SHOW, DON'T TELL: EVALUATING SPATIAL COGNITION IN GENERATIVE PIXELS RATHER THAN LLM TEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial intelligence enables agents to move beyond static semantic understanding and interact with the physical world. Many spatial tasks are grounded in continuous visual scenes, where locations, regions, and paths are more naturally expressed by pointing, marking, or drawing than by precise coordinates or discrete textual symbols. Yet existing spatial benchmarks usually require coordinates, options, or text, creating an answer-interface mismatch that makes it difficult to evaluate image-generation models under the same task semantics as text-output VLMs, despite their ability to externalize spatial judgments in pixel space. To address this problem, we propose **ProVisE** (***Pro****tocolized* ***Vis****ual* ***E****valuation*), a benchmark-agnostic framework that elicits protocol-constrained visual answers from image-generation models, parses them into structured predictions for task-specific scoring, and extends to previously unconfigured benchmarks through Agentic construction and validation of executable generation–parser protocols. We further introduce **SpatialGen-Bench**, a curated diagnostic benchmark comprising 470 samples across 14 spatial subtasks, four capability levels, and diverse answer forms. We evaluate representative text-output VLMs and image-generation models in a unified setting and validate Agentic protocol construction on six external spatial benchmarks. Experimental results show that image-generation models are competitive when spatial answers can be externalized in pixel space, while text-output VLMs retain a clear advantage in compositional spatial reasoning. These findings reveal complementary strengths of pixel-space expression and text-based reasoning, while ProVisE and SpatialGen-Bench establish a unified testbed for generative spatial evaluation and future studies of spatial cognition in image-generation models.

1 INTRODUCTION

Spatial intelligence is a foundational prerequisite for artificial agents to move beyond static semantic understanding and interact safely with the physical world (Yu et al., 2025; Liu et al., 2026; Zheng et al., 2025). It spans a broad range of perceptual and reasoning abilities, including visual grounding, affordance prediction, target-region localization, and relational spatial reasoning over positions, distances, and orientations. As vision-language models (VLMs) (Achiam et al., 2023; Comanici et al., 2025; Bai et al., 2025a) have become increasingly capable, substantial effort has been devoted to improving their spatial perception and reasoning abilities (Chen et al., 2024; Wu et al., 2025b; Cheng et al., 2024; Cai et al., 2025; Liu et al., 2025; Zhou et al., 2025; Yang et al., 2025b; Li et al., 2025b). Following progress in visual question answering (VQA) and spatial VLM evaluation (Yang et al., 2025a; Fu et al., 2024; Wang et al., 2026a; Yang et al., 2025c; Ray et al., 2024; Jia et al., 2025; Li et al., 2025a; Tong et al., 2024; Du et al., 2024; Zhang et al., 2025b), existing spatial intelligence research has largely been framed around text-output VLMs. Models are typically prompted and evaluated through symbolic or structured answer formats, such as coordinates, option labels, or textual descriptions. This paradigm makes spatial reasoning measurable for VLMs, but it also forces continuous visual judgments into textual or coordinate-based outputs, which are often unnatural for spatial tasks.

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.

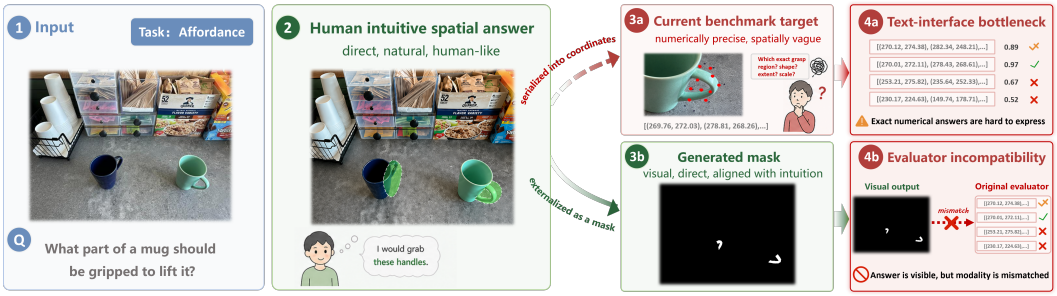


Figure 1: **Text-interface bottleneck in spatial evaluation.** Spatial judgments are naturally expressed through intuitive visual outputs, yet benchmarks often require precise coordinates or symbolic targets. As a result, text-output VLMs may be penalized for failing to report exact coordinates, while visual answers from image-generation models fall outside the original evaluator.

Recent progress in generative vision models motivates a closer evaluation of their spatial cognition. Unlike text-output VLMs, image-generation models can express spatial judgments directly in pixel space by modifying, marking, or synthesizing visual content. This makes them natural candidates for spatial tasks whose answers are continuous or visually grounded, rather than purely symbolic. Recent work casts segmentation, depth estimation, and counting as image-generation tasks (Gabeur et al., 2026; Di et al., 2026), while video-generation models exhibit zero-shot behaviors across perception, manipulation, and early visual reasoning (Wiedemer et al., 2025). Related progress in native and unified multimodal generation further suggests that visual generation may encode representations useful for understanding and world modeling (Cui et al., 2025; Xie et al., 2025; Deng et al., 2025). Controlled visual–verbal comparisons further show task-dependent reasoning benefits from visual world modeling (Wu et al., 2026).

Existing evaluation lines address complementary objectives. Text-to-image benchmarks assess prompt fidelity, compositional alignment, or spatial relations in synthesized images (Hu et al., 2023; Ghosh et al., 2023; Huang et al., 2025; Li et al., 2024; Bakr et al., 2023; Gokhale et al., 2022; Wang et al., 2026b); they evaluate whether a generated image satisfies a text prompt, not whether a model can answer a spatial question grounded in an input scene. Conversely, spatial reasoning benchmarks for VLMs focus on text-output models and score symbolic or structured predictions (Liu et al., 2023; Wu et al., 2025c; Xu et al., 2026; Stogiannidis et al., 2025; Wu et al., 2025d). These lines motivate a reusable protocol layer that aligns visual answers with heterogeneous spatial task semantics and scoring rules.

The main obstacle is the answer interface, as illustrated in Figure 1. Existing spatial benchmarks are primarily designed for text-output VLMs and typically require coordinates, option labels, or textual descriptions as final answers (Wu et al., 2025c). For tasks involving regions, paths, depths, or orientations, this requires models to express spatial judgments as numerical or symbolic outputs, such as bounding-box coordinates $[x_1, y_1, x_2, y_2]$, per-pixel depth values, or direction labels. This is analogous to asking a person to report exact pixel coordinates instead of pointing to or sketching a region. Such an interface is unnatural for text-output VLMs, which must verbalize spatial structure, and even more restrictive for image-generation models, whose native responses are visual. Although an image-generation model may mark an object, highlight a region, or draw a trajectory, these responses cannot be directly scored by benchmarks expecting textual or coordinate-based predictions. ***How can we systematically evaluate spatial cognition in image-generation models?***

A natural workaround is to use an auxiliary VLM judge to read the generated images and infer the final answer (Wu et al., 2025c; Wang et al., 2026b). However, judge-based scores depend on the judge model's own reliability and may reflect judge uncertainty rather than the generated spatial answer itself (Cho et al., 2024). Even under shared scoring rules, variation across judges and repeated evaluations can change scores and rankings, introducing evaluator-dependent uncertainty into comparisons between text-output VLMs and image-generation models.

To make generative spatial cognition measurable, we introduce an evaluation framework together with a diagnostic benchmark. **ProVisE** (***Pro**tocolized **Vis**ual **E**valuation*) is a benchmark-agnostic framework that elicits protocol-constrained visual answers from image-generation models. Instead of requiring textual or coordinate-based responses, ProVisE constrains generated outputs into pre-

defined visual formats and parses them into structured predictions for task-specific scoring. Because manually engineering a protocol for every new task does not scale, ProVisE also provides an Agentic construction path. It normalizes a new benchmark into task-level contracts, constructs an appropriate generation–parser protocol for each task, and validates the resulting artifact before model evaluation. We further construct **SpatialGen-Bench**, a curated diagnostic benchmark for spatial cognition in image-generation models. It covers 14 subtasks across perception, understanding, reasoning, and interaction, providing a capability-level view of model strengths and weaknesses. Its diverse answer forms also allow us to examine whether ProVisE can support metric-compatible visual evaluation across different answer interfaces. Together, ProVisE and SpatialGen-Bench enable systematic comparison between text-output VLMs and image-generation models under shared task semantics.

By deploying this evaluation suite, we provide a systematic comparison of spatial cognition in image-generation models and text-output VLMs. Image-generation models are most competitive when answers can be externalized as inspectable pixel-space evidence, while text-output VLMs remain stronger when tasks require compositional reasoning and precise spatial transformations. Additional experiments show that universal VLM parsing changes both scores and rankings, while most unsuccessful visual responses are classified as valid-but-incorrect by the evaluation pipeline. Experiments on six external benchmarks further show that the Agentic builder can adapt ProVisE to heterogeneous task and metric contracts. Together, these results reveal complementary strengths in visual externalization and text-based reasoning and motivate their coordinated use.

In summary, we make three contributions to generative spatial evaluation:

- **Modality Mismatch in Generative Spatial Evaluation**: We identify an answer-interface mismatch in spatial benchmarks: text- or coordinate-only responses can obscure VLM assessment and exclude image-generation models from shared task semantics and metrics.
- **A Unified Suite for Image-Generation Spatial Evaluation**: We introduce ProVisE, a benchmark-agnostic framework with an Agentic builder, and SpatialGen-Bench, a spatial-cognition benchmark supporting diverse answer forms from VLMs and image generators.
- **Systematic Comparison of Generative Models and VLMs**: We systematically compare image-generation models and text-output VLMs across spatial capabilities. Visual answering is competitive when conclusions can be externalized directly in pixels, whereas text-output VLMs remain stronger in compositional spatial reasoning.

2 Protocolized Visual Evaluation

ProVisE evaluates spatial cognition in image-generation models by converting visual answers into structured predictions for task-specific scoring. Its workflow in Figure 2 couples generation and parsing through protocols shared across target models. An Agentic builder constructs and validates protocols for new benchmarks, extending evaluation beyond manually configured tasks.

2.1 Visual Protocol Definition

A visual protocol is the basic evaluation unit of ProVisE. It couples two components: a **Generation Prompt**, which constrains how a model expresses its answer in pixel space, and a **Parser**, which recovers a structured prediction from the generated response. The protocol further declares the expected answer schema, invalid-output conditions, and task-specific scoring rule. This contract is fixed before evaluation and applied unchanged across all target models.

To cover heterogeneous spatial tasks, ProVisE maintains a task-aware protocol pool in which every visual answer format is paired with a corresponding parser. The pool spans instance and point marking, direction grids, relative-depth maps, region masks, state matching, trajectory drawing, and task-grounded evidence rendering; Appendix C.2 gives the operational parser rules, prediction formats, and invalid-output boundaries for each visual protocol.

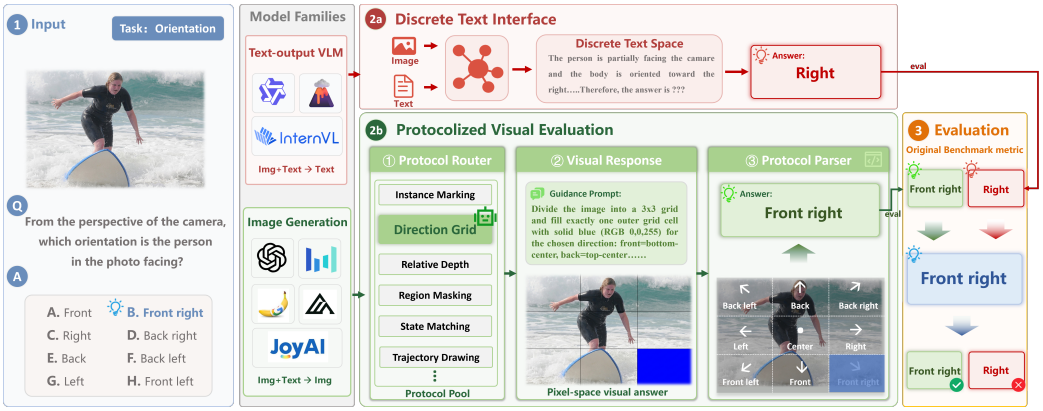


Figure 2: **Overview of ProVisE.** Text-output VLMs follow **2a**; image-generation models follow **2b**, where a routed protocol guides generation and parsing. Both use shared task-specific scoring rules.

