MINDVIEW: BENCHMARKING GENERATIVE MULTI-VIEW SPATIAL INTELLIGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial intelligence has recently been assessed through generative tasks such as spatially grounded image editing, yet language-guided generation of spatially consistent views from partial multi-view observations remains underexplored. We introduce **MindView**, a benchmark for generative multi-view spatial intelligence with 1,500 samples from 513 scenes across three indoor datasets. Its seven tasks form two families: geometry-dominant tasks specify changes in camera position and orientation, while semantics-dominant tasks require inferring target viewpoints by grounding referred objects and interpreting spatial relations and reference-frame cues. Evaluation of eleven generators reveals substantial room for improvement on MindView. A paired context analysis further shows that all eleven generators obtain lower overall scores under the expanded context setting. These findings motivate us to investigate whether task-specific supervision can improve target-view generation and whether models can use additional multi-view evidence effectively. We therefore fine-tune BAGEL on the MindView training dataset, improving its overall score by 24.3 points and yielding gains across all seven MindView tasks. Beyond these generation gains, the fine-tuned model also yields higher scores on three spatial-understanding benchmarks without training on their annotations, suggesting that learning to generate spatially grounded views may benefit spatial understanding. To improve evidence utilization under expanded context, we propose **MindCue**, a training-free spatial evidence selector guided by a target-view hypothesis. MindCue reverses the drop in overall score under expanded context and outperforms the standard-context baseline. Overall, MindView provides a concrete testbed for generative multi-view spatial intelligence. Our findings further highlight task learning and query-guided evidence selection as complementary strategies for improving performance in this setting.

1 INTRODUCTION

Benchmarks such as MindCube (Wang et al., 2025b), MMSI-Bench (Yang et al., 2026), and SAT-v2 (Ray et al., 2025) evaluate the spatial intelligence of multimodal understanding models through question answering. On the generation side, GenSpace (Wang et al., 2025c) and GSI-Bench (Zhu et al., 2026b) have begun this exploration by asking whether spatial instructions are realized in a generated or edited image. We focus on evaluating *generative multi-view spatial intelligence* through language-guided target-view generation from partial observations of the same scene (Figure 1). This task requires models to integrate evidence across views, infer the requested viewpoint, and predict what would be visible from it while remaining consistent with the observed scene.

Related work differs in how reference images are used and how target viewpoints are specified. Multi-reference benchmarks such as MICE-Bench (Luo et al., 2026) and MultiBanana (Oshima et al., 2026) emphasize preserving and combining reference content, which alone does not establish that a generated image depicts the requested viewpoint of the observed scene. Pose-conditioned methods such as ZeroNVS (Sargent et al., 2024) evaluate view synthesis with the target camera supplied, leaving language-conditioned viewpoint inference outside their task setting. More recently, Astra (Zhu et al., 2026a) uses language-controlled novel-view generation to support agentic spatial question answering and evaluates generated views against real targets for pose and content consistency. However, its generation evaluation focuses on explicit camera motions with a single dominant

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.

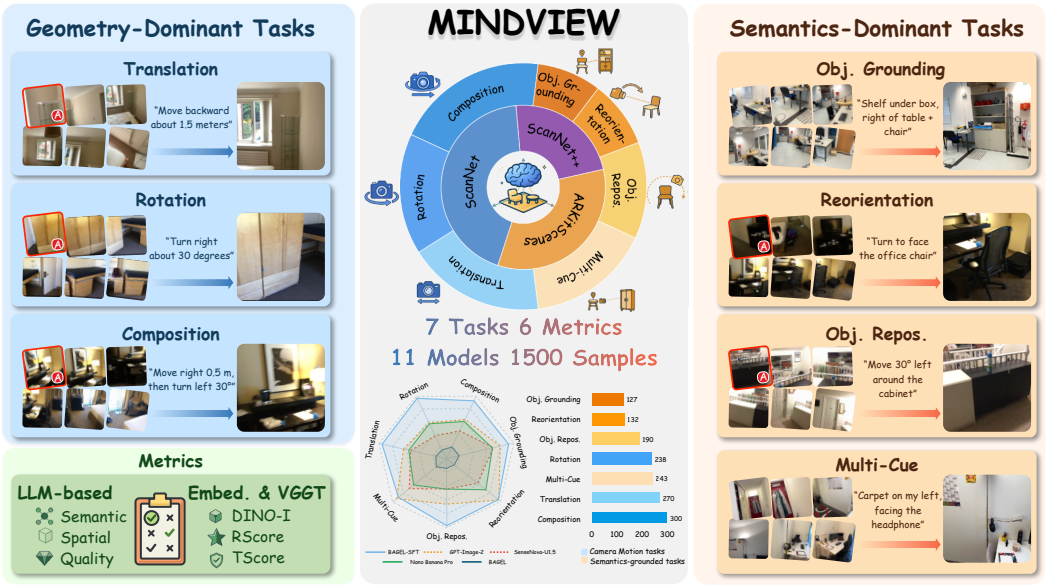


Figure 1: Overview of MindView, showing seven task examples, benchmark statistics, six metrics, and representative results. Red **A** marks anchor views, and targets are real images.

degree of freedom. We seek a unified evaluation of individual and composed camera motions and target viewpoints inferred from object references, spatial relations, and reference-frame cues.

