they do not measure.

3 The GeoR-Bench Benchmark

3.1 Principle & Motivation

The central principle of GeoR-Bench is that geoscience visual reasoning should be evaluated through scientific tasks that require both understanding and visual generation. Image editing is not the target

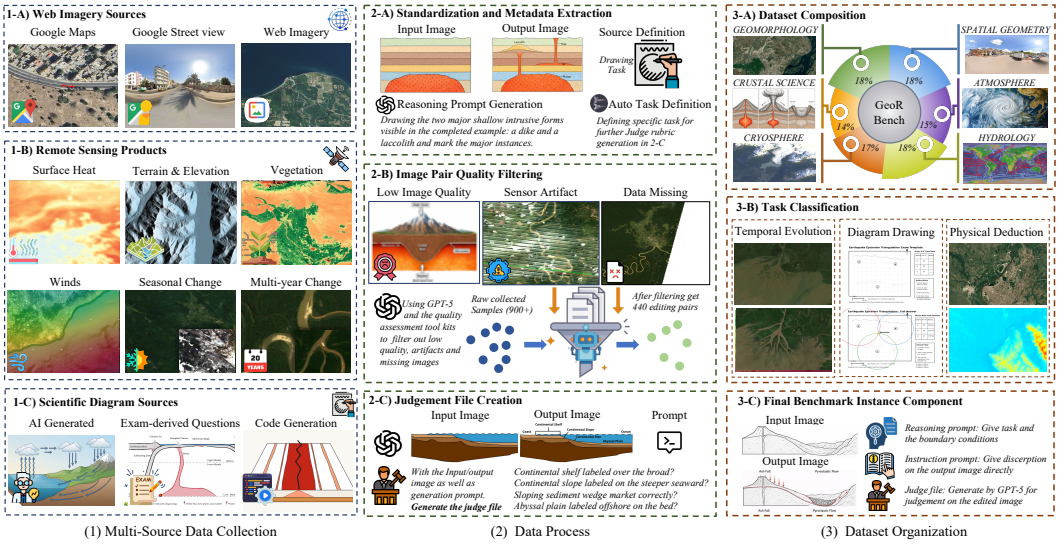


Figure 2: Overview of GeoR-Bench. The benchmark combines geoscience images from web imagery, remote sensing, scientific diagrams, and map products. The raw data are filtered and reorganized into a standardized format for model evaluation. The dataset is balanced across six geoscience categories, each with a dedicated judging rubric.

capability by itself; it is the observable interface through which a model must reveal whether it understands Earth-system structure, process, geometry, and scientific convention. A model succeeds only when it can interpret a geoscience image, infer the required generation, and generate an output image that is scientifically correct, turning image reasoning-and-generation into a benchmark of geoscience reasoning rather than a test of aesthetic plausibility alone.

Geoscience is a particularly demanding target for this framing because many tasks are process-based rather than appearance-based. The model must infer why a delta grows, why a storm rotates, where a snowline should sit, or how a tectonic cross-section should be labeled. All of these tasks are multi-scale, requiring local geometric edits, global spatial consistency, and physically meaningful background preservation within a single example. The domain also spans heterogeneous visual forms, from natural-looking satellite imagery to scientific cross-sections, topographic renderings, and cartographic products, so a benchmark that claims to measure geoscience competence must cover both real-world imagery and formal scientific representations.

This framing motivates GeoR-Bench as a purpose-built probe for geoscience visual reasoning. Existing image-editing benchmarks largely reward photorealism or local fidelity, which leaves a blind spot precisely where Earth-system competence lives: whether a model can read a geoscience scene, reason about the underlying process, and render an output that is scientifically defensible under domain-specific criteria. Our goal is to turn image generation into a measurable act of geoscience reasoning, so that progress on the benchmark translates into progress on the scientific understanding.

3.2 Dataset Construction & Composition

GeoR-Bench is built around a unified sample format while preserving raw sources. The benchmark is primarily composed of domain-curated geoscience images selected for clear visual structure and task suitability, as shown in Figure 2. In practice, these include satellite observations, map products, DEM renderings, scientific cross-sections, and structured geoscience diagrams. The release contains 440 standardized samples spanning 6 categories and 24 task types, organized so that each category contains 4 task types. The release spans the following categories: geomorphology, hydrology, atmosphere and ocean dynamics, cryosphere science, GIS and spatial geometry, and crustal science. This balanced category structure is intentional: the benchmark is designed to cover the breadth of geoscience visual reasoning rather than over-concentrate on any single modality or subfield.

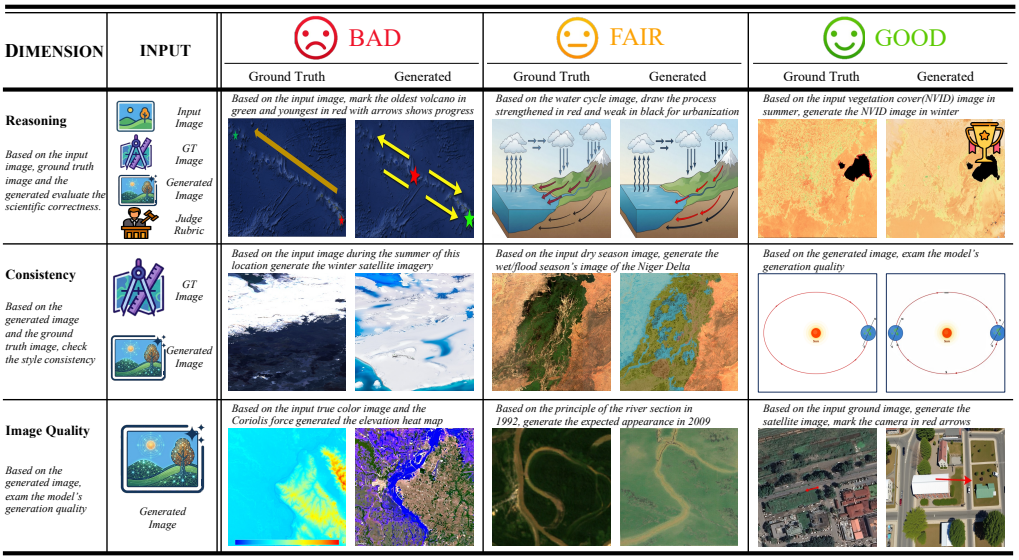


Figure 3: GeoR-Bench evaluation pipeline jointly evaluates reasoning, consistency, and quality. Reasoning uses the input, reference output, model prediction, and judging rubric to evaluate correctness. Consistency assesses the required visual relationship, while quality measures the generated output.

Within this structure, the benchmark includes both real-world Earth-observation tasks and structured scientific-figure tasks. Representative task types include river-delta evolution, oxbow-lake formation, monsoon-driven wetland fluctuation, cyclone generation and transposition, glacier retreat, snowline prediction, crater generation, street-to-satellite and satellite-to-panorama reconstruction, earthquake epicenter triangulation, seismogram interpretation, volcanic plumbing diagrams, and marine-geology labeling. This mixture is deliberate. Geoscience competence requires handling both satellite observations and formal scientific representations.