2.2 Protocolized Evaluation Workflow

Each task is first routed to one protocol. ProVisE reuses a validated configuration when available or invokes the Agentic process in Sec. 2.3 to construct one from the benchmark contracts. The accepted configuration is fixed across target models, separating protocol design from model comparison.

The selected protocol guides branch 2b in Figure 2 to produce a constrained visual response rather than a free-form image. Depending on the task, the response may contain marked instances, a direction grid, a depth map, a region mask, a grounded point, a rendered state, or a trajectory. Its parser then recovers the structured answer required for scoring. Deterministic parsing is preferred; the fallback VLM receives only the generated response and decoding specification, including the answer schema and candidate choices where applicable. It does not receive the source image or original question, so its role is to recover the expressed answer. Figure 3 shows representative protocol-constrained outputs and recovered predictions.

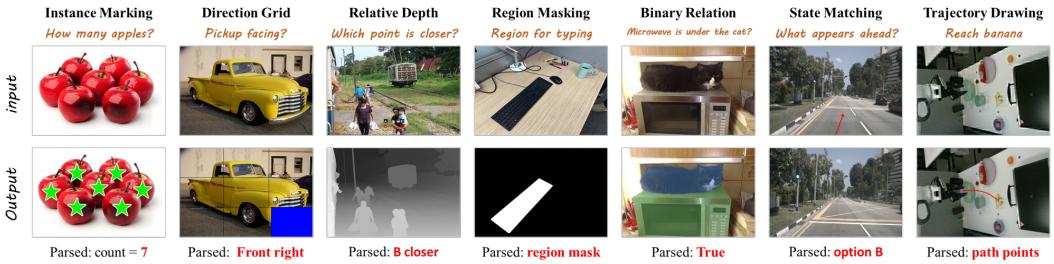


Figure 3: Representative protocol-execution cases for ProVisE. Each column shows a benchmark input (top), the protocolized output (bottom), and the recovered structured prediction.

Text and visual responses are evaluated against the same reference annotations using the task-specific scoring rules defined in Appendix D.4. Text responses are normalized directly; visual responses are decoded through their assigned protocols. Most protocols use deterministic image processing, geometry, or fixed similarity models, while Fallback uses a protocol-constrained, generated-response-only parser distinct from the universal VLM judging examined in Sec. 4.5.

2.3 Agentic Protocol Construction

For a previously unconfigured benchmark, Agentic construction produces a validated, executable protocol in the three stages shown in Figure 4. First, ProVisE identifies benchmark records, input media, task labels, answer formats, and metrics, then normalizes them into task contracts with explicit answer targets and scoring rules.

Second, the Agent examines representative task examples and selects one of three construction routes. **Reuse** selects a compatible protocol, **Build** assembles one from registered parsing compo-

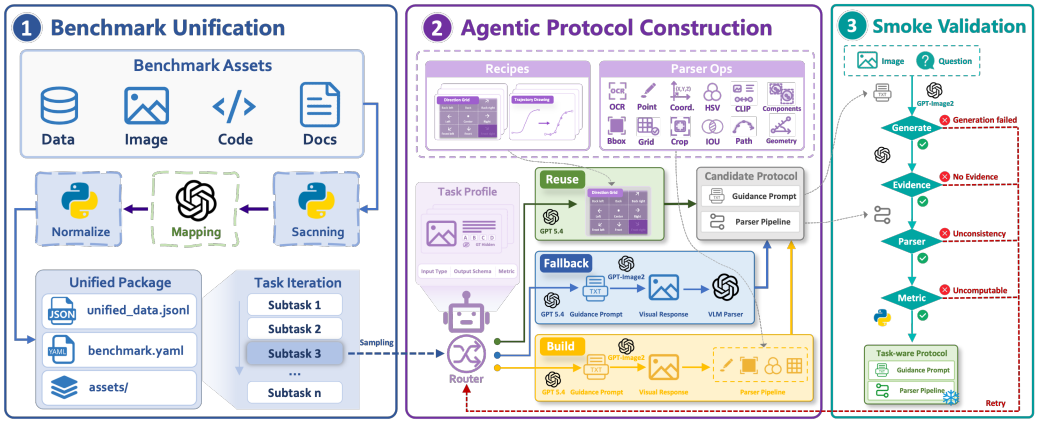


Figure 4: **Agentic protocol construction in ProVisE.** A benchmark is normalized into task contracts; the Agent constructs a protocol for each task; and the resulting generation–parser pair is automatically validated before being applied unchanged to every target model.

nents, and **Fallback** selects a constrained generated-response-only parser when deterministic components cannot recover the required answer form. Every route produces the same declarative artifact: a generation prompt, parser configuration, answer contract, invalid-output rules, and metric mapping. The Agent composes registered components rather than generating arbitrary evaluation code.

Smoke validation checks visual generation, repeated-parse agreement, recoverable evidence, and compatibility with the declared scoring rule. Failed candidates may be revised or rerouted; accepted protocols are fixed across target models to prevent model-specific optimization. Appendix C.3 details normalization, construction routes, validation criteria, and outcomes.

3 SpatialGen-Bench

3.1 Capability Hierarchy

SpatialGen-Bench is a curated diagnostic benchmark that prioritizes sample quality, capability coverage, and diverse answer forms. Its four complementary levels in Figure 5 span observable attributes, scene understanding, spatial inference, and action-oriented planning.

- **Spatial perception** reads observable spatial attributes from visual evidence through counting, relative depth estimation, orientation judgment, and object size comparison.
- **Spatial understanding** builds relational and scene-level representations through relationship verification, perspective judgment, spatial mental modeling, and spatial grounding.
- **Spatial reasoning** derives conclusions and state changes beyond directly observed evidence through multi-hop spatial reasoning, spatial state prediction, and geometric feasibility.
- **Spatial interaction** translates spatial understanding into action targets and paths through affordance grounding, navigation, and trajectory planning in the observed scene.

3.2 Benchmark Construction and Quality Control

Data sources. We curate SpatialGen-Bench from public benchmarks (Paiss et al., 2023; Fu et al., 2024; Liu et al., 2023; Jung et al., 2025; Li et al., 2025a; Wu et al., 2026; Wang et al., 2026a; Zhang et al., 2025a; Mao et al., 2016; Hao et al., 2025; Ji et al., 2025; Chow et al., 2025). We unify records, retaining source annotations and metrics; Appendix B.1 provides provenance and usage terms.

Human curation and quality control. We manually screened every retained sample for visual clarity, answerability, unique referents, and image–question–annotation consistency. Near-duplicates were removed; ambiguous or low-quality records were replaced from an independent candidate pool without using model scores. Release-wide checks verified readable media, required fields, and valid answer schemas. Appendix B.3 details review coverage and integrity checks.

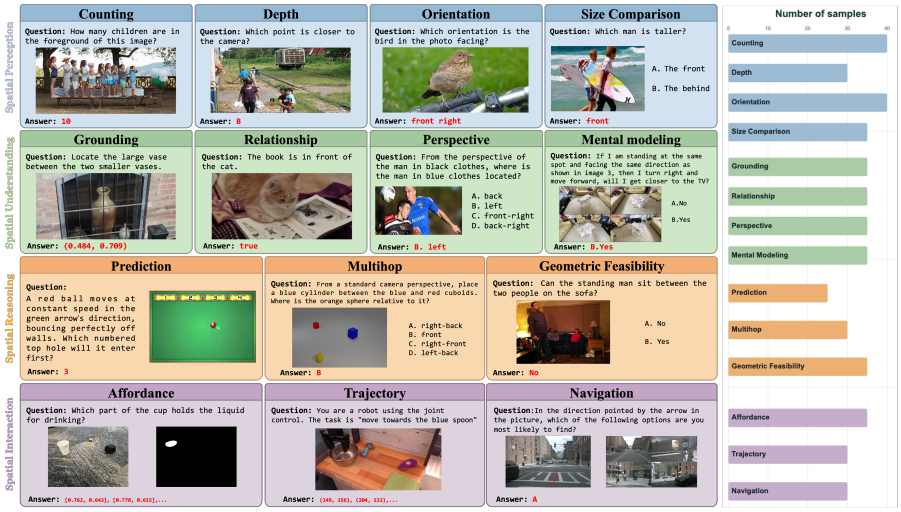


Figure 5: **SpatialGen-Bench: capability hierarchy and task coverage.** Tasks cover perception, understanding, reasoning, and interaction, with questions, answer targets, and sample counts.

4 Experiments

4.1 Models and Evaluation Settings

We evaluate 20 **Text Answering** and 11 **Visual Answering** systems, counting a unified model's interfaces separately. Visual tasks use fixed protocols shared across models without model-specific tuning; text systems retain the original answer space. Both interfaces share task targets and scoring rules. Missing or invalid outputs remain zero-scored in the denominator. Capability scores average constituent task scores; Overall averages all 14 tasks equally.

4.2 Main Results on SpatialGen-Bench

Table 1: **Text-answering results on SpatialGen-Bench (%).** Systems are grouped by training focus; bold marks the best text system per column. Human averages five full-benchmark testers.

Model	Perception	Understanding	Reasoning	Interaction	Overall
Human	97.29	85.71	77.99	87.60	87.77
General-purpose models					
GPT-5.4	74.61	44.29	55.81	**69.70**	**60.87**
SenseNova-Vision-7B	54.40	43.57	40.38	46.83	46.68
InternVL3.5-8B	51.70	36.43	**61.78**	35.87	46.10
LLaVA-OV-1.5-8B	54.23	37.86	39.40	45.56	44.52
Qwen3-VL-8B	59.79	37.14	46.42	31.43	44.38
GLM-4.6V-Flash	56.28	**45.71**	29.42	25.42	40.89
Janus-Pro-7B	39.20	26.43	35.42	24.13	31.51
Janus-1.3B	39.43	28.57	31.27	11.11	28.51
Spatial / embodied specialists					
Cosmos3-Nano	**76.82**	45.00	54.60	63.15	60.04
RynnBrain-8B	67.05	**45.71**	43.80	44.13	51.06
RoboBrain2.5-8B	58.39	34.29	47.19	50.03	47.31
SpaceR	61.04	31.43	52.60	37.38	45.70
SpaceOm	59.23	40.00	43.38	32.78	44.67
SpatialThinker-3B	57.65	41.43	39.21	36.37	44.50
GEM-2B	56.70	43.57	39.21	30.32	43.55
Cambrian-S-7B	68.01	21.43	46.69	32.22	42.46
VLM-3R-7B	56.19	21.43	37.21	31.21	36.84
Spatial-MLLM	32.77	30.00	32.72	36.51	32.77
SpatialRGPT-8B	38.90	27.86	29.23	29.52	31.66
SpatialBot-3B	33.13	32.86	31.61	10.79	27.94

Table 2: **Image-generation results on SpatialGen-Bench (%).** All models use protocolized visual answers. Bold marks the column best; metrics and the shared Human reference follow Table 1.