We introduce **MindView**, a benchmark of **1,500 samples** from **513 scenes** in three indoor datasets (Figure 1). Its seven tasks form two families: *Geometry-Dominant* tasks specify camera translations, rotations, or their ordered composition, while *Semantics-Dominant* tasks require viewpoint inference from referred objects, spatial relations, and reference-frame cues. The model receives only RGB observations and a language query, with the target image and numerical camera parameters withheld. Six complementary metrics assess content fidelity, spatial agreement, rendering quality, and geometric consistency to identify content and spatial errors an aggregate score may obscure.

Across eleven off-the-shelf generators, preserving scene content while matching the requested viewpoint remains challenging (Table 1). Expanding context from six to ten views increases target-view coverage from 88.4% to 92.0%, yet lowers Overall for all eleven generators (Figure 5). These findings raise two questions: *can task-specific supervision improve target-view generation, and can query-guided evidence selection help models benefit from additional observations?*

To address these questions, we curate a target-view generation training set and fine-tune BAGEL (Deng et al., 2025), obtaining BAGEL-SFT with a 24.3-point Overall gain and improvements across all seven tasks. BAGEL-SFT also scores higher on three spatial-understanding benchmarks without training on their annotations, providing preliminary evidence of transfer from generation training to spatial understanding. To improve evidence utilization, we propose **MindCue**, a training-free spatial evidence selector using a target-view hypothesis to select supplementary observations from the expanded context. On both BAGEL and BAGEL-SFT, MindCue reverses the expanded-context degradation and outperforms the standard-context baseline and count-matched random selection.

Our main contributions are summarized as follows:

1. We introduce **MindView**, a benchmark for generative multi-view spatial intelligence, revealing two challenges across eleven generators: limited content fidelity and spatial consistency, and performance degradation with excessive context.

2. We curate target-view generation training data and show that task-specific supervision improves BAGEL across all seven tasks, with preliminary evidence of transfer to spatial understanding.

3. We propose **MindCue**, a training-free, query-guided evidence selector that improves both BAGEL and BAGEL-SFT over fixed and count-matched random baselines.

2 Related Work

Spatial Intelligence Benchmarks and Generative Reasoning. Spatial intelligence is evaluated through both question answering and image generation. MindCube (Wang et al., 2025b), MMSI-Bench (Yang et al., 2026), and SAT-v2 (Ray et al., 2025) assess spatial understanding through question answering, whereas GenSpace (Wang et al., 2025c) and GSI-Bench (Zhu et al., 2026b) evaluate whether generated images satisfy spatial instructions. The latter also reports understanding gains from generation-focused training. Thinking with Novel Views (Zhang et al., 2026) uses generated views for spatial question answering, while UniG2U-Bench (Wen et al., 2026) examines unified models' understanding and the effects of intermediate image generation. MindView evaluates generated views of observed scenes and tests whether target-view supervision aids spatial understanding.

Novel-View Synthesis and Viewpoint Control. View generation methods differ in target-view specification, from explicit camera parameters to language instructions. NeRF (Mildenhall et al., 2020), ZeroNVS (Sargent et al., 2024), EscherNet (Kong et al., 2024), and LVSM (Jin et al., 2025) use supplied camera controls. CapFrame (Li et al., 2026) grounds language-specified viewpoints within a reconstructed 3D Gaussian scene, while Object-Uni (Tan et al., 2026) uses explicit object poses for controllable generation. Astra (Zhu et al., 2026a) generates language-guided imagined views for spatial question answering and also evaluates them against real targets. Existing methods either rely on explicit geometry or consider constrained language-specified viewpoints, leaving viewpoint inference and scene-consistent rendering from partial multi-view evidence underexplored.

Multi-Reference Image Generation and Editing. Multi-reference generation and editing use multiple input images to preserve and combine visual content. IP-Adapter (Ye et al., 2023) and PhotoMaker (Li et al., 2023) provide image- or identity-based conditioning, while MultiRef (Chen et al., 2025) explicitly supports multiple visual references. MICE-Bench (Luo et al., 2026) evaluates multi-reference creation and editing, while MultiBanana (Oshima et al., 2026) tests generation from challenging combinations of visual references. Unlike these benchmarks, MindView treats inputs as partial observations of the same scene and evaluates language-guided target-view generation.

3 MindView Benchmark

In this section, we present MindView, a benchmark designed to evaluate generative multi-view spatial intelligence. Section 3.1 formulates the target-view generation problem and introduces seven tasks across two complementary families. Section 3.2 describes the data construction pipeline, and Section 3.3 presents the evaluation protocol and metrics.

3.1 Task Formulation and Task Design

MindView evaluates language-guided target-view generation from partial observations of the same scene. Let $C_K = (I_1, \ldots, I_K)$ denote an ordered list of RGB observations and let q denote a spatial query. A generator f predicts

$$\hat{I}_{\text{target}} = f(C_K, q). \tag{1}$$

The primary benchmark setting uses $K = 6$. At inference time, the generator receives only C_K and q. The held-out real target image, camera poses, depth, and scene annotations are not provided as generation inputs. When the query is defined relative to a starting viewpoint, I_1 serves as the anchor. The generated image must preserve the observed scene while matching the requested camera position and orientation.