To evaluate models uniformly across these heterogeneous sources, each sample is reorganized into a common form: an input image, a ground-truth output image, a natural-language prompt, a task-specific judging rubric, and, when available, auxiliary context images and structured metadata. A single benchmark index enumerates all samples, and a unified dataloader exposes each one as a consistent evaluation unit regardless of its raw origin. This standardization makes it possible to mix raw source patterns, such as temporal satellite series, cross-view geospatial data, metadata-driven tasks, and structured figure-editing tasks, within one evaluation interface.

3.3 Evaluation Protocol

Our evaluation pipeline is designed specifically for geoscience image editing and generation, as shown in Figure 3. To ensure generated images are practically useful for Earth-system analysis, our framework assesses outputs through a progression of four interrelated dimensions. First, we assess Reasoning to determine whether the underlying scientific reasoning-informed visual editing is physically and domain-accurately executed. We then measure Consistency to verify that the required spatial, structural, and stylistic relationships with the source input are preserved. Next, we evaluate the foundational Quality of the artifact to ensure it is free of basic visual degradations. Finally, the strict accuracy metric unifies these independent axes and requires simultaneous success across all dimensions.

Reasoning. Reasoning evaluates whether the model performs the intended geoscience generation correctly. For each sample, the evaluator takes the input image, the ground-truth output image, the model prediction, and the task-specific `judge.md` rubric. The rubric defines scoring points aligned with the scientific objective of the task. For example, a delta-evolution task may score whether the delta front progrades seaward and whether distributary channels bifurcate correctly, while an

Table 1: Reasoning and Consistency result across 21 models, with the best score in each column highlighted in **bold** and the second- and third-best scores underlined.

Model	Evaluation Dimensions													
	GEOM		*HYDR*		*ATMO*		*CRYO*		*SPATIAL*		*CRUSTAL*		*OVERALL*	
	R	*C*	*R*	*C*	*R*	*C*	*R*	*C*	*R*	*C*	*R*	*C*	*R*	*C*
Closed Source Models														
FLUX.2 Max (Labs, 2025)	37.2	65.0	57.8	68.8	49.4	82.4	45.3	72.9	47.1	81.0	53.9	59.7	48.5	71.6
FLUX.2 Pro (Labs, 2025)	38.9	56.2	46.4	67.5	42.5	72.1	35.1	51.4	36.1	59.5	56.5	61.3	42.6	61.3
GPT-Image-1.5 (OpenAI, 2025a)	56.9	88.8	74.5	93.8	66.9	97.1	71.3	**97.1**	61.0	92.9	78.2	96.8	68.1	94.4
GPT-Image-2 (OpenAI, 2025b)	**71.7**	**95.0**	88.9	96.2	**72.4**	92.6	**82.4**	95.7	**69.2**	94.0	**90.0**	96.8	**79.1**	95.1
Nano Banana (Google, 2025b)	49.2	90.0	59.9	92.5	42.8	88.2	58.0	94.3	51.1	76.2	63.1	87.1	54.0	88.1
Nano Banana 2 (Google, 2025c)	71.6	85.0	86.9	**100**	69.3	94.1	75.7	94.3	68.7	94.0	85.0	**100**	76.2	94.6
Nano Banana Pro (Google, 2025d)	58.6	91.2	**90.6**	98.8	62.2	97.1	66.7	95.7	65.5	**96.4**	84.8	98.4	71.4	**96.3**
Seedream 4.0 (Chen et al., 2025)	45.0	52.5	36.0	77.5	48.4	80.9	43.3	55.7	22.4	47.6	37.4	35.5	38.7	58.3
Seedream 4.5 (Seedream, 2025)	50.0	68.8	55.1	72.5	45.1	82.4	51.1	81.4	29.0	63.1	48.7	59.7	46.5	71.3
Seedream 5.0 (Seedream, 2026)	50.2	81.2	76.0	88.8	56.5	**98.5**	58.3	94.3	43.1	90.5	71.0	88.7	59.2	90.3
Open Source Models														
Bagel (Deng et al., 2025)	10.1	22.5	16.5	28.8	20.7	20.6	12.3	18.6	10.6	25.0	5.81	21.0	12.7	22.7
Bagel (think) (Deng et al., 2025)	19.8	53.8	21.4	51.2	24.9	45.6	18.1	38.6	17.6	40.5	10.6	37.1	18.7	44.5
DreamOmni (Xia et al., 2025)	26.1	77.5	31.8	65.0	22.5	32.4	28.9	75.7	18.2	54.8	12.3	40.3	23.3	57.6
FLUX.2 dev (Labs, 2025)	33.0	75.0	34.1	80.0	37.4	77.9	41.4	52.9	21.1	44.0	34.4	43.5	33.6	62.2
ICEdit (Zhang et al.)	18.6	42.5	12.1	33.8	21.8	41.2	11.1	37.1	13.3	28.6	2.42	3.23	13.2	31.1
ICEdit (MoE) (Zhang et al.)	14.5	33.8	5.12	5.00	4.41	13.2	5.00	25.7	9.40	29.8	3.87	1.61	7.05	18.2
OmniGen (Xiao et al., 2025)	1.75	6.25	5.00	3.75	16.6	8.82	4.14	10.0	1.19	17.9	0.323	6.45	4.84	8.86
Qwen-Edit-2511 (Wu et al., 2025a)	44.0	60.0	57.9	71.2	44.1	67.6	51.1	68.6	31.2	32.1	65.2	75.8	48.9	62.6
Step-1x (Liu et al., 2025)	31.1	73.8	23.0	60.0	25.3	44.1	31.4	55.7	28.7	41.7	8.55	27.4	24.7	50.4
Step-1x (think) (Liu et al., 2025)	22.6	66.2	24.9	55.0	27.1	42.6	30.9	54.3	23.5	22.6	9.19	35.5	23.0	46.0
Step-1x (think+reflect) (Liu et al., 2025)	23.0	56.2	20.6	57.5	29.0	35.3	24.6	45.7	15.4	27.4	15.5	32.3	21.3	42.4

epicenter task may score whether the circles are drawn at the correct radii and intersect at the correct point. The weighted sum yields a reasoning score in $[0, 1]$, which is then scaled to $[0, 100]$.

Consistency. Consistency evaluates the required visual relationship between the generated output, the input image, and the ground-truth target. For each sample, the evaluator takes the input image, the ground-truth output image, and the model prediction, and measures whether the prediction preserves the visual relationship implied by the task. In style-sensitive tasks, this means preserving the sensor or rendering style of the source while matching the target visual appearance. In localized tasks, consistency emphasizes preserving content outside the instructed edit. In independent-output tasks, consistency verifies that the generated output remains visually compatible with the expected target representation rather than identical in style to the input.

Quality. Quality evaluates the generated output using an image-quality assessment method applied to the model prediction itself. This component measures whether the generated image is visually well formed as an output artifact, independent of whether the underlying geoscience reasoning is correct. It captures degradations such as blur, distortion, artifacts, and other failures that reduce the usability of the generated result.