Model	Perception	Understanding	Reasoning	Interaction	Overall
GPT Image 2	69.58	**47.14**	**41.83**	**61.25**	**55.44**
Nano Banana 2	**72.92**	42.14	24.92	51.95	49.35
GPT-5 Image Mini	64.40	33.57	27.00	33.57	40.97
Seedream 4.5	51.43	35.71	28.69	36.37	38.84
JoyAI-Image	41.90	35.00	19.50	45.95	36.00
FLUX.2 [klein] 9B	47.68	26.43	22.80	42.87	35.25
SenseNova-Vision-7B	32.44	45.71	12.44	33.71	32.22
FLUX.2 [klein] 4B	41.93	27.14	14.17	30.60	29.33
OmniGen-v1	29.85	25.00	26.36	28.54	27.44
BAGEL-7B	27.38	23.57	24.44	34.10	27.10
Qwen-Image-Edit-2511	30.65	24.29	29.62	22.69	26.91

A human–model gap persists across both answer interfaces. Even the strongest system at each capability level trails the human reference in Tables 1 and 2. The largest gap is in spatial understanding, which requires relational and scene-level representations rather than isolated attribute judgments. Strong task-specific scores thus remain insufficient for consistently human-level performance across spatial capabilities; bootstrap intervals appear in Table 11.

Visual expression is competitive, but compositional reasoning remains a bottleneck. Image generators are competitive in spatial perception and spatial interaction, where answers often involve localized visual operations. Text-output VLMs lead most clearly in spatial reasoning, which requires combining relations and tracking state changes under spatial constraints. This contrast suggests that expressing a spatial judgment does not ensure consistency across dependent reasoning steps.

4.3 Capability Differences and Complementarity

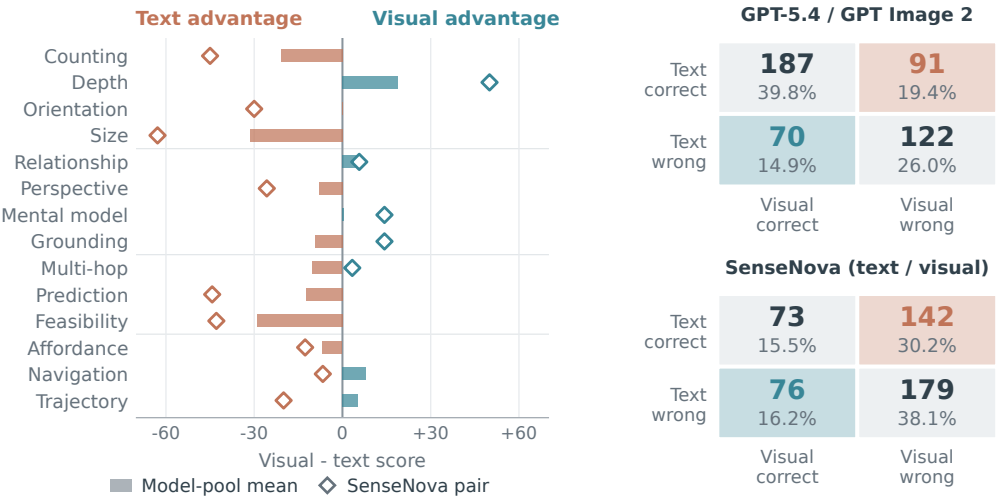


Figure 6: **Where does showing help?** Left: visual-minus-text scores for model-pool means and the SenseNova pair; positive values favor visual answering. Right: paired correctness for GPT-5.4/GPT Image 2 and SenseNova on all 470 items. Percentages are diagnostic shares, not task-macro scores.

Visual gains concentrate on directly expressed spatial states. In the left panel of Figure 6, both comparisons favor visual answering on depth and relationship, where depth fields and grounded overlays preserve spatial structure in the answer. Text leads on size, prediction, and geometric feasibility, which require comparisons, state updates, or constraint checks. Navigation reverses direction across comparisons, so visual expressibility alone does not ensure an advantage.

Stronger text models still miss cases solved visually. Both pairs in the right panel of Figure 6 contain visual-only successes despite higher text overall scores. Aggregate rankings thus conceal complementary successes on individual samples. These are behavioral differences, not evidence for a causal modality effect, since model pools and decoding paths differ.

4.4 Agentic Adaptation Across Benchmarks

ProVisE extends visual evaluation beyond SpatialGen-Bench. Agentic protocols cover six external benchmarks (Du et al., 2024; Jia et al., 2025; Liao et al., 2024; Song et al., 2025; Ray et al., 2024; Tang et al., 2025), with all 23 task units passing smoke validation. Separately, the GPT Image 2 pilots in Table 3 achieve 95.00% parse success across 300 visual outputs. Together, these results demonstrate adaptation across heterogeneous spatial tasks and answer formats while retaining the original task-specific scoring criteria.

Table 3: **Cross-benchmark transfer (%).** Models share 50 records per benchmark. Avg. is the unweighted six-benchmark mean; failed outputs score zero.

Model	EmbSpatial	OmniSpatial	Q-Spatial+	RoboSpatial	SAT	RoboAfford	Avg.
GPT Image 2	42.00	44.00	68.00	46.00	**74.00**	**64.00**	56.33
GPT-5.4	82.00	38.00	72.00	**58.00**	66.00	48.00	**60.67**
Qwen3-VL-8B	**84.00**	**46.00**	**74.00**	46.00	50.00	50.00	58.33
InternVL3.5-8B	78.00	44.00	56.00	56.00	42.00	36.00	52.00

4.5 Evaluator Variability and Protocol Stability

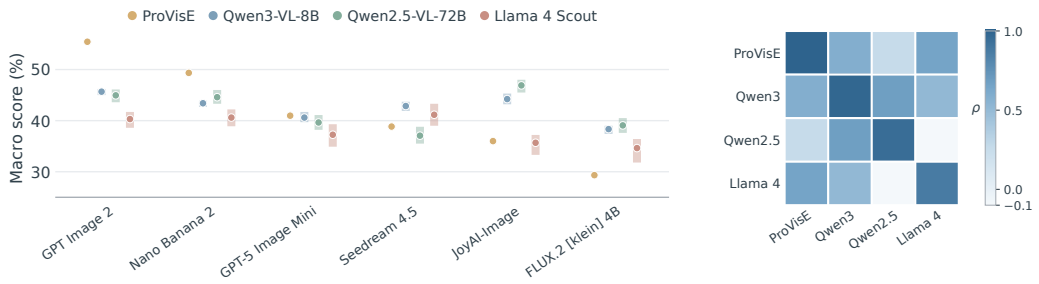


Figure 7: **Universal-parser variability and rank agreement.** Left: five-run VLM means and ranges on fixed outputs, with a fixed ProVisE score reference. Right: mean Spearman ρ within or across evaluators.

VLM judges make scores and rankings evaluator-dependent. We evaluate fixed outputs five times with Qwen3-VL-8B, Qwen2.5-VL-72B, and Llama 4 Scout. Figure 7 shows both disagreement across judges and variation within the same judge. With generated responses and scoring rules held fixed, these shifts reflect evaluator variability rather than changes in model performance, complicating comparisons based on a single judge or run. ProVisE supplies a fixed-score reference, not an independently verified ground truth; this experiment measures variability rather than decoding accuracy. Evaluator inputs and model references appear in Appendix E.2.

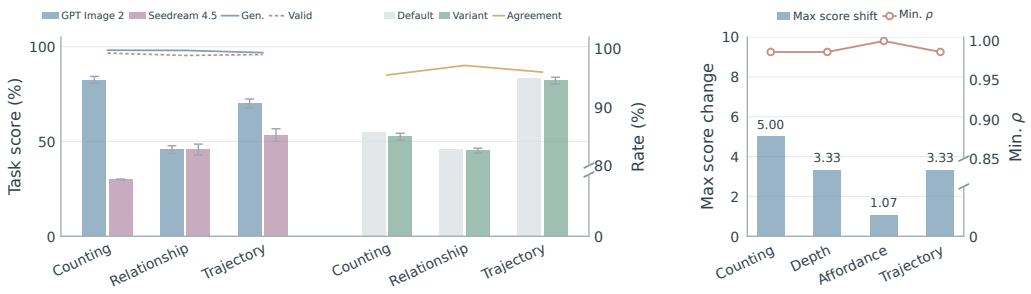


Figure 8: **Stability controls.** Left: generation repeats; middle: default protocols and variants; right: fixed-output sweeps. Error bars show five-run SD; parser bars show maximum shifts in percentage points.

Repeated generation and fixed-output parsing remain stable in the tested settings. In Figure 8, five generation runs on Counting, Relationship, and Trajectory yield high generation and valid-output rates with small score variation. Protocol variants show small run-to-run dispersion and mean shifts of at most 2.50 points. On fixed outputs, parser-setting and scoring-threshold sweeps largely preserve model rankings. Together, these controls support repeatable evaluation within the tested generation settings and low ranking sensitivity to the examined parser parameters.

4.6 Failure Attribution

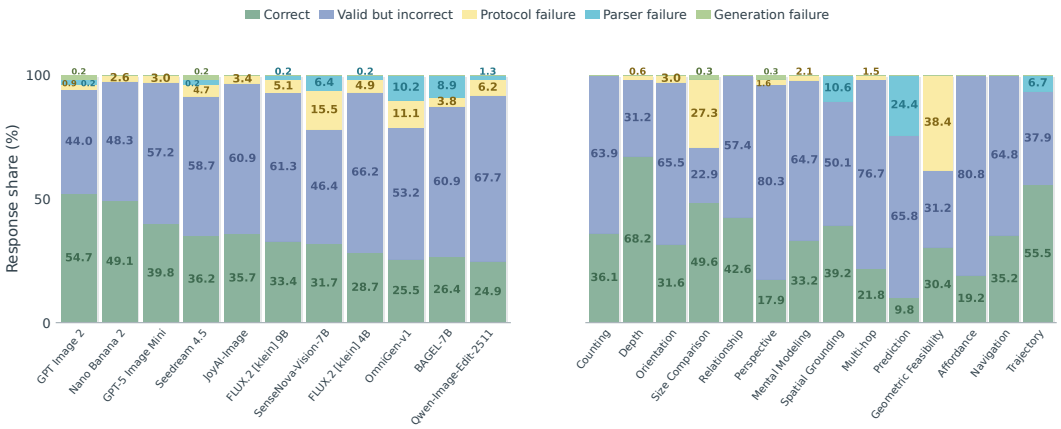


Figure 9: **Failure attribution by model and task.** Schematic breakdown of correct answers and four failure categories across models (left) and tasks (right); labels report percentages of all responses.

Most failures are valid but incorrect spatial predictions. The evaluation pipeline classifies 87.51% of failures in Figure 9 as valid but incorrect predictions. An image is produced and a task-compatible answer is recovered, yet the answer is wrong. Visual expression does not replace the comparisons, constraint composition, or state updates required to derive a correct spatial answer.

Protocol and parsing failures concentrate in different tasks. Protocol violations dominate failures in size comparison and geometric feasibility, where the output must show comparison or placement evidence. Prediction and spatial grounding instead account for most parser failures, where endpoints and locations must be encoded precisely enough for reliable recovery.

5 Limitations

Given the computational and monetary costs of image generation, SpatialGen-Bench prioritizes sample quality through item-level manual review, with broad task and answer-format coverage in a compact diagnostic set. Fine-grained rankings remain subject to sampling uncertainty. Cross-interface results characterize current systems, not causal modality effects, given differences in architecture, scale, and training data. Agentic construction depends on its backends and protocol library, while fallback parsing may inherit provider variability. Extending the static-image setting to video and closed-loop interaction requires protocol validation and temporal grounding.

6 Conclusion

We presented ProVisE, a benchmark-agnostic framework that parses protocol-constrained visual answers into structured predictions for task-specific scoring, and SpatialGen-Bench for evaluating heterogeneous spatial tasks. ProVisE's Agentic builder adapts protocols to new benchmarks. Results reveal task-dependent strengths: image-generation models are competitive when answers remain inspectable in pixels, whereas text-output VLMs lead on compositional relations and spatial transformations. Paired predictions show complementary successes, while most unsuccessful visual responses are classified as valid-but-incorrect by the evaluation pipeline. Future work will evaluate spatial intelligence in video-generation models and explore ways to combine the complementary strengths of visual generation and text-output VLMs for spatial problem solving.