The seven tasks share this interface and differ in how the target viewpoint is specified. **Geometry-Dominant Tasks** express it through egocentric camera motions from an anchor. Translation (Trans.) changes camera position through relative displacements while preserving heading. Rotation (Rot.) changes heading by a specified angle while keeping position fixed. Composition (Comp.) combines translation and rotation in the order specified by the query.

Semantics-Dominant Tasks specify the target through referred objects, spatial relations, and reference-frame cues. Relational Object Grounding (ROG) identifies the referred object through relations and generates a view satisfying distance, centering, and visibility constraints, without an anchor. Target-Facing Reorientation (TFR) requires locating the target object and turning to face it

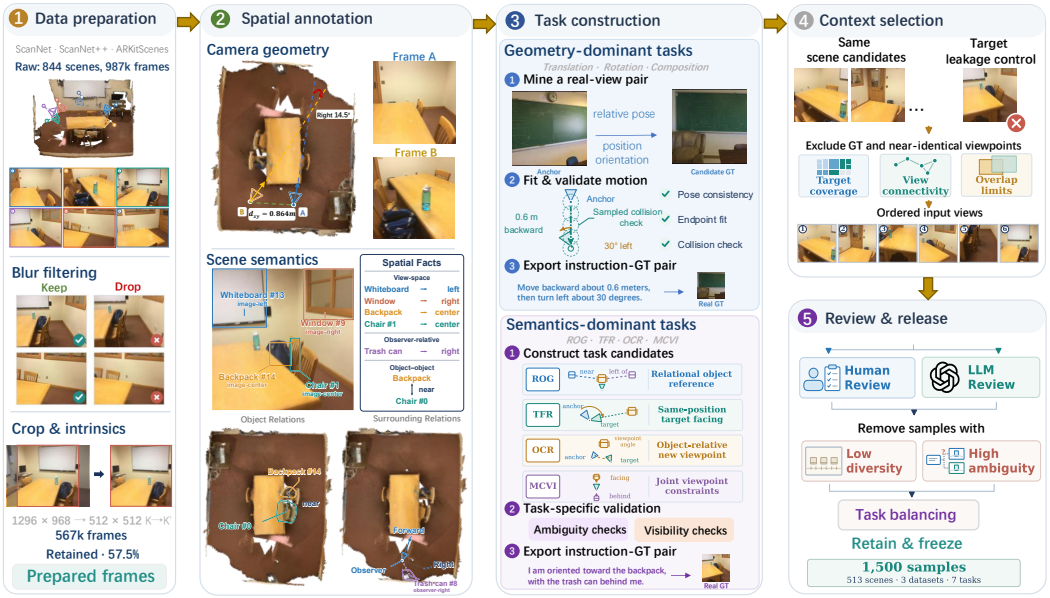


Figure 2: Five-stage offline construction pipeline for MindView.

while remaining at the anchor position. Object-Centric Repositioning (OCR) instead requires moving to a specified side or angular offset around the object and facing it, with directions defined relative to the anchor view. Multi-Cue Viewpoint Inference (MCVI) jointly infers position and heading from observer-relative scene cues, without an anchor. Together, the two families test complementary forms of target-view specification under a common generation interface.

Appendix A.4 provides real examples and full generation prompts for all seven tasks.

3.2 Data Construction

We construct MindView from ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), and ARKitScenes (Baruch et al., 2021). Figure 2 summarizes the five-stage offline construction pipeline.

Data preparation and spatial annotation. We first apply blur filtering while preserving temporal coverage along each recorded sequence. We then account for cropping and resizing in the camera model used for geometric checks. Using camera poses, depth, scene geometry, and instance annotations, we derive relative viewpoints, object relations, visibility, and object locations in each image. These annotations form the spatial basis for constructing queries and targets.

Task construction. For Geometry-Dominant Tasks, relative poses between recorded views yield motion instructions. Candidates are filtered by task-specific pose constraints, and endpoints reconstructed from rounded instructions are checked against recorded targets. Where point clouds are available, Translation and Composition also use discrete collision checks along instructed segments. For Semantics-Dominant Tasks, identities, relations, and reference-frame cues define query-target candidates. Visibility and viewpoint checks limit ambiguity across recorded views. Each query uses a real captured target and, when required, an anchor. Appendix A details the task-specific criteria.

Offline context selection. For each query-target pair, we select six unique same-scene views as a connected path. The target is excluded, any required anchor comes first, and other candidates pass temporal and pose-separation filters. Selection enforces pairwise-overlap limits and coverage of target-visible content and the broader scene. MCVI also requires every object cue constraining observer location to appear across the selected views.