Strict accuracy. For sample i, using normalized scores, we define

$$\mathrm{Pass}_i = \mathbf{1}\left[R_i = 1 \wedge C_i = 1 \wedge Q_i > 0.7\right], \qquad \mathrm{StrictAcc} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Pass}_i\,. \tag{1}$$

Here, $R_i = 1$ means that every criterion in the task-specific correctness rubric is satisfied, $C_i = 1$ means that every consistency criterion is satisfied, and $Q_i > 0.7$ is the minimum human-readability gate. Correctness and consistency therefore use point-by-point rubric satisfaction rather than tuned continuous thresholds. The rubric evaluates scientific constraints rather than pixel identity, so a scientifically valid prediction may differ visually from the reference output.

4 Experiments

4.1 Experimental Setup

We evaluate 21 image-editing and image-generation systems spanning the current closed- and open-source landscape. The models are categorized into two groups: (1) closed-source models, including Nano Banana Pro (Google, 2025d), Nano Banana 2 (Google, 2025c), Seedream 5.0 (Seedream,

Table 2: Quality and Accuracy result across 21 models, with the best score in each column highlighted in **bold** and the second- and third-best scores underlined.

Model	Evaluation Dimensions													
	GEOM		*HYDR*		*ATMO*		*CRYO*		*SPATIAL*		*CRUSTAL*		*OVERALL*	
	Q	*A*	*Q*	*A*	*Q*	*A*	*Q*	*A*	*Q*	*A*	*Q*	*A*	*Q*	*A*
Closed Source Models														
FLUX.2 Max (Labs, 2025)	69.7	2.50	77.3	10.0	79.3	14.7	74.0	5.71	73.0	19.0	**81.6**	9.68	75.8	10.3
FLUX.2 Pro (Labs, 2025)	77.3	0.00	**79.7**	2.50	80.9	5.88	**77.4**	8.57	**79.4**	7.14	79.8	9.68	**79.1**	5.63
GPT-Image-1.5 (OpenAI, 2025a)	61.1	7.50	71.6	22.5	75.0	23.5	71.7	34.3	71.2	33.3	80.5	35.5	71.9	26.1
GPT-Image-2 (OpenAI, 2025b)	62.3	20.0	72.7	52.5	78.0	**35.3**	70.1	**54.3**	68.8	**52.4**	79.1	41.9	71.8	**42.7**
Nano Banana (Google, 2025b)	61.4	5.00	72.8	22.5	76.2	2.94	69.1	25.7	71.4	11.9	79.0	19.4	71.7	14.6
Nano Banana 2 (Google, 2025c)	69.6	**25.0**	77.3	50.0	78.0	26.5	75.5	40.0	71.6	45.2	80.5	**48.4**	75.4	39.2
Nano Banana Pro (Google, 2025d)	56.9	20.0	75.8	**57.5**	76.0	20.6	71.8	34.3	71.8	35.7	79.7	41.9	72.0	35.0
Seedream 4.0 (Chen et al., 2025)	75.8	5.00	75.5	2.50	80.5	5.88	74.2	8.57	74.1	4.76	81.1	3.23	76.9	4.99
Seedream 4.5 (Seedream, 2025)	77.4	10.0	78.5	17.5	**84.5**	5.88	75.8	20.0	74.9	7.14	81.2	9.68	78.7	11.7
Seedream 5.0 (Seedream, 2026)	69.0	15.0	72.7	32.5	76.8	8.82	74.3	28.6	73.3	16.7	78.4	16.1	74.1	19.6
Open Source Models														
Bagel (Deng et al., 2025)	59.9	0.00	63.8	0.00	66.5	0.00	72.5	2.86	66.5	0.00	68.5	0.00	66.3	0.476
Bagel (think) (Deng et al., 2025)	49.9	0.00	58.7	5.00	68.0	0.00	62.1	2.86	56.8	2.38	66.7	3.23	60.4	2.24
DreamOmni (Xia et al., 2025)	44.3	0.00	62.9	7.50	69.1	0.00	61.8	11.4	66.8	0.00	73.0	0.00	63.0	3.15
FLUX.2 dev (Labs, 2025)	60.0	0.00	67.9	7.50	71.3	5.88	70.2	11.4	61.4	2.38	76.2	3.23	67.8	5.07
ICEdit (Zhang et al.)	49.7	5.00	59.3	0.00	63.9	0.00	60.9	2.86	57.3	0.00	67.7	0.00	59.8	1.31
ICEdit (MoE) (Zhang et al.)	31.6	0.00	15.8	0.00	30.7	0.00	35.9	0.00	31.5	0.00	33.1	0.00	29.8	0.00
OmniGen (Xiao et al., 2025)	60.7	0.00	63.7	0.00	59.2	0.00	58.7	0.00	63.5	0.00	63.1	0.00	61.5	0.00
Qwen-Edit-2511 (Wu et al., 2025a)	**78.6**	5.00	78.1	17.5	80.2	5.88	77.2	17.1	77.4	0.00	77.3	16.1	78.1	10.3
Step-1x (Liu et al., 2025)	53.0	2.50	65.8	2.50	75.5	0.00	69.1	8.57	70.9	7.14	73.4	0.00	68.0	3.45
Step-1x (think) (Liu et al., 2025)	56.6	0.00	64.9	2.50	75.1	2.94	68.4	5.71	71.6	0.00	74.6	0.00	68.5	1.86
Step-1x (think+reflect) (Liu et al., 2025)	56.5	0.00	68.4	5.00	76.3	2.94	66.9	5.71	71.0	2.38	72.4	0.00	68.6	2.67

2026), GPT-Image-1.5 (OpenAI, 2025a), FLUX.2 Max (Labs, 2025), Nano Banana (Google, 2025b), Seedream 4.5 (Seedream, 2025), GPT-Image-2 (OpenAI, 2025b), FLUX.2 Pro (Labs, 2025), and Seedream 4.0 (Chen et al., 2025); and (2) open-source models, including Qwen-Edit-2511 (Wu et al., 2025a), Step-1x (Liu et al., 2025), DreamOmni (Xia et al., 2025), Bagel (Deng et al., 2025), ICEdit (Zhang et al.), FLUX.2 dev (Labs, 2025), and OmniGen (Xiao et al., 2025).

Closed-source models are accessed directly through official APIs, while all open-source models are deployed locally on NVIDIA H200 GPUs using their default inference configurations. Following the GeoR-Bench protocol, each model receives the standardized input image, prompt, and any task-defined context images to generate a single predicted output. These generated outputs are evaluated using Gemini-3-Flash (Google, 2025a) for reasoning and consistency on a normalized 0–100 scale, while Q-Align (Wu et al., 2023) evaluates image quality. The main evaluation metric is strict accuracy, which requires a generated sample to pass all three dimensions simultaneously.

4.2 Main Results

Tables 1 and 2 present the full leaderboard. GPT-Image-2 is the strongest system overall, reaching 42.7% strict accuracy with the highest reasoning score of 79.1 and near-saturated consistency at 95.1. Nano Banana 2 follows at 39.2% strict accuracy, scoring 76.2 in reasoning and 94.6 in consistency. Nano Banana Pro achieves 35.0% accuracy while securing the best overall consistency score of 96.3. No evaluated system exceeds 50% strict accuracy, confirming that benchmark-defined geoscience visual reasoning remains far from solved even for frontier closed-source systems.