AI USE STATEMENT

We used generative AI tools for manuscript revision, evaluation code development and testing, dataset consistency checks, and feedback on experimental methodology and result interpretation. The authors are responsible for verifying AI-assisted outputs and the paper's final content.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025. URL https://arxiv.org/abs/2509.23661.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models, 2023. URL https://arxiv.org/abs/2304.05390.

Hunar Batra, Haoqin Tu, Hardy Chen, Yuanze Lin, Cihang Xie, and Ronald Clark. Spatialthinker: Reinforcing scene graph-grounded spatial reasoning via dense rewards, 2025. URL https://arxiv.org/abs/2511.07403.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9490–9498. IEEE, 2025.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation, 2024. URL https://arxiv.org/abs/2310.18235.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding, 2025. URL https://arxiv.org/abs/2501.16411.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, Jiangpin Liu, Yunxuan Mao, Zhikai Wang, Yuqian Yuan, et al. Rynnbrain: Open embodied foundation models. *arXiv preprint arXiv:2602.14979*, 2026. URL https://arxiv.org/abs/2602.14979.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Shangzhe Di, Zhaokai Wang, and Weidi Xie. Are image generators zero-shot perceivers? a rigorous evaluation, 2026. URL https://arxiv.org/abs/2609.07884.

Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, 2024.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, et al. VLM-3R: Vision-language models augmented with instruction-aligned 3D reconstruction, 2025. URL https://arxiv.org/abs/2505.20279.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T Barron, et al. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026.

Yunfei Ge, Anbang Liu, Qineng Wang, Johnalbert Garnica, Jianwen Lyu, Zihan Wang, Reuben Tan, Jianfeng Gao, Ruohan Zhang, Yining Hong, Jiajun Wu, and Manling Li. MindTopo: Can foundation models reason in topological space?, 2026. URL https://arxiv.org/abs/2609.11900.

Robert Geirhos, Yuxuan Li, Thaddäus Wiedemer, Neha Kalibhat, Zi Wang, Mani Malek, Oyvind Tafjord, Kevin Swersky, Been Kim, and Priyank Jaini. Visual prompt engineering for video models, 2026. URL https://arxiv.org/abs/2607.25537.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022.

Xiaoyang Han, Jianhua Li, Kewang Deng, Zukai Chen, Xuanke Shi, Sihan Wang, Boxuan Li, Linyan Wang, Siyi Xie, Xin You, et al. Vision as unified multimodal generation, 2026. URL https://arxiv.org/abs/2607.06560.

Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang, Yanbiao Ma, Yunfeng Diao, Ziyu Jia, Wenbo Ding, Hangjun Ye, and Long Chen. Roboafford++: A generative ai-enhanced dataset for multimodal affordance learning in robotic manipulation and navigation. *arXiv preprint arXiv:2511.12436*, 2025.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025.

Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:2502.21257*, 2025.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025.

Joy Future Academy, JD. Joyai-image: Awakening spatial intelligence in unified multimodal understanding and generation. Technical report, 2026. URL https://joyai-image.s3.cn-north-1.jdcloud-oss.com/JoyAI-Image.pdf.

Ji Hyeok Jung, Eun Tae Kim, Seoyeon Kim, Joo Ho Lee, Bumsoo Kim, and Buru Chang. Is 'right'right? enhancing object orientation understanding in multimodal large language models through egocentric instruction tuning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14257–14267. IEEE, 2025.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025a.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025b.

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models, 2024. URL https://arxiv.org/abs/2409.09788.

Disheng Liu, Tuo Liang, Zhe Hu, Jierui Peng, Yiren Lu, Yi Xu, Yun Fu, and Yu Yin. Spatial intelligence in vision-language models: A comprehensive survey, 2026. URL https://doi.org/10.21203/rs.3.rs-8919250/v1.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. Ssr: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. *arXiv preprint arXiv:2505.12448*, 2025.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. doi: 10.1109/CVPR.2016.9. URL https://doi.org/10.1109/CVPR.2016.9.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI. Technical report, 2026. URL https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing MLLMs in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025. URL https://arxiv.org/abs/2504.01805.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3170–3180, 2023.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. Oral Presentation.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A. Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models, 2025. URL https://arxiv.org/abs/2503.19707.

Huajie Tan, Enshen Zhou, Zhiyu Li, Yijie Xu, Yuheng Ji, Xiansheng Chen, Cheng Chi, Pengwei Wang, Huizhu Jia, Yulong Ao, et al. Robobrain 2.5: Depth in sight, time in mind. *arXiv preprint arXiv:2601.14352*, 2026. URL https://arxiv.org/abs/2601.14352.

Yingbo Tang, Lingfeng Zhang, Shuyi Zhang, Yinuo Zhao, and Xiaoshuai Hao. Roboafford: A dataset and benchmark for enhancing object and spatial affordance learning in robot manipulation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12706–12713, 2025. doi: 10.1145/3746027.3758209. URL https://doi.org/10.1145/3746027.3758209.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. Mindcube: Spatial mental modeling from limited views, 2026a. URL https://arxiv.org/abs/2506.21458.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. Everything in its place: Benchmarking spatial intelligence of text-to-image models. *arXiv preprint arXiv:2601.20354*, 2026b.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners, 2025. URL https://arxiv.org/abs/2509.20328.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation, 2024. URL https://arxiv.org/abs/2410.13848.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025b.

Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. Spatialscore: Towards comprehensive evaluation for spatial intelligence. *arXiv preprint arXiv:2505.17012*, 2025c.

Jialong Wu, Xiaoying Zhang, Hongyi Yuan, Xiangcheng Zhang, Tianhao Huang, Changjing He, Chaoyi Deng, Renrui Zhang, Youbin Wu, and Mingsheng Long. Visual generation unlocks human-like reasoning through multimodal world models, 2026. URL https://arxiv.org/abs/2601.19834.

Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. Vsp: Diagnosing the dual challenges of perception and reasoning in spatial planning tasks for mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2270–2280, October 2025d.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation, 2024. URL https://arxiv.org/abs/2409.11340.

Yuxi Xiao, Longfei Li, Shen Yan, Xinhang Liu, Sida Peng, Yunchao Wei, Xiaowei Zhou, and Bingyi Kang. Spatialtree: How spatial abilities branch out in mllms, 2025. URL https://arxiv.org/abs/2512.20617.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, Gege Qi, and Yunjian Zhang. Spatialbench: Benchmarking multimodal large language models for spatial cognition, 2026. URL https://arxiv.org/abs/2511.21471.

Cheng Yang, Chufan Shi, Bo Shui, Yaokang Wu, Muzi Tao, Huijuan Wang, Ivan Yee Lee, Yong Liu, Xuezhe Ma, and Taylor Berg-Kirkpatrick. UReason: Benchmarking reasoning-to-generation alignment in unified multimodal models, 2026. URL https://arxiv.org/abs/2602.08336.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025a.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025b. URL https://arxiv.org/abs/2511.04670.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *ICLR*, 2025c.

Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zaibin Zhang, et al. How far are vlms from visual spatial intelligence? a benchmark-driven perspective. *arXiv preprint arXiv:2509.18905*, 2025.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. *ACL*, 2025a.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. Spinbench: Perspective and rotation as a lens on spatial reasoning in vlms. *arXiv preprint arXiv:2509.25390*, 2025b.

Ruowen Zhao, Bangguo Li, Zuyan Liu, Yinan Liang, Junliang Ye, Fangfu Liu, Diankun Wu, Zhengyi Wang, Xumin Yu, Yongming Rao, Han Hu, and Jun Zhu. GEM: Generative supervision helps embodied intelligence. *arXiv preprint arXiv:2605.28548*, 2026. URL https://arxiv.org/abs/2605.28548.

Xu Zheng, Zihao Dongfang, Lutao Jiang, Boyuan Zheng, Yulong Guo, Zhenquan Zhang, Giuliano Albanese, Runyi Yang, Mengjiao Ma, Zixin Zhang, et al. Multimodal spatial reasoning in the large model era: A survey and benchmarks. *arXiv preprint arXiv:2510.25760*, 2025.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025.

APPENDIX

- **Section A.** Related work on spatial benchmarks and generative evaluation.
- **Section B.** Dataset sources and usage terms, the released instance format, and automatic integrity checks alongside manual quality review.
- **Section C.** Visual protocol design, parser rules, Agentic construction routes, and validation coverage across external spatial benchmarks.
- **Section D.** Human references, model and prompting settings, task-specific metrics, score aggregation, and statistical uncertainty estimates.
- **Section E.** External transfer, evaluator sensitivity, Depth ablation, and controls.
- **Section F.** Qualitative text–visual comparisons with source images, model responses, complete questions, and reference answers for individual examples.

A RELATED WORK

Spatial reasoning benchmarks and models. Spatial evaluation has progressed from single-image relation and perception tests, including VSR, BLINK, and ViewSpatial (Liu et al., 2023; Fu et al., 2024; Li et al., 2025a), to broader taxonomies such as SpatialBench, OmniSpatial, and SpatialScore (Xu et al., 2026; Jia et al., 2025; Wu et al., 2025c). Planning and embodied suites probe action-oriented reasoning (Ray et al., 2024; Du et al., 2024; Wu et al., 2025d), while VSI-Bench, MindCube, and MMSI extend evaluation to video, sparse views, and multiple images (Yang et al., 2025a; Wang et al., 2026a; Yang et al., 2025c). MindTopo tests topological reasoning and interactive planning, including agents augmented with visual generation (Ge et al., 2026). Specialized models further improve grounding through geometric supervision and spatial post-training (Chen et al., 2024; Cheng et al., 2024; Wu et al., 2025b; Ouyang et al., 2025).

Generative and unified visual systems. Unified models such as Janus, BAGEL, Emu3.5, and Show-o2 combine visual understanding and generation (Chen et al., 2025; Deng et al., 2025; Cui et al., 2025; Xie et al., 2025). Conventional generative evaluations measure prompt fidelity and composition (Hu et al., 2023; Ghosh et al., 2023; Huang et al., 2025; Cho et al., 2024). More directly, ProbeGen uses visual prompts and task-specific extractors for zero-shot depth, segmentation, and counting (Di et al., 2026), while UReason diagnoses whether textual reasoning is realized in generated images (Yang et al., 2026). VIPE improves video reasoning by modifying input images (Geirhos et al., 2026), whereas ProVisE constrains output answers. Our focus is a reusable, validated generation–parser protocol layer spanning heterogeneous spatial task contracts.

Table 4 compares spatial benchmarks, including SpatialTree and SpatialGenEval (Xiao et al., 2025; Wang et al., 2026b). SpatialGen-Bench supports hierarchical diagnosis and diverse answer forms under shared text–visual task semantics and metrics.

Table 4: Representative spatial benchmarks by scope, scale, hierarchy, answer forms, and shared text–visual evaluation.

Benchmark	Scope	Scale	Hier.	Ans.	Shared eval.
BLINK	Visual perception	3.8K / 14 tasks	✗	✗	✗
ViewSpatial-Bench	Multi-view localization	5.7K+ / 5 types	△	✗	✗
MindCube	Mental modeling	21K+ / 3.2K images	✗	△	✗
OmniSpatial	Broad VLM spatial reasoning	8K+ / 50 subcats.	✓	✗	✗
SpatialScore	Multimodal spatial intelligence	5K / 30 tasks	✓	△	✗
SpatialTree	Hierarchical spatial ability	6.3K / 11 tasks	✓	△	✗
SpatialGenEval	T2I spatial reasoning	1,230 prompts	✓	✗	△
SpatialGen-Bench	**Generative spatial cognition**	**470 / 14 tasks**	✓	✓	✓

✓: full; △: partial; ✗: not targeted. *Hier.*: hierarchical diagnosis; *Ans.*: multiple evaluated answer forms; *Shared*: text and visual answers use shared task semantics and metrics.

B Dataset Construction and Quality Control

B.1 Data Sources and Usage Terms

Table 5 lists task-level sources, sample counts, usage terms, and official links. All samples retain provenance in `source`, `source_id`, and `metadata_json`. Usage terms follow official project pages, repositories, or Hugging Face dataset cards and may not cover every upstream media asset. The table records provenance without reinterpreting source rights. "N/A" indicates that no explicit license field was found, not that the source is unrestricted.