Review and release. Candidate groups undergo human and LLM-assisted screening for residual ambiguity and limited within-sample visual diversity. Appendix A.3 provides the exact GPT-5.6 (OpenAI, 2026b) screening prompt. The frozen benchmark contains 1,500 samples from 513 scenes across the three datasets (Figure 3). We also provide 20,000 training examples from 3,406 scenes,

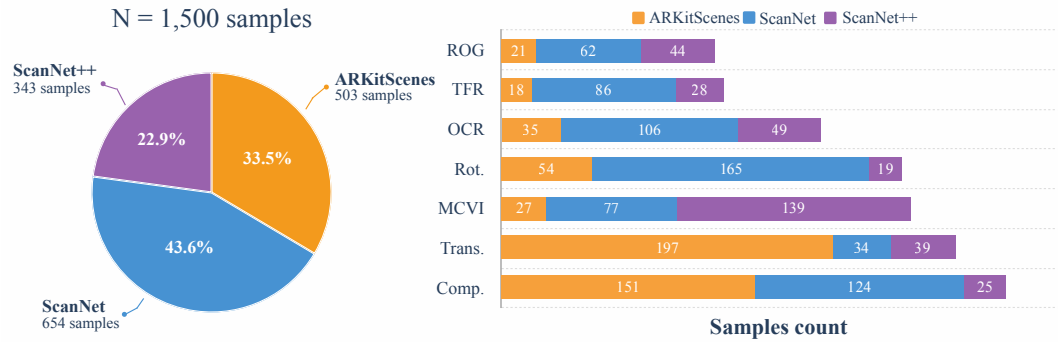


Figure 3: Distribution of the 1,500 MindView samples across source datasets and tasks.

using the same interface for all seven tasks. These examples use the source datasets' training splits and share no scenes with the benchmark (Appendix C.1).

3.3 Evaluation Protocol

Image-feature similarity. DINO-I probes overall visual-content fidelity through DINOv2-Large feature correspondence (Oquab et al., 2023). We use the cosine similarity between generated and real target features, scaled by 100 without clipping.

VLM-based assessment. Qwen3-VL-30B-A3B-Instruct (Qwen Team, 2025b) judges whether scene content (**Semantic**) and visible viewpoint and layout (**Spatial**) match the real target; **Quality** assesses internal coherence and rendering defects in the generation alone. The judge receives no query or context; integer 0–10 ratings are scaled to 0–100. See Appendices B.2 and B.6 for inputs and prompts.

Geometric consistency. We introduce a VGGT-based protocol (Wang et al., 2025a) for MindView to assess target-viewpoint fidelity through estimated camera position and orientation. Joint inference on fixed references and generated (gen) and real (real) target images yields rotations $\hat{R}$ and camera centers $\hat{\mathbf{c}}$ in a shared frame. For sample i, the pose errors are

$$\left(e_{\mathrm{rot},i}, e_{\mathrm{trans},i}\right) = \left(d_{\mathrm{SO(3)}}\left(\hat{R}_{\mathrm{gen}}, \hat{R}_{\mathrm{real}}\right),\ a_i \left\|\hat{\mathbf{c}}_{\mathrm{gen}} - \hat{\mathbf{c}}_{\mathrm{real}}\right\|_2\right). \tag{2}$$

Here, $d_{\mathrm{SO(3)}}$ is angular distance in degrees, and $a_i > 0$ converts VGGT lengths to meters using real-image sensor depths, without ground-truth poses. Per-sample RScore and TScore decrease linearly from 100 at zero error to 0 at 60° and 2 m, respectively, and remain zero beyond these thresholds. Appendix B.3 details the reference inputs, calibration, and scoring.

All scores are higher-is-better. We average each component within tasks, then equally across the seven tasks; Overall is the unweighted mean of the six task-macro scores.

Evaluator reliability. We assess evaluator reliability through comparisons with human ratings and reconstruction stability tests. Three multimodal experts independently rate outputs from four generators. Aggregated human and VLM scores yield identical model rankings for Semantic, Spatial, and Quality. For VGGT, we vary the generated image while keeping all real views fixed. Across samples, the median changes in the reconstructed target pose are 0.71° in relative rotation and 2.75 cm in reference-aligned position. Together, these experiments support the reliability of our VLM-based and geometric metrics for comparing generators (Appendix B.4).

4 Experiments

We first evaluate current generators on MindView and analyze their task-level performance. We then examine whether additional observations improve target-view generation. Motivated by these findings, we investigate task-specific adaptation through supervised fine-tuning and training-free, query-guided evidence selection through MindCue.

Table 1: **Quantitative comparison on MindView.** Task-macro scores over 1,500 samples. and mark best and second-best across displayed models, respectively; † denotes tiled context views.

Model	DINO-I	VLM Judge			Geometry		Overall
		Semantic	Spatial	Quality	RScore	TScore	
Open-Source Models							
BAGEL	48.3	44.5	33.6	47.1	34.7	31.9	40.0
FLUX.2	52.2	49.5	39.4	80.4	35.5	38.6	49.3
Qwen-Image-Edit-2511†	52.4	49.2	38.6	69.5	37.2	40.3	47.9
Emu3.5†	44.9	41.0	31.2	67.2	36.0	49.9	45.0
DreamOmni2	41.7	36.0	28.8	65.5	25.2	42.7	40.0
SenseNova-U1	53.6	48.2	38.3	63.5	25.6	33.2	43.7
SenseNova-U1.5	58.0	52.8	42.6	69.7	41.7	43.2	51.3
Closed-Source Models							
Nano Banana	56.5	54.0	41.7	66.2	39.0	40.7	49.7
Nano Banana Pro	60.0	57.2	45.4	80.0	38.8	39.0	53.4
GPT-Image-1.5	58.8	58.9	46.3	82.2	39.9	33.5	53.3
GPT-Image-2	63.0	61.5	49.0	**84.6**	49.9	39.1	57.9
Task-Specific Adaptation							
BAGEL-SFT	**68.0**	**71.0**	**57.1**	61.8	**61.7**	**66.5**	**64.4**



Figure 4: Qualitative comparison on Composition and Multi-Cue Viewpoint Inference.