The gap to open-source models is large. The best open-source system, Qwen-Edit-2511, reaches 10.3% strict accuracy, roughly four times lower than the top closed-source model. Five of the eleven open-source systems score below 2%, and OmniGen scores exactly 0%. This spread is far wider than typical natural-image editing benchmarks report, indicating that GeoR-Bench probes capabilities that open-source editing models currently lack.

4.3 Bottleneck Analysis

A consistent pattern emerges from the component decomposition. Quality clusters tightly across the leaderboard. Almost every system lands in the 60 to 80 range, led by FLUX.2 Pro at 79.1 and followed closely by three models within two points. Consistency is more spread out, but the top seven systems all achieve scores above 90. Reasoning, by contrast, spans from 4.8 to 79.1 and covers more than an order of magnitude across the same model set. Reasoning is therefore the dimension

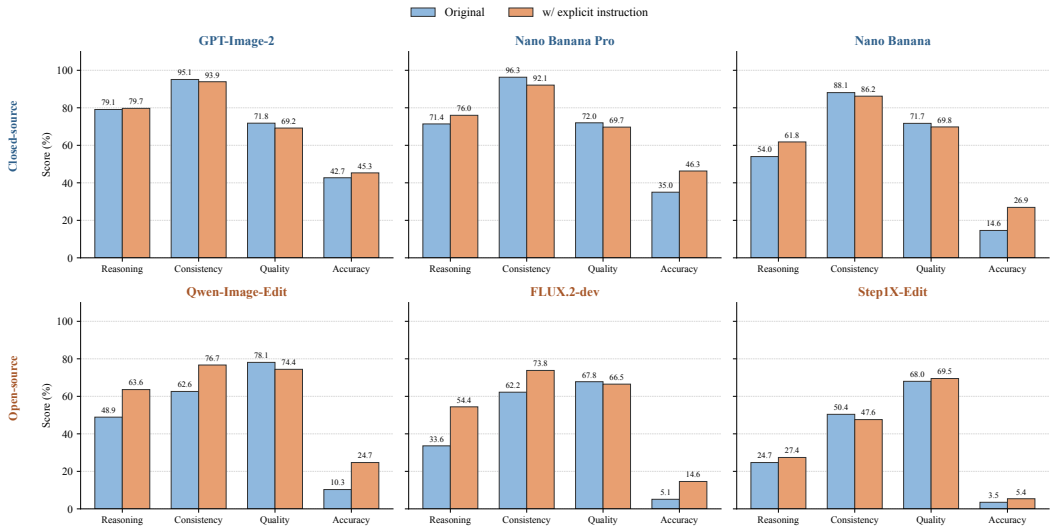


Figure 4: Comparison of model performance across different prompt settings.

that actually discriminates geoscience competence, as the other evaluation components saturate long before it does.

Two models make this concrete. FLUX.2 Pro holds the highest quality score in the table yet produces correct scientific transformations on only 5.6% of samples because its reasoning is mediocre at 42.6 and its consistency is weak at 61.3. Qwen-Edit-2511 reaches a quality level of 78.1 to perform on par with frontier closed source systems and surpass Nano Banana 2, but its reasoning is capped at 48.9, which collapses its strict accuracy down to 10.3%. These are precisely the visually plausible but scientifically wrong failures.

4.4 Category-level Analysis

Category level accuracy in Table 2 reveals strong domain asymmetries. Hydrology is the most tractable slice. Nano Banana Pro reaches 57.5% strict accuracy here, and three closed source systems exceed 50%. This is consistent with hydrologic task types being grounded in recurring temporal satellite series such as oxbow lakes and wetland fluctuation, where strong appearance priors align with correct geoscience behavior. The cryosphere category is similarly amenable, where GPT-Image-2 achieves 54.3% strict accuracy since glacier retreat and snowline shifts exhibit clear directional signals.

Atmosphere and ocean dynamics form the hardest slice. Even GPT-Image-2 reaches only 35.3% here, and most systems fall below 10%. These tasks, including cyclone generation, transposition, and monsoon wet to dry transitions, require reasoning about rotational sense, the Coriolis effect, and circulation structure, none of which can be spoofed by surface visual priors. Geomorphology is also surprisingly demanding at the top, where the best performance reaches only 25.0%, because delta progradation, dune migration, and oxbow cutoff are process based. The correct answer is determined by sediment transport and channel geometry rather than scene appearance. GIS and spatial geometry together with crustal science form a middle regime. GPT-Image-2 performs strongly by scoring 52.4 in spatial tasks and 41.9 in crustal science, and Nano Banana 2 follows closely with scores of 45.2 and 48.4 in these respective domains. However, weaker systems fall below 10% in these categories due to the combined demands of viewpoint, scale, and symbol convention.

4.5 Open-source Landscape

Within open source systems we observe three additional patterns. First, Qwen-Edit-2511 is a clear leader with a reasoning score of 48.9 and a quality score of 78.1, making it the only open source system to break 10% strict accuracy. Second, adding a chain of thought mechanism does not uniformly

Table 3: Alignment between automated evaluation and expert ratings. Lower is better.

	Reasoning		Consistency		Image Quality	
Judge model	MAE↓	STD↓	MAE↓	STD↓	MAE↓	STD↓
Qwen3-VL-235B Team (2025)	0.216	0.341	0.178	0.361	0.188	0.207
GPT-5 Singh et al. (2025)	0.168	0.298	0.194	0.384	0.172	0.208
Gemini-3-Flash Google (2025a)	**0.113**	**0.244**	**0.131**	**0.300**	0.228	0.272
Q-Align Wu et al. (2023)	–	–	–	–	**0.097**	**0.121**

help. Step-1x with a think and reflect configuration reaches a reasoning score of 21.3 and an accuracy of 2.67%, which falls marginally below the base Step-1x reasoning score of 24.7 and accuracy of 3.45%. In contrast, Bagel benefits from explicit reasoning steps by improving its reasoning score from 12.7 to 18.7. The value of explicit reasoning steps is therefore dependent on the architecture rather than the specific task. Third, the Mixture of Experts variant of ICEdit is an outlier at the bottom of the leaderboard with a reasoning score of 7.1, consistency of 18.2, and quality of 29.8. This performs substantially worse than the base ICEdit scores of 13.2, 31.1, and 59.8 respectively. The quality collapse is especially informative because it shows the Mixture of Experts variant is not just reasoning weak but also generation unstable on geoscience inputs. Together these observations suggest that current open source editing systems remain architecturally tuned for natural image edits and require dedicated alignment for scientific content.

4.6 Ablation Study

Instruction explicitness. Many GeoR-Bench prompts intentionally require implicit geoscience reasoning by asking the model to evolve a delta rather than explicitly instructing it to add distributary channels seaward. To test how much models benefit from explicit guidance, we rewrite a matched subset of prompts into procedural instructions while keeping the input and target output fixed. Figure 4 shows a clear asymmetry. Open-source systems gain the most from explicit instructions. Qwen-Edit-2511 improves from 10.3% to 24.7% strict accuracy, which is a 14.4-point absolute gain. FLUX.2-dev increases from 5.1% to 14.6% for a 9.5-point gain, and Step1X-Edit sees a modest 1.9-point improvement. This diagnostic compares task interpretation and visual execution under fixed inputs; it does not provide a causal decomposition. Closed-source systems gain unevenly. GPT-Image-2 improves from 42.7% to 45.3%, while Nano Banana Pro improves from 35.0% to 46.3%.