Table 5: Data sources, sample counts (N), and usage terms for SpatialGen-Bench. Links identify the official project, repository, or dataset card for each source.

Task	N	**Source benchmark**	**Terms**	**Source**
Counting	40	CountBench	CC-BY-4.0	Project
Depth	30	BLINK	Apache-2.0	HF
Orientation	40	EgoOrientBench	MIT	HF
Object Size	35	SPHERE-VLM	COCO terms	HF
Relationship	35	VSR	Apache-2.0	GitHub
Perspective	35	ViewSpatial-Bench	Apache-2.0	HF
Mental Modeling	35	MindCube	MIT	HF
Spatial Grounding	35	RefCOCOg	N/A	HF
Multi-hop	30	VisWorld-Eval	N/A	HF
Prediction	25	VisWorld-Eval	N/A	HF
Geometric Feasibility	35	SPHERE-VLM	COCO terms	HF
Affordance	35	RoboAfford-Eval	CC-BY-4.0	HF
Navigation	30	PhysBench	Apache-2.0	HF
Trajectory	30	ShareRobot-Bench	Apache-2.0	HF
Total	**470**	*14 spatial tasks*		

B.2 Unified Instance Format

Each JSONL row normalizes field names, media paths, option IDs, and labels while retaining source answer semantics. Table 6 separates model inputs, evaluator targets, and provenance in the public schema. Prediction's `num_holes` is scoring metadata inside decoded `metadata_json`, not a top-level column. Dataset revision: `2dfb851d0f80baec198b6b43858b76d9a05ae11b`.

Table 6: Released SpatialGen-Bench JSONL schema. Paths are relative to the dataset root. `mask_path` and evaluator metadata are scoring targets, not model inputs.

Field	**Type**	**Meaning**
`id`	string	Globally unique sample identifier.
`capability`	string	One of perception, understanding, reasoning, or interaction.
`task`	string	One of the 14 normalized subtask names.
`image_path`	string	Relative path to the primary input image.
`media_paths`	list[string]	Ordered model-input media, including any required views or candidates.
`mask_path`	string	Optional evaluator-only target-mask path; empty when unused.
`question`	string	Normalized instruction shown to the model.
`answer`	string	Canonical serialized answer in the source task's answer space.
`answer_type`	string	Declared answer schema, such as integer, choice, boolean, point, mask, or path.
`choices`	list[string]	Ordered candidates; empty when the task is not choice-based.
`source`	string	Source benchmark identifier.
`source_id`	string	Identifier retained from the source benchmark.
`metadata_json`	string	JSON-encoded provenance and task-specific auxiliary metadata.

B.3 Data Integrity and Review Coverage

Deterministic integrity checks verify unique IDs, relative paths, readable primary media, answer-schema validity, placeholder-like assets, and required task inputs or evaluator annotations across the released split; every record passes. Manual review covers every item and checks that its instruction, visual evidence, annotation, and scoring contract remain mutually consistent.

C Protocol and Agentic Construction Details

This section specifies protocol constraints, parser rules, Agentic routes, and validation evidence. Exact generation prompts and executable protocol configurations are available in the released code.

C.1 Protocol Design

Per-sample prompts substitute only declared fields: `question`, `choices`, object references, candidate states, or trajectory start points. New Agentic protocols follow four invariants:

- **Task-grounded evidence.** The response must expose a location, region, relation, state, measurement, or path that is meaningful for the source task.
- **Minimal intervention.** Editing protocols require preservation of unrelated scene content; synthesis protocols specify the scene and viewpoint constraints that generated outputs must satisfy.
- **Executable parser.** Every prompt is paired with a declared deterministic, fixed-auxiliary, or generated-image-only parser and explicit invalid-output conditions.
- **No answer shortcut.** The builder rejects option letters, corner codes, verdict symbols, or prose-only outputs when these merely encode the answer without spatial evidence.

SpatialGen-Bench uses fixed-slot protocols for Mental Modeling and Multi-hop, and a trajectory-guided endpoint for Prediction. The latter draws one reflected yellow path and circles one destination hole; a fixed auxiliary VLM reads the hole index and candidate count from the generated image alone. The scoring denominator uses source-scene metadata, not the decoded count. These evaluated protocols differ from Agentic routes for external benchmarks. Figures 2 and 3 summarize the protocol pool and representative outputs.

C.2 Parser Rules and Thresholds

Table 8 separates reusable parsers from those instantiated or selected during Agentic construction. It records each operational rule, decoded answer type, and invalid-output boundary, and identifies the CLIP, OCR, or VLM components used for answer recovery from generated responses.

C.3 Agentic Construction Details and Coverage

For each task, the construction Agent receives answer-hidden examples, input and answer schemas, choices, the metric contract, and registered protocols and Parser Ops. It proposes generation guidance, an executable parser, invalid-output rules, and metric mapping. Static checks reject unknown operators, contract mismatches, missing media, and answer leakage. Smoke tests require successful generation, recoverable visual evidence, repeated-parse agreement, and metric-compatible predictions. Ground truth and correctness feedback remain hidden during construction.

Accepted protocols are fixed and reused across target models without rerouting. Table 7 summarizes construction and pilot validity: after answer-contract audit, all 23 task units pass smoke validation, with four Reuse, 11 Build, and eight generated-image-only Fallback routes.

Table 7: Agentic protocol coverage after contract audit on six spatial benchmarks. Pilot and Parse summarize archived GPT Image 2 runs; Smoke records protocol validation.

Benchmark	Records	Tasks	Reuse	Build	Fallback	Pilot	Parse (%)	Smoke
EmbSpatial-Bench	3,640	1	0	1	0	50	100.00	1/1
OmniSpatial	1,533	5	0	0	5	50	82.00	5/5
Q-Spatial+	101	3	0	3	0	50	100.00	3/3
RoboSpatial-Home	350	3	1	2	0	50	100.00	3/3
SAT	150	8	0	5	3	50	92.00	8/8
RoboAfford	338	3	3	0	0	50	96.00	3/3
Total	**6,112**	**23**	**4**	**11**	**8**	**300**	**95.00**	**23/23**

Table 8: Parser rules for reusable and Agentic protocols. HSV ranges use OpenCV hue values; fixed auxiliary models and generated-image-only Fallback are named explicitly.

Protocol	Parser rule	Prediction	Invalid-output checks / reliability risks
Reusable SpatialGen-Bench protocols			
Instance Marking	Threshold green pixels (H=35–90, S/V$\geq$80), apply 3×3 morphology, and retain plausible connected components by area and solidity.	Integer count.	No valid component; risks: merged markers or incorrect target grounding.
Direction Grid	Threshold blue pixels (H=100–130, S$\geq$120, V$\geq$50), aggregate evidence over the eight outer cells of a 3×3 grid, and require at least 50 blue pixels.	Direction label.	No active cell, center-cell marking, or conflicting evidence across cells.
Fixed-Slot Coding	Threshold magenta pixels (H=140–175, S/V$\geq$80) in the protocol's predefined slots and require at least 30 pixels.	Choice label or numeric slot.	Missing mark, multiple active slots, or a mark outside the code layout.
Binary Relation	Recover green subject and blue object overlays using fixed HSV ranges and minimum color support.	Boolean label.	Absent required masks; semantic risk: incorrect entity grounding; color risk: spurious scene signals.
Relative Depth	Sample 5×5 local grayscale means at the benchmark A/B locations and select the brighter location as closer.	A/B depth label.	Unreadable depth map, invalid locations, indistinguishable local values, or violation of the declared brightness convention.
Region Masking	Resize to the target resolution, threshold grayscale at > 127, and pass the binary mask to the source metric.	Binary mask.	Empty or unreadable mask, missing target metadata, or non-binary artifacts.
State Matching	Encode the generated state and candidate states with fixed CLIP ViT-B/32 and select the maximum cosine similarity.	Candidate label.	Missing candidates, unreadable images, or a generated state not comparable to the candidates.
Trajectory Drawing	Select the red component nearest the public start point, skeletonize and order it, then sample 20 normalized path points.	Normalized polyline.	Unreadable image, no eligible red component, or no recovered path; archived invalid outputs remain zero.
Trajectory-Guided Endpoint	A fixed auxiliary VLM reads the numbered hole enclosed by the blue circle and the total hole count from the generated yellow-trajectory response alone.	Indexed endpoint.	Missing or ambiguous endpoint circle, unreadable hole numbering, or a non-positive endpoint/count.
Point Marker	Threshold novel cyan pixels (H=80–100, S/V$\geq$70), select one compact component, and normalize its centroid.	Normalized point.	Missing marker, multiple similarly prominent marks, or a component too small or diffuse to localize.
Agentic Build and Fallback parsers			
Marked-Object Choice	Align the output to the source, subtract pre-existing magenta, recover one contour, crop the enclosed source object, and match it to textual choices with fixed CLIP.	Choice label.	Missing or ambiguous contour, invalid crop, or insufficient CLIP separation.
Grounded Dimension	Recover magenta/cyan anchors and the yellow dimension line by color differencing; local OCR parses the adjacent scalar-unit tag.	Scalar and unit.	Required anchors or the dimension line are missing; OCR fails; the unit is unsupported; or the label is detached.
Dual-Anchor Relation	Recover role-colored subject and reference masks, then compute the configured horizontal, vertical, dominant-axis, or distance relation.	Choice label.	Either anchor is missing or ambiguous, or the recovered geometry cannot map uniquely to a choice.
Relation-Zone Boolean	Recover subject, reference, and valid-relation-region masks; predict true exactly when the subject lies in the rendered valid region.	Boolean label.	Missing anchors, degenerate relation region, or a verdict-only symbol instead of spatial evidence.
Size Fallback	A constrained parser maps only visible object-comparison or measurement evidence in the generated response to the A/B schema.	A/B size label.	Bare label, missing comparison evidence, contradictory marks, or ambiguous referenced objects.
Feasibility Fallback	A constrained parser reads only the depicted placement, clearance, overlap, or collision state.	Boolean label.	Attempted placement absent, unrealistic or ambiguous scale, verdict symbol, or no visible fit/collision evidence.
Generic Fallback	A task-specific parser prompt recovers the declared answer schema solely from the generated response; choices may only map visible evidence to labels.	Contract-specific answer.	Evidence absent or contradictory, output contains only prose/label, or reading would require the source image or original question.

D EXPERIMENTAL SETTINGS AND SCORING

This section specifies the evaluation interface used after the benchmark split and visual protocols are fixed. It covers text-output prompts, model runtime settings, and score aggregation. Protocol constraints and the released generation-prompt locations are documented in Appendix C.

D.1 HUMAN EVALUATION

Five human testers each complete all 470 samples and are scored with the same task-specific metrics used for model outputs. Capability and Overall values use the same macro-aggregation in Eq. 5; Table 9 reports each tester and their mean. The five-tester mean supplies the Human row throughout the result tables. This task evaluation is separate from the source-data quality audit in Appendix B.3.

Table 9: Five human testers on all 470 samples; task-macro scores are reported with their mean.

Reference	N	Perception	Understanding	Reasoning	Interaction	Overall
Tester 1	470	97.41	88.57	80.32	90.29	89.70
Tester 2	470	97.41	85.71	78.25	87.60	87.86
Tester 3	470	96.79	82.86	74.86	84.92	85.56
Tester 4	470	97.41	85.71	78.25	87.69	87.88
Tester 5	470	97.41	85.71	78.25	87.52	87.84
Human Mean	5×470	97.29	85.71	77.99	87.60	87.77

D.2 TEXT-OUTPUT VLM PROMPT FORMATS

Text-output VLMs are prompted to return only the final answer in the original benchmark answer space. The response constraint is matched to the target answer type: a single integer for counting, A/B for relative depth and object size, a normalized direction label for orientation and perspective, `true`/`false` for relation and feasibility decisions, an option label for multiple-choice tasks, normalized points for spatial grounding, and the structured textual form required by the corresponding original evaluator for mask or trajectory targets. This keeps the text baseline and the visual protocol scored against the same target format and task-specific scoring rules for every task.