4.1 Experimental Setup

Evaluation set and models. All MindView generation comparisons use the frozen release of 1,500 aligned scene–query–target samples. The standard benchmark input consists of six context images and a language query per sample. All MindView generation results follow the evaluation protocol in Section 3.3. We compare eleven off-the-shelf generators: seven open-source models (BAGEL (Deng et al., 2025), FLUX.2 (Black Forest Labs, 2026), Qwen-Image-Edit-2511 (Wu et al., 2025; Qwen Team, 2025a), Emu3.5 (Cui et al., 2025), DreamOmni2 (Xia et al., 2025), SenseNova-U1 (Diao et al., 2026b), and SenseNova-U1.5 (Diao et al., 2026a)) and four closed-source models (Nano Banana (Fortin et al., 2025), Nano Banana Pro (Raisinghani, 2025), GPT-Image-1.5 (OpenAI, 2025), and GPT-Image-2 (OpenAI, 2026a)). We additionally evaluate BAGEL-SFT, obtained by supervised fine-tuning of BAGEL on our curated training dataset for 10,000 steps.

4.2 Benchmark Results

Quantitative results. Among the eleven off-the-shelf generators, GPT-Image-2 achieves the highest Overall score (57.9), while SenseNova-U1.5 leads the open-source group (51.3; Table 1). GPT-Image-2 also leads on all seven tasks, but task-level weaknesses recur across models: Composition is the lowest-scoring task for eight generators, and Multi-Cue Viewpoint Inference is lowest for the remaining three (Table 2). These results highlight recurring difficulties with composed camera motions and viewpoints constrained by multiple scene cues.

Content fidelity and geometric consistency also yield different model rankings. Nano Banana Pro scores higher than SenseNova-U1.5 on DINO-I and Semantic, yet lower on RScore (38.8 versus 41.7) and TScore (39.0 versus 43.2). This discrepancy underscores the value of evaluating target-view geometry alongside content fidelity when assessing language-guided view generation.

Qualitative results. Consistent with the quantitative findings, Figure 4 illustrates viewpoint and content errors across models on two challenging tasks, Composition and Multi-Cue Viewpoint Inference. In Composition, Nano Banana Pro emphasizes the painting but differs from the target in

Table 2: **Task-wise comparison on MindView.** Per-task Overall scores across all displayed models. Highlighting and † follow Table 1; abbreviations are defined in Section 3.1.

Model	Geometry-Dominant Tasks			Semantics-Dominant Tasks			
	Trans.	Rot.	Comp.	ROG	TFR	OCR	MCVI
Open-Source Models							
BAGEL	41.7	38.8	36.1	43.2	38.7	44.1	37.6
FLUX.2	53.0	47.0	44.1	50.5	49.9	53.9	46.6
Qwen-Image-Edit-2511†	53.4	45.9	41.6	49.9	46.6	57.0	40.7
Emu3.5†	55.7	43.6	40.7	42.5	45.0	52.3	35.5
DreamOmni2	54.4	40.9	35.8	33.7	35.0	47.9	32.2
SenseNova-U1	44.2	40.1	37.9	49.4	46.8	44.6	43.3
SenseNova-U1.5	53.6	46.8	43.2	56.7	53.4	54.7	50.9
Closed-Source Models							
Nano Banana	49.3	48.1	45.8	53.4	55.7	48.3	47.3
Nano Banana Pro	54.5	53.7	47.3	56.3	58.6	54.7	48.7
GPT-Image-1.5	52.1	47.9	46.3	59.2	59.4	54.3	53.7
GPT-Image-2	57.0	54.2	48.2	59.4	**66.8**	62.4	**57.0**
Task-Specific Adaptation							
BAGEL-SFT	**67.1**	**68.2**	**56.3**	**62.7**	66.5	**73.4**	56.2

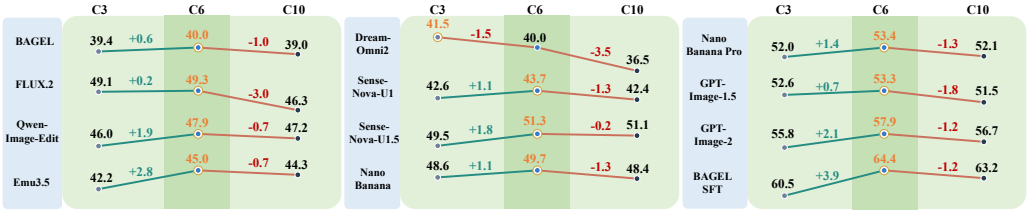


Figure 5: Overall score trajectories across nested context settings.

both viewing angle and the painting's content. In Multi-Cue Viewpoint Inference, GPT-Image-2 and FLUX.2 include the desk, chair, and heater, but depict the chair as smaller and farther back in the image than in the target view. Appendix Figure 15 provides comparisons across all seven tasks.