Judge alignment. We assess automated judging against human ratings on 48 benchmark instances spanning the six categories. Each instance contributes all 21 model outputs, yielding $48 \times 21 = 1{,}008$ output-level evaluations. Eight trained geography undergraduates each rated 126 distinct, non-overlapping outputs using the same public point-by-point rubrics. Because no output received ratings from multiple people, this audit does not support conventional inter-rater agreement or adjudication; it measures judge-versus-human alignment across 1,008 outputs. Table 3 reports judge-level mean absolute error and standard deviation on each dimension. Gemini-3-Flash attains the lowest MAE on both reasoning (0.113) and consistency (0.131), and Q-Align the lowest image-quality MAE at 0.097. The single-rater allocation is a limitation of this audit.

5 Conclusion

We present GeoR-Bench, a benchmark that evaluates geoscience intelligence through structured image reasoning and generation tasks. This dataset stems from the observation that geoscience images carry strict scientific meaning, so evaluation must extend beyond standard visual realism. By grounding the evaluation in curated geoscience imagery across 440 samples spanning six categories and 24 task types, GeoR-Bench measures multimodal models' geoscience visual reasoning rather than aesthetics alone. Our framework jointly assesses reasoning, consistency, and quality, revealing significant limitations in current multimodal systems: no evaluated model achieves 50% strict accuracy. Ultimately, we hope GeoR-Bench guides future research toward physically grounded and scientifically faithful image generation for critical Earth-system applications.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic benchmark images and for generating the one-sentence scene descriptions used in the benchmark prompts as well as LLM as judge. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in writing proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, translation, cleaning or reformatting datasets, qualitative or thematic data analysis, or interpreting results; proof-related tasks are not applicable to this work. Additionally, we used generative AI tools for editing and polishing the paper to improve readability. We have reviewed all AI-assisted work. We checked the generated scene descriptions and benchmark images for consistency with the specified maps, prompts, and experiment protocol, and checked the AI-assisted edits for accuracy and consistency with the intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Yifan Chang, Yukang Feng, Jianwen Sun, Jiaxin Ai, Chuanhao Li, S. Kevin Zhou, and Kaipeng Zhang. Sridbench: Benchmark of scientific research illustration drawing of image generation model. *ArXiv*, abs/2505.22126, 2025. URL https://api.semanticscholar.org/CorpusID:278960126.

Yun Chen, Yu Gao, Lixue Gong, Meng-Hao Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, Xiaowen Jian, Huafeng Kuang, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yanzuo Lu, Zhengxiong Luo, Tongtong Ou, Guangchao Shi, Yichun Shi, Shiqi Sun, Yu-Chen Tian, Zhi Tian, Peng Wang, Rui Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Wenxu Wu, Yonghui Wu, Xin Xia, Xuefeng Xiao, Shuang Xu, Xin Yan, Ceyuan Yang, Jianchao Yang, Zhonghua Zhai, Chenlin Zhang, Heng Zhang, Qi Zhang, Xinyu Zhang, Yuwei Zhang, Shijia Zhao, Wenliang Zhao, and W. Zhu. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiao-Ping Nie, Ziang Song, Shi Guang, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Google. Gemini 3 flash, 2025a.

Google. Introducing gemini 2.5 flash image (nano banana). https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, 2025b.

Google. Nano banana 2: Combining pro capabilities with lightning-fast speed. https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/, 2025c.

Google. Introducing nano banana pro. https://blog.google/innovation-and-ai/products/nano-banana-pro/, 2025d.

Jingliang Hu, Rong Liu, Danfeng Hong, Andrés Camero, Jing Yao, Matthias Schneider, Franz Kurz, Karl Segl, and Xiaoxiang Zhu. Mdas: a new multimodal benchmark dataset for remote sensing. *Earth System Science Data*, 15(1):113–133, 2023.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding.

Mingxin Liu, Ziqian Fan, Zhaokai Wang, Leyao Gu, Zirun Zhu, Yiguo He, Yuchen Yang, Changyao Tian, Xiangyu Zhao, Ning Liao, Shaofeng Zhang, Qibing Ren, Zhihang Zhong, Xuanhe Zhou, Junchi Yan, and Xue Yang. Grade: Benchmarking discipline-informed reasoning in image editing, 2026. URL https://arxiv.org/abs/2603.12264.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Jialin Luo, Yuanzhi Wang, Ziqi Gu, Yide Qiu, Shu-qing Yao, Fuyun Wang, Chunyan Xu, Wenhua Zhang, Dan Wang, and Zhen Cui. Mmm-rs: A multi-modal, multi-gsd, multi-scene remote sensing dataset and benchmark for text-to-image generation. *arXiv preprint arXiv:2410.22362*, 2024.

OpenAI. GPT-Image-1.5. https://openai.com/index/new-chatgpt-images-is-here/, 2025a.

OpenAI. Gpt-image-2.0. https://developers.openai.com/api/docs/models/gpt-image-2, 2025b.

Seedream. Seedream 4.5. https://seed.bytedance.com/en/seedream4_5/, 2025.

Seedream. Seedream 5.0. https://seed.bytedance.com/en/seedream5_0_lite/, 2026.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi.

Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. Kris-bench: Benchmarking next-level intelligent image editing models. 2025b. URL https://arxiv.org/abs/2505.16707.

Bowen Xia, Yuhang Zhang, Jiacheng Li, Cheng Wang, Yang Wang, et al. Dreamomni: Unified image generation and editing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025.

Zitong Xu, Huiyu Duan, Zhongpeng Ji, Xinyun Zhang, Yutao Liu, Xiongkuo Min, Ke Gu, Jian Zhang, Shusong Xu, Jinwei Chen, et al. Edithf-1m: A million-scale rich human preference feedback for image editing. *arXiv preprint arXiv:2603.14916*, 2026.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. 2024.

Daoan Zhang, Che Jiang, Ruoshi Xu, Biaoxiang Chen, Zijian Jin, Yutian Lu, Jianguo Zhang, Liang Yong, Jiebo Luo, and Shengda Luo. Worldgenbench: A world-knowledge-integrated benchmark for reasoning-driven text-to-image generation. 2025. URL https://arxiv.org/abs/2505.01490.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Y Tang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. arxiv 2025. *arXiv preprint arXiv:2504.20690*.

Xiangyu Zhao, Wanghan Xu, Bo Liu, Yuhao Zhou, Fenghua Ling, Ben Fei, Xiaoyu Yue, Lei Bai, Wenlong Zhang, and Xiao-Ming Wu. Msearth: A multimodal scientific dataset and benchmark for phenomena uncovering in earth science. *arXiv preprint arXiv:2505.20740*, 2025.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *Advances in Neural Information Processing Systems*, 38, 2026.

A Additional Details

A.1 Release statistics

The GeoR-Bench release contains 440 samples spanning 6 categories and 24 task types. Each category contains 4 task types, and each task type includes at least 8 samples. The 6 categories are geomorphology, hydrology, atmosphere and ocean dynamics, cryosphere science, GIS and spatial geometry, and crustal science.