D.3 MODEL AND SCORING SETTINGS

Local experiments use one NVIDIA RTX PRO 6000 GPU; APIs use the listed endpoints. Local VLMs decode deterministically at batch size one. The main evaluation retains at most one visual response per model and sample, without correctness-based retries or manual selection; cached outputs are reused. Edits preserve source aspect ratio, target 1K when supported, and resize to source resolution before parsing. Missing or invalid outputs receive the declared score. Table 10 records model families and runtime settings used throughout benchmark evaluation.

D.4 SCORE AGGREGATION

All parsed predictions are first converted into per-sample scores $s_i \in [0, 1]$. Counting, relative-depth choice, orientation, object size, relationship, perspective, mental modeling, multihop reasoning, geometric feasibility, and navigation use exact-match accuracy after answer normalization. The remaining tasks use the task-specific scores below; all reported numbers are percentages.

For spatial grounding, point-in-mask accuracy compares the reference target mask M with the normalized point $\hat{p}$ returned by the text model or recovered from the visual marker:

$$s_{\text{ground}}(\hat{p}, M) = \mathbf{1}[M(\hat{p}) > 0]. \tag{1}$$

For affordance grounding, M denotes the annotated target region. The prediction $\hat{M}$ is the parsed binary mask for image-output models, or the set of predicted points rasterized to image coordinates for text-output VLMs. The per-sample score is precision:

$$s_{\text{aff}}(\hat{M}, M) = \frac{|\hat{M} \cap M|}{|\hat{M}| + \epsilon}, \tag{2}$$

where ϵ prevents division by zero for empty predictions. This score measures how much of the predicted actionable support lies inside the annotated region; the reported benchmark score uses the mean precision across samples in that benchmark subtask.

For ball-state prediction, $\hat{y}$ and y are the predicted and ground-truth hole indices. H is the number of candidate holes annotated in the source scene and stored in the dataset metadata. The same per-item H is used for text and visual predictions:

$$s_{\text{pred}}(\hat{y}, y, H) = \begin{cases} 1, & |\hat{y} - y| = 0, \\ 0.8 - \frac{1}{H}, & |\hat{y} - y| = 1, \\ 0, & |\hat{y} - y| \geq 2, \end{cases} \tag{3}$$

Equation 3 gives full credit to correct endpoints and partial credit to adjacent endpoints. Invalid or missing predictions receive zero credit.

For trajectory planning, let $\hat{P} = (\hat{p}_1, \ldots, \hat{p}_n)$ and $P = (p_1, \ldots, p_m)$ be the parsed and ground-truth point sequences after normalization to image coordinates. We compute Discrete Fréchet Distance and threshold it as in Eq. 4 for every sample:

$$d_{\text{DFD}}(\hat{P}, P) = \min_{C} \max_{(\hat{p},p)\in C} \|\hat{p} - p\|_2, \qquad s_{\text{traj}} = \mathbf{1}[d_{\text{DFD}}(\hat{P}, P) < 0.4], \tag{4}$$

where C is a monotone coupling; the trajectory score is binary geometric-path agreement. Visual paths use 20 arc-length samples from the red component nearest the public start point, while reference paths retain their original vertices. Empty paths receive zero. The same fixed extraction is used for all visual systems and the support-budget baseline.

Equation 5 defines the aggregation rule. For task t, N_t is the number of samples; for capability c, $\mathcal{T}_c$ is its subtask set:

$$S_t = \frac{1}{N_t} \sum_{i=1}^{N_t} s_i, \qquad S_c = \frac{1}{|\mathcal{T}_c|} \sum_{t\in\mathcal{T}_c} S_t, \qquad S_{\text{overall}} = \frac{1}{14} \sum_{t=1}^{14} S_t. \tag{5}$$

Task means include all assigned samples; capability and Overall scores use unweighted task averages. The binary diagnostics in Figures 6 and 9 use precision ≥ 0.5 for Affordance and exact endpoints for Prediction. These diagnostic shares do not replace the task-specific scores.

D.5 Statistical Uncertainty

We estimate sample-selection uncertainty on the complete benchmark split. Each of 10,000 replicates resamples examples with replacement within every task, recomputes task means, and then applies the same task-macro aggregation as the leaderboard. Intervals are percentile 95% confidence intervals for fixed cached outputs; they do not capture generation stochasticity, provider-side model updates, or other deployment drift.

Table 10: Complete model list and runtime settings for SpatialGen-Bench evaluation.

Model	Runtime setting
Text-output VLMs *Direct answers in the original benchmark answer space.*	
Qwen3-VL-8B (Bai et al., 2025a)	Local inference; deterministic decoding; batch size 1; max 512 new tokens.
InternVL3.5-8B (Wang et al., 2025)	Local BF16 inference; dynamic image tiling; deterministic decoding.
LLaVA-OV-1.5-8B (An et al., 2025)	Local BF16 inference; deterministic decoding.
GLM-4.6V-Flash (Team et al., 2025)	Local BF16 inference; deterministic decoding.
GPT-5.4	OpenAI-compatible chat API; request model `gpt-5.4`; direct structured answer; deterministic request settings where supported.
Cosmos3-Nano; RynnBrain-8B; RoboBrain2.5-8B (NVIDIA, 2026; Dang et al., 2026; Tan et al., 2026)	Local multimodal inference; model-native preprocessing; deterministic decoding; batch size 1.
SpaceR; Cambrian-S-7B; GEM-2B; SpatialThinker-3B (Ouyang et al., 2025; Yang et al., 2025b; Zhao et al., 2026; Batra et al., 2025); SpaceOm	Local spatial-specialized VLM inference; released checkpoint and processor; deterministic decoding.
VLM-3R-7B; SpatialRGPT-8B; Spatial-MLLM; SpatialBot-3B (Fan et al., 2025; Cheng et al., 2024; Wu et al., 2025b; Cai et al., 2025)	Local spatial-specialized VLM inference; released checkpoint and task-compatible prompt format.
SenseNova-Vision-7B; Janus-Pro-7B; Janus-1.3B (Han et al., 2026; Chen et al., 2025; Wu et al., 2024)	Local unified-model text interface; deterministic decoding; batch size 1.
Output processing: raw responses are normalized to the original answer format; invalid formats are scored as incorrect.	
Image-output generative models *Protocol-constrained visual responses parsed back to structured predictions.*	
JoyAI-Image (Joy Future Academy, JD, 2026)	Local image editing; seed 42; 50 inference steps; guidance 5.0; base size 1024.
FLUX.2 [klein] 4B	Local image editing; seed 42; 4 inference steps; guidance 1.0; max side 1024.
FLUX.2 [klein] 9B	Local image editing with the 9B checkpoint; one protocol-constrained image per sample.
SenseNova-Vision-7B; BAGEL-7B (Han et al., 2026; Deng et al., 2025)	Local unified-model visual interface; one protocol-constrained image per sample.
Qwen-Image-Edit-2511 (Wu et al., 2025a)	Local BF16 image editing; model offload; seed 0; 40 steps; true CFG 4.0; max side 1024.
OmniGen-v1 (Xiao et al., 2024)	Local BF16 image editing; seed 42; 50 steps; guidance 2.5; image guidance 1.6; max side 1024.
Seedream 4.5	Provider-compatible image API; request route `bytedance-seed/seedream-4.5`; one protocol-constrained image per sample.
GPT-5 Image Mini	OpenRouter image route `openai/gpt-5-image-mini`; provider-composed request alias; one protocol-constrained image per sample.
Nano Banana 2	Provider-compatible image route `google/gemini-3.1-flash-image-preview` (dated alias `google/gemini-3.1-flash-image-preview-20260226`); one image per sample.
GPT Image 2	OpenAI-compatible image-edit API; request model `gpt-image-2`; returned alias `gpt-image-2-codex`; source image plus protocol prompt; one image per sample.
Output processing: one output image per sample; outputs are resized to the source resolution if needed and parsed automatically with the assigned protocol. Missing, unreadable, or unparseable outputs are marked invalid.	

Table 11: Bootstrap 95% intervals for all model-interface systems; entries are point estimate [2.5th, 97.5th percentile].

Model	Overall	Perception	Understanding	Reasoning	Interaction
Text Answering					
GPT-5.4	60.87 [56.96, 64.76]	74.61 [67.26, 81.58]	44.29 [35.71, 52.86]	55.81 [46.66, 64.57]	69.70 [62.13, 76.71]
Cosmos3-Nano	60.04 [56.32, 63.71]	76.82 [70.15, 82.98]	45.00 [37.86, 52.14]	54.60 [45.66, 63.66]	63.15 [56.09, 70.31]
RynnBrain-8B	51.06 [47.20, 54.82]	67.05 [60.00, 73.81]	45.71 [38.57, 52.86]	43.80 [34.32, 53.15]	44.13 [36.19, 52.54]
RoboBrain2.5-8B	47.31 [43.75, 50.96]	58.39 [51.93, 64.61]	34.29 [27.86, 40.71]	47.19 [38.12, 56.22]	50.03 [42.01, 58.10]
SenseNova-Vision-7B	46.68 [42.53, 50.90]	54.40 [46.93, 62.05]	43.57 [36.43, 51.43]	40.38 [31.00, 49.68]	46.83 [37.62, 56.03]
InternVL3.5-8B	46.10 [42.41, 49.90]	51.70 [45.45, 57.92]	36.43 [29.29, 43.57]	61.78 [53.05, 70.55]	35.87 [27.30, 44.76]
SpaceR	45.70 [41.68, 49.69]	61.04 [53.51, 68.51]	31.43 [24.29, 38.57]	52.60 [43.13, 61.80]	37.38 [29.13, 45.63]
SpaceOm	44.67 [40.61, 48.71]	59.23 [51.64, 66.82]	40.00 [32.86, 47.14]	43.38 [34.45, 52.71]	32.78 [24.21, 41.11]
LLaVA-OV-1.5-8B	44.52 [40.57, 48.52]	54.23 [47.50, 60.86]	37.86 [30.00, 45.71]	39.40 [30.42, 48.75]	45.56 [37.30, 53.65]
SpatialThinker-3B	44.50 [40.52, 48.60]	57.65 [50.33, 64.91]	41.43 [34.29, 49.29]	39.21 [30.00, 48.46]	36.37 [27.98, 44.92]
Qwen3-VL-8B	44.38 [40.41, 48.32]	59.79 [52.17, 67.35]	37.14 [30.00, 44.29]	46.42 [37.45, 55.27]	31.43 [23.33, 39.68]
GEM-2B	43.55 [39.48, 47.68]	56.70 [48.84, 64.35]	43.57 [35.71, 51.43]	39.21 [30.59, 48.16]	30.32 [22.22, 38.89]
Cambrian-S-7B	42.46 [38.59, 46.33]	68.01 [60.65, 75.18]	21.43 [15.00, 28.57]	46.69 [38.08, 55.53]	32.22 [23.81, 40.95]
GLM-4.6V-Flash	40.89 [37.29, 44.66]	56.28 [49.70, 62.92]	45.71 [37.86, 53.57]	29.42 [22.53, 36.51]	25.42 [17.98, 33.42]
VLM-3R-7B	36.84 [33.16, 40.53]	56.19 [49.46, 62.86]	21.43 [15.00, 28.57]	37.21 [28.83, 45.84]	31.21 [22.69, 39.75]
Spatial-MLLM	32.77 [29.15, 36.47]	32.77 [27.53, 38.27]	30.00 [22.86, 37.86]	32.72 [24.29, 41.53]	36.51 [27.94, 44.92]
SpatialRGPT-8B	31.66 [28.00, 35.43]	38.90 [32.38, 45.39]	27.86 [20.71, 35.00]	29.23 [21.19, 37.73]	29.52 [22.06, 37.14]
Janus-Pro-7B	31.51 [27.99, 35.14]	39.20 [32.98, 45.09]	26.43 [19.29, 33.57]	35.42 [27.13, 43.69]	24.13 [16.67, 31.75]
Janus-1.3B	28.51 [25.02, 32.06]	39.43 [32.26, 46.61]	28.57 [22.14, 35.00]	31.27 [23.37, 39.20]	11.11 [5.56, 17.78]
SpatialBot-3B	27.94 [24.36, 31.55]	33.13 [26.85, 39.55]	32.86 [25.00, 40.71]	31.61 [23.12, 40.18]	10.79 [5.40, 16.98]
Visual Answering					
GPT Image 2	55.44 [51.39, 59.54]	69.58 [62.11, 76.76]	47.14 [39.29, 55.00]	41.83 [32.62, 51.25]	61.25 [53.13, 69.36]
Nano Banana 2	49.35 [45.39, 53.40]	72.92 [65.74, 79.64]	42.14 [34.29, 50.00]	24.92 [16.95, 33.40]	51.95 [43.00, 61.16]
GPT-5 Image Mini	40.97 [37.00, 44.96]	64.40 [56.73, 71.90]	33.57 [25.71, 41.43]	27.00 [18.95, 35.22]	33.57 [25.14, 42.16]
Seedream 4.5	38.84 [35.04, 42.65]	51.43 [44.29, 58.69]	35.71 [27.86, 43.57]	28.69 [20.86, 36.78]	36.37 [28.22, 44.60]
JoyAI-Image	36.00 [32.15, 39.82]	41.90 [34.37, 49.35]	35.00 [27.86, 42.86]	19.50 [12.07, 27.41]	45.95 [37.39, 54.58]
FLUX.2 [klein] 9B	35.25 [31.47, 39.01]	47.68 [40.54, 54.64]	26.43 [20.00, 33.57]	22.80 [14.70, 31.00]	42.87 [33.87, 51.79]
SenseNova-Vision-7B	32.22 [28.81, 35.65]	32.44 [26.52, 38.33]	45.71 [39.29, 52.86]	12.44 [6.67, 18.89]	33.71 [25.42, 42.26]
FLUX.2 [klein] 4B	29.33 [25.62, 33.22]	41.93 [34.61, 49.52]	27.14 [20.71, 34.29]	14.17 [7.37, 21.46]	30.60 [22.38, 39.17]
OmniGen-v1	27.44 [24.11, 30.95]	29.85 [25.56, 34.43]	25.00 [18.57, 31.43]	26.36 [17.86, 35.37]	28.54 [20.51, 37.03]
BAGEL-7B	27.10 [23.51, 30.80]	27.38 [20.57, 34.55]	23.57 [17.14, 30.00]	24.44 [16.98, 32.54]	34.10 [25.87, 42.24]
Qwen-Image-Edit-2511	26.91 [23.17, 30.72]	30.65 [23.84, 37.77]	24.29 [17.86, 31.43]	29.62 [20.97, 38.60]	22.69 [15.14, 30.89]