4.3 Context Scaling and Evidence Utilization

We examine context expansion on the same 1,500 scene–query–target samples using nested three-, six-, and ten-view settings, with C_6 as the standard input. C_3 is an ordered subsequence of C_6, which is an ordered subsequence of C_{10}. Target images are excluded from generator inputs, and VGGT uses the same ten evaluation references across settings. Computed only for offline analysis, target coverage measures the fraction of content visible in the real target that is also observed in at least one context view. Its sample mean rises from 57.6% at C_3 to 88.4% at C_6 and 92.0% at C_{10}. Despite this increase, all eleven off-the-shelf generators and BAGEL-SFT obtain lower Overall at C_{10} than at C_6 (Figure 5). Greater coverage therefore does not ensure better generation under this protocol. This observation motivates MindCue, which selects query-relevant evidence from the four C_{10} additions while retaining the complete C_6 core (Section 4.5). Full component scores appear in Appendix Table 13.

4.4 Supervised Adaptation and Transfer to Spatial Understanding

Supervised adaptation. To examine whether task-specific supervision improves target-view generation, we fine-tune BAGEL (Deng et al., 2025) on our curated training dataset for 10,000 steps. Adaptation raises Overall from 40.0 to 64.4, a gain of 24.3 points, as shown in Table 1, and all six metrics improve on every task (Appendix Tables 9–12). The adapted model's task-macro advantage over GPT-Image-2 is larger on geometry-dominant than semantics-dominant tasks (10.8 versus 3.3 points; Table 2), possibly because explicit motion supervision provides more regular patterns to learn, whereas semantics-dominant tasks demand more complex semantic interpretation and spatial reasoning. In the Composition example in Figure 4, BAGEL-SFT more closely matches the GT's painting-dominated close-up than GPT-Image-2, which instead produces a broad room view. Training details are provided in Appendix C.

Table 3: **Spatial-understanding transfer.** We report accuracy (%) and gains in percentage points.

Benchmark	Protocol	Eval. Size	Accuracy (%)		Difference (pp)
			BAGEL	BAGEL-SFT	
SAT-v2	Micro accuracy	300	58.0	62.7	**+4.7**
	Both-order accuracy	150	27.3	41.3	**+14.0**
MMSI-Bench	Vanilla accuracy	1,000	31.0	32.2	**+1.2**
MindCube-Tiny	Raw-QA accuracy	1,050	34.7	36.5	**+1.8**

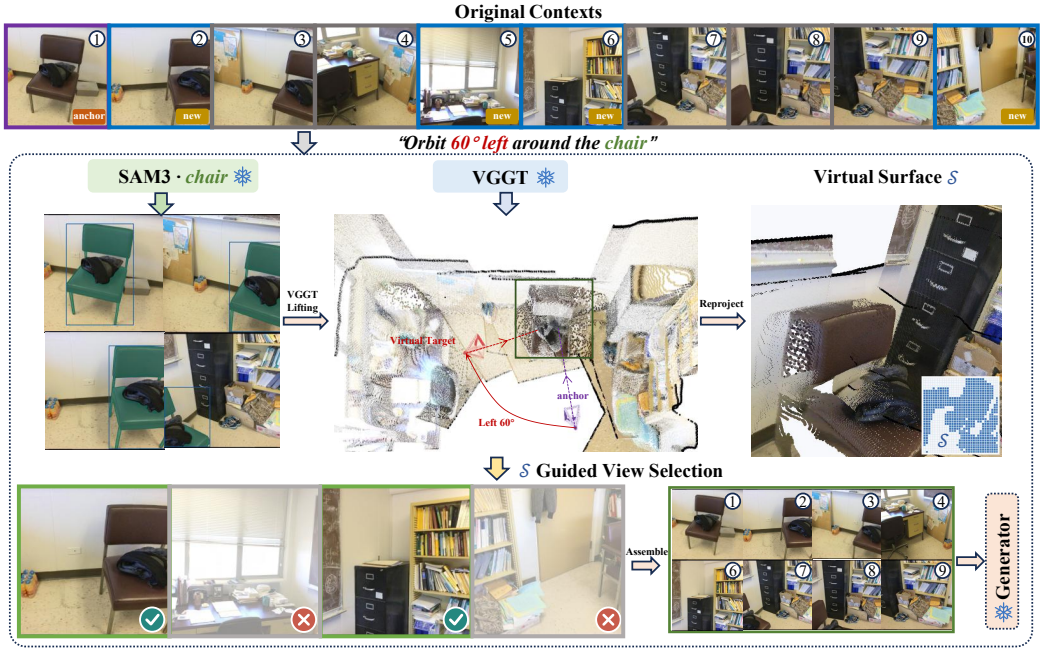


Figure 6: Overview of the MindCue framework.

Transfer to spatial understanding. We next examine whether learning to generate target views also benefits spatial understanding. To this end, we evaluate BAGEL and BAGEL-SFT on SAT-v2 (Ray et al., 2025), MMSI-Bench (Yang et al., 2026), and MindCube-Tiny (Wang et al., 2025b), strictly following their respective input-construction and answer-scoring protocols (Appendix D). Our fine-tuning stage uses no annotations from these benchmarks. All results are obtained in our own evaluations. Our reproduced BAGEL accuracies on MMSI-Bench and MindCube-Tiny match those reported by SenseNova-SI (Cai et al., 2026) at the reported precision.