The benchmark includes style-consistent, independence-consistent, and localized-consistent tasks. This distribution reflects the fact that many geoscience tasks involve either temporal forecasting with preserved sensor style or cross-view reconstruction where output fidelity should follow geoscience structure rather than raw appearance.

A.2 Task taxonomy and coverage

The taxonomy allocates four task families to each reporting category. It is a balanced diagnostic scope over process reasoning, geometry, topology, cross-view transformation, and scientific conventions; it is not intended as a complete ontology of geoscience.

Table 4: Task-family coverage in GeoR-Bench. The capability column states the primary scientific constraint probed by each task family.

Category	Task family	Primary scientific constraint
GEOM	Oxbow-lake formation	Channel topology and temporal evolution
	River-delta evolution	Sediment transport and spatial growth
	Dune migration	Wind-driven transport and direction
	Coriolis bank asymmetry	Physical direction and cross-section geometry
HYDR	Flow-direction inference	Drainage direction from spatial evidence
	Water-cycle completion	Directed fluxes and process relations
	Monsoon lake fluctuation	Seasonal extent and temporal change
	Ocean-current mapping	Circulation paths and map conventions
ATMO	Cyclone generation/transposition	Circulation orientation and hemisphere
	Meteorite burn-up control	Physical conditions for an impact event
	Sea–land breeze patterns	Thermal contrast and circulation direction
	Monsoon map drawing	Seasonal circulation and spatial relations
CRYO	Glacier/proglacial-lake evolution	Ice retreat and associated lake extent
	Snowline prediction	Elevation-dependent snow/ice boundary
	Snow/green-up phenology	Seasonal appearance under environmental change
	Periglacial diagrams	Process constraints in scientific representations
SPATIAL	Earth rotation/revolution	Viewpoint, motion, and illumination geometry
	Canyon shadows	Terrain and illumination geometry
	Street-to-satellite reconstruction	Cross-view location and spatial layout
	Satellite-to-street reconstruction	Cross-view geometry and scene structure
CRUSTAL	Seismology	Wave interpretation and source geometry
	Volcanology	Volcanic structure and process conventions
	Structural geology	Geological geometry and labeling
	Marine geology	Seafloor structure and scientific labeling

A.3 Geographic coverage

The geolocated subset contains 129 terrestrial sites in 48 countries on six continents, together with seven open-ocean scenes. These counts describe coverage rather than uniform site density, and they do not characterize non-geolocated diagrams. Russia is counted under Europe.

Table 5: Geographic coverage of the geolocated subset.

Region	Number of countries or scenes
Asia	12 countries
Africa	14 countries
South America	6 countries
North America	5 countries
Europe	9 countries
Oceania	2 countries
Open ocean	7 scenes
Total	48 countries + 7 ocean scenes

A.4 Source provenance and release treatment

The release distinguishes assets whose pixels can be redistributed from restricted source material. Public-domain or appropriately licensed observational assets and author-created material are distributed with source-specific attribution. Google Maps and Street View inputs are represented by source ID and reconstruction parameters rather than image pixels when required by provider terms. Each sample's metadata records the source platform, asset identifier, license or terms of use, attribution, and redistribution status.

Table 6: Source and license information documented for the GeoR-Bench release. T1 denotes pixels or derived products that can be distributed under the listed terms; T2 denotes ID-only release with reconstruction metadata.

Category	Task(s)	Source and asset identifiers	License / release tier
GEOM	Oxbow lakes; delta evolution; dune migration	Landsat 5/7/8–9 C02 and Sentinel-2 (`COPERNICUS/S2_SR_HARMONIZED`) via Google Earth Engine	USGS public domain; Copernicus free/open; T1
GEOM	Coriolis bank asymmetry	Google Static Maps views; SRTM DEM (`USGS/SRTMGL1_003`) renders	Google terms, T2; SRTM public domain, T1
HYDR	Flow direction	Landsat Collection 2 via Google Earth Engine	USGS public domain; T1
HYDR	Water cycle; ocean currents	Author-drawn schematics; current overlays on a NOAA public-domain base map	NOAA public domain; author-created CC BY 4.0; T1
HYDR	Monsoon lake fluctuation	Sentinel-2 and MODIS via Google Earth Engine	Copernicus and NASA terms; T1
ATMO	Cyclone generation/transposition	NASA Worldview (MODIS/VIIRS) and NOAA GOES	NASA/NOAA public domain; source URL per image; T1
ATMO	Meteorite burn-up control	Sentinel-2 crater scenes; Copernicus GLO-30 depth renders; author-built counterfactuals	Copernicus free/open; author-created CC BY 4.0; T1
ATMO	Sea–land breeze	MODIS, NOAA OISST, Sentinel-2, and Landsat-8 via Google Earth Engine	NASA/NOAA/Copernicus terms; T1
ATMO	Monsoon map drawing	Author-drawn base and answer maps; question ideas adapted from cited public exams	Author-created CC BY 4.0; cited exam sources; T1
CRYO	Proglacial lakes; glacier retreat	Sentinel-2 SR Harmonized and Landsat 5 via Google Earth Engine	Copernicus; USGS public domain; T1
CRYO	Snowline prediction	Copernicus GLO-30 hillshade input; ELA output computed by the authors	Derived-product redistribution with credit; T1
CRYO	Snow/green-up phenology	Sentinel-2 via Google Earth Engine	Copernicus free/open; T1
CRYO	Periglacial diagrams	Author-drawn diagrams	Author-created CC BY 4.0; T1
SPATIAL	Earth rotation/revolution	Renders from NASA Blue Marble texture	NASA free with credit; T1
SPATIAL	Canyon shadow simulation	SRTM DEM via Google Earth Engine; shadows simulated by the authors' code	Public-domain input; author-created derived product; T1
SPATIAL	Street-satellite cross-view	Google Street View panoramas and satellite orthophotos	Google Maps terms; IDs, coordinates, headings, and script only; T2
CRUSTAL	Seismology; volcanology; structural geology; marine geology	Original diagrams and questions created by the author team	Author-created CC BY 4.0; T1

A.5 Human audit accounting

The human audit covers 48 benchmark instances and all 21 model outputs for each instance, giving $48 \times 21 = 1{,}008$ output-level evaluations. Eight trained geography undergraduates each evaluated

126 distinct outputs with the same public point-by-point rubrics. The allocation was non-overlapping: each output was rated once, so conventional inter-rater agreement and adjudication are unavailable. The audit therefore measures judge-versus-human alignment across 1,008 outputs rather than inter-rater reliability across repeated ratings.

A.6 STANDARDIZED SAMPLE FORMAT

Each task is standardized into a directory of the following form:

```
{task_id}/
|-- input.png
|-- output.png
|-- prompt.txt
`-- judge.md
```

A top-level `manifest.jsonl` indexes all tasks. The dataloader returns a dictionary with `task_id`, `category`, `task_type`, `input_image`, `prompt`, `output_image`, `judge_prompt`, `context_images`, and `metadata`.