D.6 Complete Task-Level Results

Table 12: Complete task-level SpatialGen-Bench results (%). Tasks are grouped by capability; aggregates appear in Tables 1 and 2.

Model	Perception				Understanding				Reasoning			Interaction		
	Cnt.	Dep.	Ori.	Size	Rel.	Pers.	Mental	Ground.	M-Hop	Pred.	Feas.	Aff.	Nav.	Traj.
Human	99.50	100.00	92.50	97.14	88.57	91.43	74.29	88.57	63.33	79.20	91.43	89.48	83.33	90.00
Text Answering														
GPT-5.4	80.00	76.67	67.50	74.29	42.86	31.43	45.71	57.14	50.00	46.01	71.43	42.43	83.33	83.33
Cosmos3-Nano	55.00	83.33	77.50	91.43	22.86	37.14	34.29	85.71	46.67	59.98	57.14	76.13	23.33	90.00
RynnBrain-8B	70.00	80.00	32.50	85.71	37.14	22.86	40.00	82.86	43.33	33.79	54.29	85.71	26.67	20.00
RoboBrain2.5-8B	55.00	80.00	10.00	88.57	25.71	20.00	11.43	80.00	33.33	45.39	62.86	80.08	13.33	56.67
SenseNova-Vision-7B	67.50	33.33	42.50	74.29	42.86	34.29	22.86	74.29	30.00	48.28	42.86	63.81	26.67	50.00
InternVL3.5-8B	32.50	70.00	10.00	94.29	20.00	20.00	28.57	77.14	56.67	74.39	54.29	40.95	26.67	40.00
SpaceR	60.00	36.67	67.50	80.00	57.14	20.00	22.86	25.71	43.33	54.48	60.00	18.81	26.67	66.67
SpaceOm	52.50	43.33	52.50	88.57	31.43	25.71	25.71	77.14	33.33	31.09	65.71	5.00	40.00	53.33
LLaVA-OV-1.5-8B	57.50	63.33	7.50	88.57	42.86	31.43	31.43	45.71	33.33	30.60	54.29	26.67	30.00	80.00
SpatialThinker-3B	47.50	36.67	55.00	91.43	37.14	20.00	42.86	65.71	36.67	26.69	54.29	5.77	43.33	60.00
Qwen3-VL-8B	55.00	66.67	37.50	80.00	22.86	25.71	22.86	77.14	33.33	37.35	68.57	40.95	16.67	36.67
GEM-2B	65.00	40.00	47.50	74.29	54.29	22.86	42.86	54.29	33.33	21.43	62.86	27.62	16.67	46.67
Cambrian-S-7B	82.50	46.67	60.00	82.86	28.57	22.86	28.57	5.71	16.67	57.69	65.71	20.00	30.00	46.67
GLM-4.6V-Flash	25.00	73.33	32.50	94.29	48.57	25.71	40.00	68.57	3.33	19.22	65.71	36.27	16.67	23.33
VLM-3R-7B	67.50	53.33	12.50	91.43	22.86	25.71	25.71	11.43	33.33	12.60	65.71	13.64	33.33	46.67
Spatial-MLLM	27.50	10.00	5.00	88.57	25.71	25.71	34.29	34.29	36.67	10.06	51.43	16.19	33.33	60.00
SpatialRGPT-8B	62.50	16.67	5.00	71.43	25.71	37.14	31.43	17.14	23.33	10.06	54.29	8.57	16.67	63.33
Janus-Pro-7B	67.50	10.00	5.00	74.29	37.14	17.14	42.86	8.57	13.33	27.22	65.71	5.71	16.67	50.00
Janus-1.3B	52.50	56.67	0.00	48.57	54.29	22.86	37.14	0.00	6.67	27.13	60.00	0.00	16.67	16.67
SpatialBot-3B	57.50	10.00	5.00	60.00	37.14	31.43	40.00	22.86	33.33	10.06	51.43	9.05	6.67	16.67
Visual Answering														
GPT Image 2	82.50	63.33	52.50	80.00	45.71	40.00	25.71	77.14	40.00	54.07	31.43	60.40	40.00	83.33
Nano Banana 2	82.50	86.67	62.50	60.00	45.71	25.71	31.43	65.71	16.67	20.95	37.14	42.50	40.00	73.33
GPT-5 Image Mini	52.50	73.33	77.50	54.29	42.86	28.57	28.57	34.29	10.00	30.99	40.00	14.05	36.67	50.00
Seedream 4.5	30.00	90.00	40.00	45.71	45.71	20.00	25.71	51.43	10.00	36.08	40.00	9.12	36.67	63.33
JoyAI-Image	55.00	53.33	5.00	54.29	40.00	17.14	25.71	57.14	10.00	14.20	34.29	34.51	33.33	70.00
FLUX.2 [klein] 9B	25.00	80.00	20.00	65.71	40.00	8.57	45.71	11.43	23.33	16.50	28.57	28.61	36.67	63.33
SenseNova-Vision-7B	22.50	83.33	12.50	11.43	48.57	8.57	37.14	88.57	33.33	4.00	0.00	51.13	20.00	30.00
FLUX.2 [klein] 4B	20.00	46.67	32.50	68.57	40.00	14.29	31.43	22.86	6.67	18.70	17.14	11.79	36.67	43.33
OmniGen-v1	7.50	93.33	10.00	8.57	48.57	5.71	42.86	2.86	33.33	17.16	28.57	5.63	40.00	40.00
BAGEL-7B	10.00	36.67	20.00	42.86	40.00	14.29	37.14	2.86	33.33	0.00	40.00	5.63	33.33	63.33
Qwen-Image-Edit-2511	10.00	43.33	15.00	54.29	31.43	14.29	34.29	17.14	23.33	28.40	37.14	4.74	33.33	30.00

E Additional Quantitative Results

E.1 Cross-Benchmark Evaluation Details

Table 3 reports success rates; Q-Spatial+ uses $\max(\hat{y}/y, y/\hat{y}) < 2$ for same-unit predictions $\hat{y}$ and targets y. GPT Image 2 parsing succeeds on 285/300 outputs (95.00%); per-benchmark rates appear in Table 7. The OmniSpatial pilot covers four task units, versus five in current construction.

E.2 Universal VLM Parser Sensitivity Details

Figure 7 reports five VLM evaluations of fixed outputs against a fixed ProVisE score reference; Table 14 gives the separate single-pass comparison. The reference reuses current parsed predictions, not five fresh end-to-end parsing runs, and uses the main evaluation's Trajectory extraction. Invalid answers receive zero. GPT Image 2's ProVisE score in Table 14 matches the complete main evaluation at 55.44.

In the archived Depth route, universal VLM judges view the source marker image and generated depth map and directly return A/B; ProVisE uses annotated coordinates and deterministic grayscale comparison. This exception differs from the generated-response-only fallback in the framework.

In the single-pass comparison, Qwen3-VL-8B (Bai et al., 2025a), Qwen2.5-VL-72B (Bai et al., 2025b), and Llama 4 Scout achieve 99.47/98.47/92.02% mean valid-parse rates among available images. Excluding Depth retains Spearman correlations of 0.49/0.37/0.89 and mean absolute score shifts of 6.30/5.67/6.18 points relative to ProVisE. Thus, evaluator-dependent scores and rankings are not explained solely by invalid formats or the Depth route.

E.3 Depth Ablation and Sensitivity Settings

Table 13 compares the prior mixed Depth route with deterministic parsing on 330 fixed outputs from 11 visual models. The latter compares local grayscale means at normalized A/B coordinates under a fixed brighter-is-closer convention. Accuracy varies by only 1.21 points across the tested windows; the default remains 5×5. Without coordinates, a VLM may locate objects in the source image but does not judge depth.

In Figure 8 (left), generation and valid-output rates pool the two models. Protocol variants pair Counting with JoyAI-Image, Relationship with Seedream 4.5, and Trajectory with GPT Image 2. Default scores match the main tables; variant means and standard deviations summarize five runs. Agreement pools per-sample correct/incorrect matches to the default over five runs. Parser sweeps vary Affordance mask thresholds over 96–160 and Trajectory rules and DFD thresholds over 0.35–0.45; the latter changes the scoring criterion. The repeated-generation and parser sweeps retain their archived fixed configurations; their Trajectory outputs were not rescored with the updated main extractor. Table 16 tests discretization with this extractor on the same fixed outputs.

Table 13: **Depth route and window ablation.** All rates use 330 fixed outputs from 11 visual models. Windows sample benchmark A/B coordinates; shading marks the default 5×5 window.

Metric (%)	**Prior**	1×1	3×3	5×5	7×7	9×9	11×11
Parse	78.48	98.79	99.39	99.39	99.39	99.39	99.39
Accuracy	55.45	68.79	68.18	68.18	68.18	69.09	69.39

Table 14: Task-level universal VLM scores on fixed generated outputs, with ProVisE main-table references; unavailable and invalid responses remain zero-scored.