As shown in Table 3, BAGEL-SFT obtains higher scores in all three comparisons. On SAT-v2, micro accuracy increases from 58.0% to 62.7%, while both-order accuracy rises from 27.3% to 41.3%, indicating that more questions are answered correctly under both evaluated answer orders. Its scores on MMSI-Bench and MindCube-Tiny also exceed the reproduced BAGEL baselines by 1.2 and 1.8 percentage points, respectively. Together, these results provide preliminary evidence that learning to generate spatially grounded views may also benefit spatial understanding, even without direct supervision from the evaluated benchmarks.

4.5 MindCue: A Training-Free Spatial Evidence Selector

Motivated by the context-scaling results, **MindCue** uses a target-view hypothesis to select additional spatial evidence without updating model weights (Figure 6).

Method. For sample i, let C_i^{pool} be the ordered observations, q_i the query, and $\mathcal{F}_i, \mathcal{D}_i$ the core and additional-view index sets. Frozen VGGT (Wang et al., 2025a) provides confidence-filtered scene points $\hat{\mathcal{P}}_i$; for semantics-dominant tasks, frozen SAM 3 (Carion et al., 2026) segments queried entities. A task-specific solver uses this evidence, the query, and core/reference-view metadata to

Table 4: **Task-macro Overall under context selection.** Parentheses give mean view counts; and mark the best and second-best policies per model, respectively. ΔC_k is MindCue's score-point gain over Fixed C_k, computed before rounding.

Model	Fixed C_6 (6.0)	Fixed C_{10} (10.0)	CounterSel (8.6)	RandDrop (7.4)	MindCue (7.4)	ΔC_6	ΔC_{10}
BAGEL	40.0	39.0	37.1	39.6	**41.1**	+1.1	+2.1
BAGEL-SFT	64.4	63.2	62.6	64.1	**65.1**	+0.8	+1.9

hypothesize a target camera $\hat{\mathcal{C}}_i^\star$ (pose and intrinsics). Its virtual surface is

$$\mathcal{S}_i = \mathrm{Occ}\Big[\mathrm{ZBuf}\Big(\Pi_{\hat{\mathcal{C}}_i^\star}\Big(\hat{\mathcal{P}}_i\Big)\Big)\Big]. \tag{3}$$

Here Π projects points, ZBuf retains the nearest valid depth per grid cell, and Occ returns the occupied cells.

For nonempty $\mathcal{S}_i$, let $R_i(v)$ measure additional view v's reliable coverage, capping credit for geometry supported only by that view, and let τ be the coverage threshold. $X_i \subseteq \mathcal{D}_i$ contains at most one below-threshold candidate passing connectivity and nonredundancy tests; without an explicit anchor, it is empty. The retained additional-view indices are

$$A_i = \begin{cases} \varnothing, & \mathcal{S}_i = \varnothing, \\ \{v \in \mathcal{D}_i : R_i(v) \geq \tau\} \cup X_i, & \mathcal{S}_i \neq \varnothing. \end{cases} \tag{4}$$

The generator receives the original query and pool views indexed by $\mathcal{F}_i \cup A_i$, preserving their order and the complete core. Selection uses no held-out targets, ground-truth geometry, generated outputs, or evaluation scores. Appendix E specifies the support and selection rules; Table 7 details the task-specific cameras.

Evaluation. We compare vanilla and fine-tuned BAGEL on the same 1,500 samples with fixed checkpoints and no-thinking inference. Fixed C_6 and C_{10} supply the core and full pool, respectively. Both controls preserve the core: RandDrop randomly selects the same number of additions as MindCue per sample, while CounterSel retains rejected additions and is not count-matched. Table 4 reports Overall and view counts; Appendix Table 14 gives all six components.

Results. MindCue exceeds both fixed-context baselines on BAGEL and BAGEL-SFT, reversing the expanded-context decline and improving on the standard input. Its gains over count-matched RandDrop are 1.6 and 1.1 points for vanilla and fine-tuned BAGEL, respectively, supporting selection beyond count reduction. The improvement over RandDrop extends to five of the six component scores for BAGEL and all six for BAGEL-SFT. CounterSel scores lowest among all five policies for both models, consistent with the value of target-guided retention.

Runtime. On a single A100 80GB, MindCue adds 1.44 s on average, or 3.4% of BAGEL inference time with identical selected views and input order. The task-averaged costs of VGGT reconstruction and SAM 3 segmentation with mask lifting are 0.57 s and 0.82 s, respectively. SAM 3 is used only for semantics-dominant tasks. The remaining selection steps add 0.05 s. We sum separately timed stages with preloaded models, excluding preprocessing (Appendix E.6). Thus, query-guided selection adds only modest computational overhead relative to generation.

5 Conclusion

We introduce MindView, a benchmark for generative multi-view spatial intelligence through language-guided target-view generation from partial multi-view observations. Evaluation of eleven generators reveals challenges in content fidelity and viewpoint consistency, while expanding from six to ten views lowers Overall for all eleven. Fine-tuning BAGEL on our training set improves generation across all seven tasks and provides preliminary evidence of transfer to spatial understanding. Complementing task-specific adaptation, MindCue uses a target-view hypothesis for training-free evidence selection and improves both BAGEL variants over standard and expanded inputs and count-matched random selection. These findings highlight task learning and query-guided evidence selection as complementary directions for advancing generative multi-view spatial intelligence.