A.7 REPRESENTATIVE TASK FAMILIES

Representative task families in the release include river delta evolution, oxbow lake formation, monsoon-driven wetland fluctuation, cyclone generation and transposition, glacier retreat, snowline altitude prediction, crater generation, marine geology labeling, seismology diagram completion, volcanic and igneous petrology cross-sections, and cross-view satellite reconstruction. This mixture is important: geoscience competence includes both image understanding in the wild and correct editing of formal scientific representations.

B HUMAN EXPERT SCORING INTERFACE

This is the human expert scoring interface for GeoR-Bench. Each sample is scored along three dimensions: reasoning, consistency, and quality. The interface displays the input image, model prediction, ground-truth output, and task-specific judging rubric to guide the expert's evaluation.

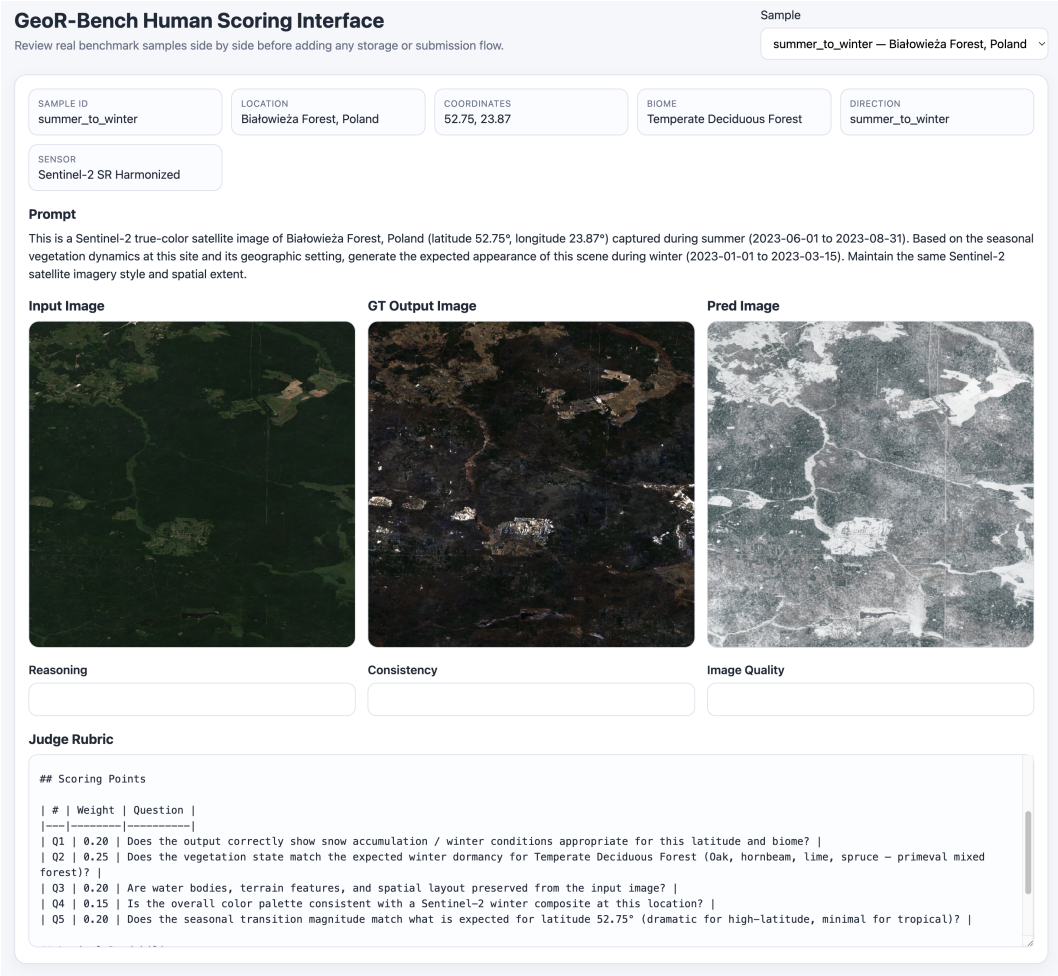
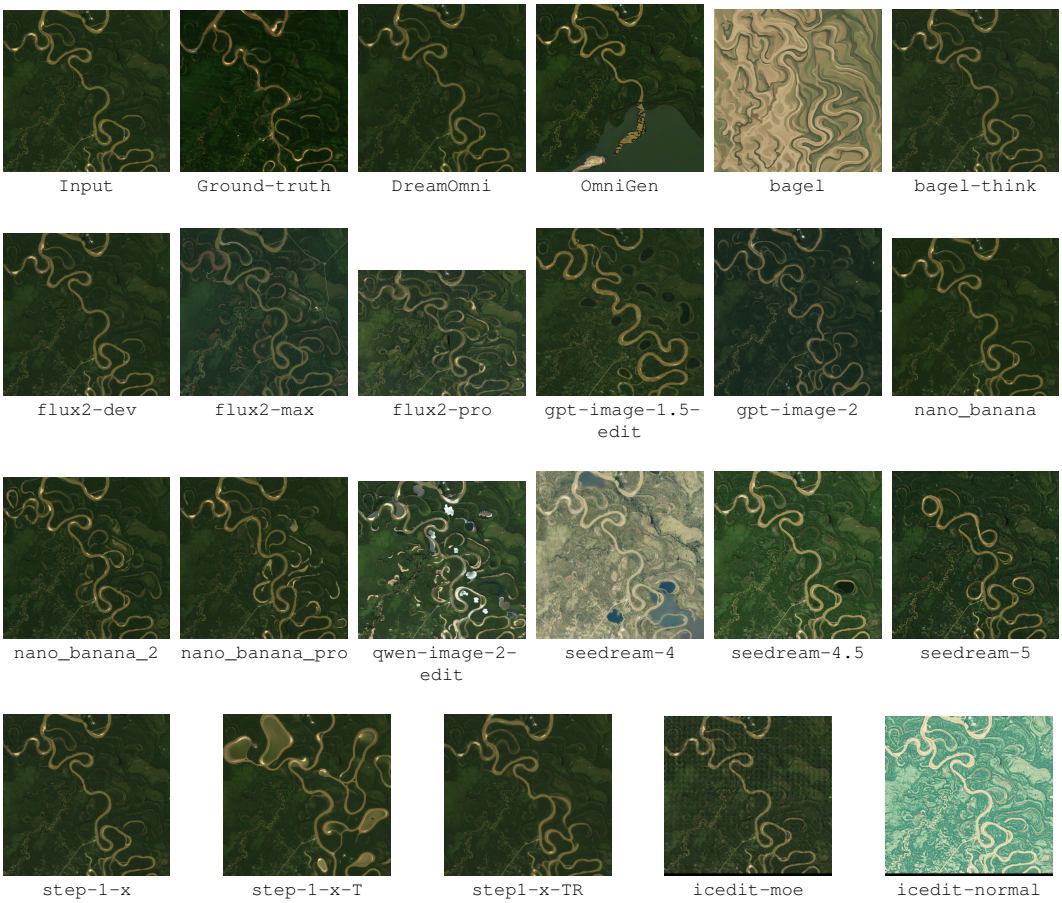


Figure 5: The human expert scoring interface for GeoR-Bench. Each sample is scored along three dimensions: reasoning, consistency, and quality. The interface displays the input image, model prediction, ground-truth output, and task-specific judging rubric to guide the expert's evaluation.