Image model	Scorer	Overall	Rank	Parse	Cnt.	Dep.	Ori.	Size	Rel.	Pers.	Mental	Ground	M-Hop	Pred.	Feas.	Aff.	Nav.	Traj.
GPT Image 2	ProVisE	55.44	1	–	82.5	63.3	52.5	80.0	45.7	40.0	25.7	77.1	40.0	54.1	31.4	60.4	40.0	83.3
	Qwen3-VL-8B	47.81	1	99.4	70.0	53.3	40.0	80.0	42.9	28.6	40.0	68.6	30.0	56.3	45.7	30.7	16.7	66.7
	Qwen2.5-VL-72B	47.77	1	98.1	75.0	50.0	55.0	82.9	51.4	48.6	31.4	62.9	36.7	50.1	71.4	26.9	10.0	16.7
	Llama 4 Scout	42.69	1	92.3	67.5	80.0	52.5	74.3	40.0	28.6	37.1	22.9	33.3	58.1	37.1	12.9	23.3	30.0
Nano Banana 2	ProVisE	49.35	2	–	82.5	86.7	62.5	60.0	45.7	25.7	31.4	65.7	16.7	21.0	37.1	42.5	40.0	73.3
	Qwen3-VL-8B	43.40	3	98.5	67.5	56.7	50.0	77.1	37.1	25.7	40.0	60.0	30.0	19.2	51.4	22.9	23.3	46.7
	Qwen2.5-VL-72B	44.59	3	98.1	72.5	50.0	55.0	77.1	48.6	28.6	42.9	54.3	33.3	30.1	57.1	18.1	10.0	46.7
	Llama 4 Scout	40.60	3	91.5	67.5	76.7	50.0	60.0	40.0	42.9	37.1	37.1	16.7	39.3	37.1	13.9	26.7	23.3
GPT-5 Image Mini	ProVisE	40.97	3	–	52.5	73.3	77.5	54.3	42.9	28.6	28.6	34.3	10.0	31.0	40.0	14.1	36.7	50.0
	Qwen3-VL-8B	40.61	5	99.8	62.5	53.3	50.0	68.6	45.7	40.0	34.3	34.3	33.3	35.4	51.4	16.4	23.3	20.0
	Qwen2.5-VL-72B	39.62	4	98.7	55.0	63.3	52.5	62.9	48.6	28.6	31.4	34.3	36.7	34.8	68.6	14.8	10.0	13.3
	Llama 4 Scout	37.25	4	92.3	52.5	66.7	77.5	57.1	40.0	28.6	28.6	22.9	20.0	42.6	37.1	14.7	20.0	13.3
Seedream 4.5	ProVisE	38.84	4	–	30.0	90.0	40.0	45.7	45.7	20.0	25.7	51.4	10.0	36.1	40.0	9.1	36.7	63.3
	Qwen3-VL-8B	42.95	4	99.8	50.0	50.0	35.0	82.9	51.4	31.4	37.1	37.1	36.7	43.1	48.6	27.9	30.0	40.0
	Qwen2.5-VL-72B	37.53	6	98.3	45.0	40.0	32.5	74.3	51.4	40.0	22.9	37.1	26.7	32.9	57.1	15.6	10.0	40.0
	Llama 4 Scout	41.15	2	93.8	47.5	86.7	37.5	74.3	37.1	40.0	31.4	31.4	20.0	51.5	48.6	16.7	20.0	33.3
JoyAI-Image	ProVisE	36.00	5	–	55.0	53.3	5.0	54.3	40.0	17.1	25.7	57.1	10.0	14.2	34.3	34.5	33.3	70.0
	Qwen3-VL-8B	44.23	2	99.8	62.5	60.0	45.0	74.3	60.0	37.1	31.4	57.1	30.0	20.7	45.7	31.9	20.0	43.3
	Qwen2.5-VL-72B	46.92	2	98.9	55.0	63.3	60.0	77.1	54.3	31.4	42.9	37.1	36.7	31.0	62.9	35.2	23.3	46.7
	Llama 4 Scout	35.70	5	90.9	42.5	73.3	42.5	68.6	42.9	31.4	34.3	22.9	20.0	32.3	34.3	11.5	20.0	23.3
FLUX.2 [klein] 4B	ProVisE	29.33	6	–	20.0	46.7	32.5	68.6	40.0	14.3	31.4	22.9	6.7	18.7	17.1	11.8	36.7	43.3
	Qwen3-VL-8B	38.43	6	99.4	47.5	53.3	20.0	71.4	45.7	37.1	45.7	31.4	30.0	32.0	51.4	29.0	20.0	23.3
	Qwen2.5-VL-72B	39.17	5	98.5	57.5	50.0	37.5	65.7	40.0	28.6	37.1	40.0	26.7	31.5	68.6	18.6	20.0	26.7
	Llama 4 Scout	34.74	6	91.1	42.5	53.3	45.0	57.1	42.9	37.1	34.3	28.6	23.3	28.2	37.1	10.3	23.3	23.3

Cnt.: Counting; Dep.: Depth; Ori.: Orientation; Rel.: Relationship; Pers.: Perspective; Ground: Grounding; M-Hop: Multi-hop; Pred.: Prediction; Feas.: Feasibility; Aff.: Affordance; Nav.: Navigation; Traj.: Trajectory.

E.4 FIXED-OUTPUT SUPPORT-BUDGET CONTROLS

We isolate support sampling and path discretization on cached outputs from all 20 text and 11 visual systems, using all 35 Affordance and 30 Trajectory samples per system. No model is queried again; empty or previously invalid outputs remain zero-scored. These controls match downstream sample counts, not the information content or generation cost of the two answer interfaces.

Affordance support sampling. We first reproduce all 1,085 original scores from the saved point lists and generated masks. For each output, we then draw K support locations uniformly with replacement and apply the same point-in-mask test, repeating this process 2,000 times. Table 15 reports pool means and sampling SD. The estimator has expectation equal to the original support precision, so increasing K reduces sampling variance rather than introducing a systematic score advantage. The largest observed model-level mean shift is 0.27 points. This control does not equate selecting a few favorable points with describing an entire region.

Table 15: Affordance support sampling (%). Pool means $\pm$ sampling SD over 2,000 resamples (35 samples per system). Original uses the full saved support; K denotes draws per output.

Pool	**Original**	$K = 1$	$K = 5$	$K = 20$	$K = 50$
Text (20 systems)	31.17	31.17 ± 0.93	31.17 ± 0.44	31.17 ± 0.22	31.17 ± 0.14
Visual (11 systems)	24.37	24.34 ± 1.42	24.37 ± 0.64	24.38 ± 0.32	24.37 ± 0.20

Trajectory discretization. Text trajectories reuse saved normalized coordinates. Visual trajectories use the main evaluation's uniform red-path extractor: 20 points anchored at the task-provided start. Table 16 uses these same paths as Baseline. After freezing them, we resample prediction and reference to K equally spaced arc-length positions and apply the same DFD threshold of 0.4. Across $K = 5$–40, model-rank correlation with Baseline remains at least 0.9982, although individual scores can change. This stability is conditional on the fixed readout and does not imply invariance to the extraction procedure.

Table 16: Trajectory discretization (%). All 30 samples per system; K resamples both paths. Shifts (points) and Spearman ρ are relative to the main Baseline.

Pool / statistic	**Baseline**	$K = 5$	$K = 10$	$K = 20$	$K = 40$
Text (20 systems)	50.33	50.33	50.67	50.50	50.33
Visual (11 systems)	55.45	56.36	56.06	56.06	55.45
Max. model $\lvert\Delta\rvert$	–	3.33	6.67	3.33	0.00
Model-rank ρ	–	0.9982	0.9984	0.9987	1.0000

F QUALITATIVE INTERFACE COMPLEMENTARITY

Figures 10, 11, and 12 complement the paired outcome counts in Fig. 6 (right). Each pairs an incorrect response from the named text-output baseline with a displayed visual response that passes the unchanged task metric. The captions record the complete question, answer space, text response, and benchmark answer. Source inputs and generated responses are shown without cropping or aspect-ratio normalization.



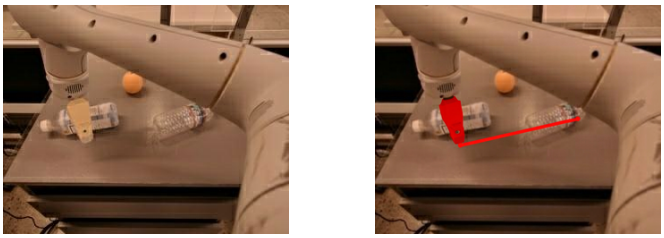
Source input GPT Image 2 response

Figure 10: **Depth. Question:** Which point is closer to the camera? **Options:** (A) A is closer; (B) B is closer. **GPT-5.4:** A. **Ground truth:** B. The generated depth field supports B at the benchmark coordinates; GPT Image 2 and all 11 visual systems pass this item.



Source input GPT-5 Image Mini response

Figure 11: **Relationship. Question:** Is the following statement true or false: The car is in front of the cat. **Options:** True; False. **GPT-5.4:** False. **Ground truth:** True. GPT-5 Image Mini marks the cat and foreground vehicle as distinct, parser-readable regions; 8 of 11 visual systems pass.



Source input JoyAI-Image response

Figure 12: **Trajectory. Question:** You are a Franka robot using the end effector control. The task is pick blue plastic bottle. **Answer space:** a five-point trajectory polyline in normalized image coordinates (no discrete options). **GPT-5.4:** $(0.266, 0.465) \rightarrow (0.330, 0.495) \rightarrow (0.395, 0.525) \rightarrow (0.455, 0.550) \rightarrow (0.510, 0.575)$. **Ground truth:** $(0.266, 0.465) \rightarrow (0.278, 0.465) \rightarrow (0.250, 0.422) \rightarrow (0.209, 0.348) \rightarrow (0.200, 0.320)$. The JoyAI-Image path passes the DFD criterion; 5 of 11 visual systems pass.

Additional task examples. Figures 13–16 cover Size, Grounding, Feasibility, and Orientation; the last shows the direction-grid protocol.



Source input GPT Image 2 response

Figure 13: **Size. Question:** Is the second giraffe from the right the shortest? **Options:** (A) no; (B) yes. **GPT-5.4:** A (no). **Ground truth:** B (yes). GPT Image 2 outlines the queried giraffe and gives an affirmative visual mark; 9 of 11 visual systems pass.



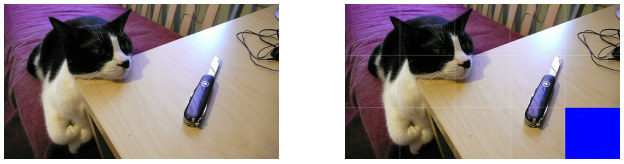
Source input GPT-5.4 text point (overlaid) GPT Image 2 response

Figure 14: **Spatial Grounding. Question:** Locate the object described as *a bus in front of others*. Return one normalized point inside the target object. **Answer space:** a normalized point (no discrete options). The ground-truth target is the RefCOCOg segmentation mask of the foreground bus. GPT-5.4's point $(0.677, 0.634)$, overlaid in magenta, falls just outside it. The parser reads $(0.603, 0.768)$ from GPT Image 2, placing the cyan marker inside the mask; 10 of 11 visual systems pass.



Source input GPT-5 Image Mini response

Figure 15: **Geometric Feasibility. Question:** Is there enough space for another person to sit in the ski lift? **Options:** (A) no; (B) yes. **RoboBrain2.5-8B:** A (no). **Ground truth:** B (yes). GPT-5 Image Mini adds a second passenger with clearance; 8 of 11 visual systems pass.



Source input Nano Banana 2 response

Figure 16: **Orientation. Question:** From the camera's perspective, which direction is the cat facing? **Options:** (A) front; (B) front-right; (C) right; (D) back-right; (E) back; (F) back-left; (G) left; (H) front-left. **GPT-5.4:** C (right). **Ground truth:** B (front-right). Nano Banana 2 fills the bottom-right grid cell, which the deterministic parser maps to front-right; 7 of 11 visual systems pass.