C SELECTED CASE STUDIES

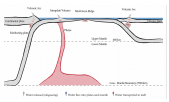
CASE STUDY 1

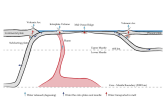
Prompt. This is a Landsat 5 satellite image of the Ichilo River in Bolivia (16.80 degS, 64.80 degW) captured in 1992. The image shows a section of a meandering tropical river flowing through the Bolivian lowlands of the Amazon basin, characterized by active lateral migration and frequent channel avulsion in its alluvial floodplain. Based on the principles of fluvial geomorphology, generate the expected appearance of this river section in 2009, after 17 years of continued meander migration, bend tightening, cutoff development, and oxbow lake formation. Maintain the same satellite imagery style and spatial extent.

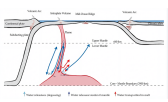


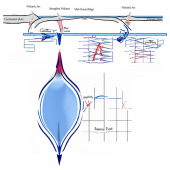
CASE STUDY 2

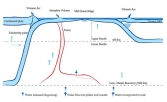
Prompt. This diagram shows a tectonic cross-section, but the arrows for the deep water cycle are missing. Draw the deep water cycle on this figure. Use dark blue arrows for water flux into plates and mantle, light blue arrows for water released by degassing, and red arrows for water transported in melt. Keep the existing cross-section geometry, labels, and all other elements unchanged.

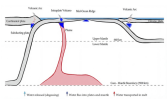

Input

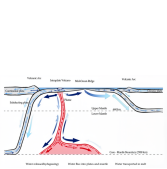

Ground-truth

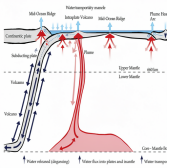

DreamOmni

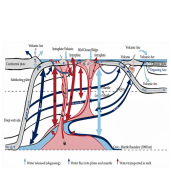

OmniGen

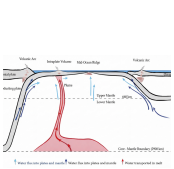

bagel

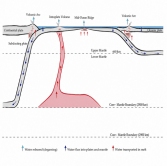

bagel-think

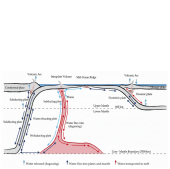

flux2-dev

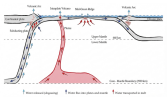

flux2-max

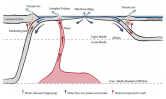

flux2-pro

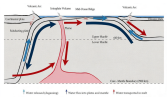

gpt-image-1.5-edit

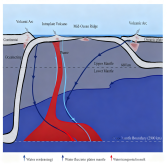

gpt-image-2

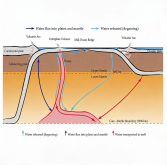

nano_banana

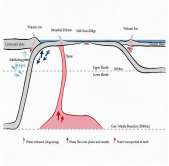

nano_banana_2

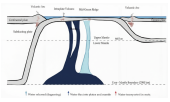

nano_banana_pro

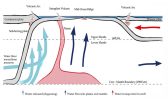

qwen-image-2-edit

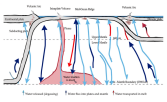

seedream-4


seedream-4.5

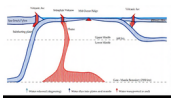

seedream-5


step-1-x


step-1-x-T

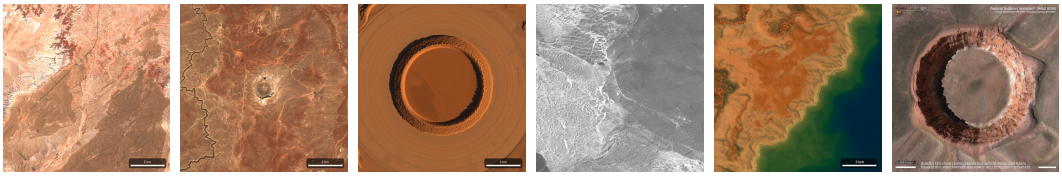

step1-x-TR


icedit-moe

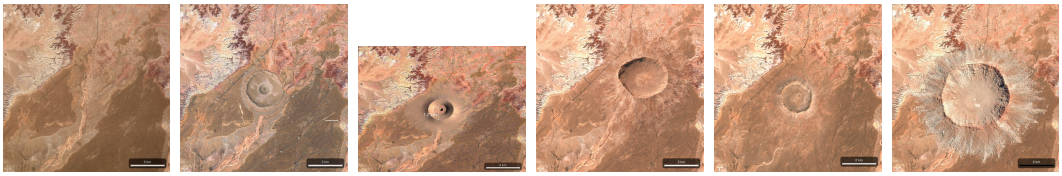

icedit-normal

CASE STUDY 3

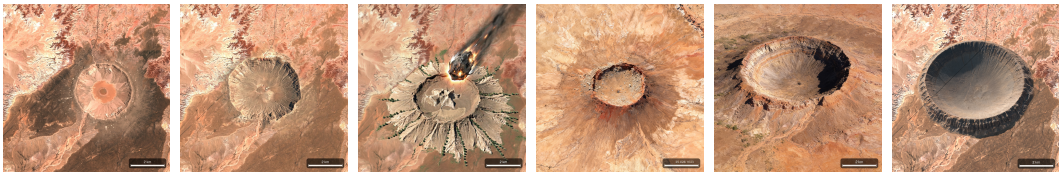
Prompt. This is a true-color Sentinel-2 satellite image of undisturbed terrain near Barringer (Meteor Crater), United States (35.028, -111.023), selected to match the pre-impact geology of the Colorado Plateau. Assume the scene has the same Permian-Triassic target sequence as the impact site, including quartz sandstone, sandy dolomite, and siltstone. Using the meteorite entry conditions - a 50 m iron body traveling at 12.8 km/s and entering at 45 deg - infer the landform and surface expression that would develop in this setting and render the resulting landscape as it would appear after about 50,000 years of preservation. Maintain the same true-color satellite style and spatial extent.



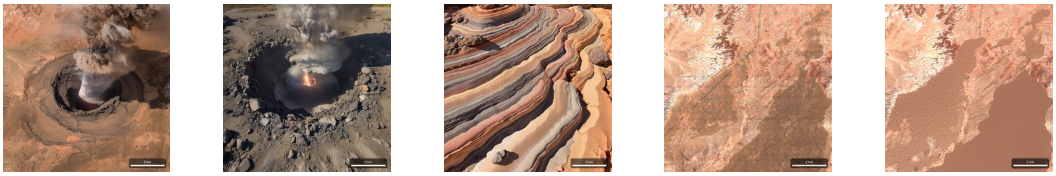
Input | Ground-truth | DreamOmni | OmniGen | bagel | bagel-think



flux2-dev | flux2-max | flux2-pro | gpt-image-1.5-edit | gpt-image-2 | nano_banana



nano_banana_2 | nano_banana_pro | qwen-image-2-edit | seedream-4 | seedream-4.5 | seedream-5



step-1-x | step-1-x-T | step1-x-TR | icedit-moe | icedit-normal