AI USE STATEMENT

We used AI tools to polish the language and improve the presentation of the manuscript. GPT-5.6 screened and filtered candidate benchmark samples for target ambiguity and insufficient context-view diversity before expert review (Appendix A.3). Our use of Qwen3-VL for model-based evaluation is described in Appendix B.2. The authors take responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

We will release the data, code, and detailed documentation needed for faithful reproduction of our main experimental findings. Appendix A details benchmark construction and review, including the exact screening prompt. Appendix B specifies the evaluation metrics, aggregation rules, and evaluator prompts. Training and inference settings, spatial-understanding evaluation protocols, and MindCue implementation are provided in Appendices C, D, and E, respectively.

REFERENCES

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Black Forest Labs. FLUX.2 [klein]: Towards interactive visual intelligence. https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence, January 2026. Evaluated checkpoint: https://huggingface.co/black-forest-labs/FLUX.2-klein-9B. Accessed 2026-09-19.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, Qingping Sun, Hui En Pang, Jiaqi Li, Oscar Qian, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7879–7890, June 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/e0982cbc81401df3430ee1ff780dc7a2-Paper-Conference.pdf.

Ruoxi Chen, Dongping Chen, Siyuan Wu, Sinan Wang, Shiyun Lang, Petr Sushko, Gaoyang Jiang, Yao Wan, and Ranjay Krishna. Multiref: Controllable image generation with multiple visual references. *arXiv preprint arXiv:2508.06905*, 2025.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, et al. SenseNova-U1.5: Towards native unified visual intelligence. *arXiv preprint arXiv:2609.11929*, 2026a.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. SenseNova-U1: Unifying multimodal understanding and generation with NEO-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026b.

Alisa Fortin, Guillaume Vernade, Kat Kampf, and Ammaar Reshi. Introducing Gemini 2.5 Flash Image, our state-of-the-art image model. Google Developers Blog, August 2025. URL https://developers.googleblog.com/introducing-gemini-2-5-flash-image/. Official announcement of Nano Banana. Published August 26, 2025. Accessed 2026-09-19.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242*, 2025.

Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J. Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jirong Li, Satoshi Ikehata, Shuhei Kurita, and Ikuro Sato. CapFrame: Text-Instructed Viewpoint Grounding in 3D Gaussian Scenes via Geometric Pseudo Labels. *arXiv preprint arXiv:2608.30342*, 2026.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Siqi Luo, Huayu Zheng, Jianghan Shen, Yi Xin, Luxin Xu, Jiyao Liu, Xinyu Zhang, Hang Zhou, Pengyu Xie, Xiaohui Li, Shuo Cao, Yuandong Pu, Junjun He, Bin Fu, Yihao Liu, Yu Qiao, Guangtao Zhai, Yuewen Cao, and Xiaohong Liu. MICE-Bench: A challenging and comprehensive benchmark for multi-reference image creation and editing. In *International Conference on Machine Learning*, 2026.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 2020.

OpenAI. GPT-Image-1.5. https://developers.openai.com/api/docs/models/gpt-image-1.5, 2025. Official model documentation. Accessed 2026-09-19.

OpenAI. GPT-Image-2. https://developers.openai.com/api/docs/models/gpt-image-2, 2026a. Official model documentation. Accessed 2026-09-19.

OpenAI. GPT-5.6 Sol. https://developers.openai.com/api/docs/models/gpt-5.6-sol, 2026b. Official model documentation. Accessed 2026-09-24.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Yuta Oshima, Daiki Miyake, Kohsei Matsutani, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. MultiBanana: A challenging benchmark for multi-reference text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Qwen Team. Qwen-Image-Edit-2511 model card. https://huggingface.co/Qwen/Qwen-Image-Edit-2511, 2025a. Accessed 2026-09-19.

Qwen Team. Qwen3-VL-30B-A3B-Instruct model card. https://huggingface.co/Qwen/Qwen3-VL-30B-A3B-Instruct, 2025b. Accessed 2026-09-16.

Naina Raisinghani. Introducing Nano Banana Pro. Google, The Keyword, November 2025. URL https://blog.google/innovation-and-ai/products/nano-banana-pro/. Official announcement of Gemini 3 Pro Image. Published November 20, 2025. Accessed 2026-09-19.

Arijit Ray, Jiafei Duan, Ellis L. Brown, II, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Dynamic spatial aptitude training for multimodal language models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=DW8U8ZWa1U.

A. Santos, C. Ortiz de Solórzano, J. J. Vaquero, J. M. Peña, N. Malpica, and F. del Pozo. Evaluation of autofocus functions in molecular cytogenetic analysis. *Journal of Microscopy*, 188(3):264–272, 1997. doi: 10.1046/j.1365-2818.1997.2630819.x.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. Zeronvs: Zero-shot 360-degree view synthesis from a single image. *arXiv preprint arXiv:2310.17994*, 2024.

Mining Tan, Yinuo Wang, Ziqi Zhou, Weize Quan, Sifei Li, Jingdong Chen, DanDan Zheng, Libin Wang, and Weiming Dong. Object-Uni: A Unified Model for Object-Centric Spatial Understanding and Controllable Generation. *arXiv preprint arXiv:2608.22757*, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025a.

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. *arXiv preprint arXiv:2506.21458*, 2025b.

Zehan Wang, Jiayang Xu, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Genspace: Benchmarking spatially-aware image generation. *arXiv preprint arXiv:2505.24870*, 2025c.

Zimo Wen, Boxiu Li, Wanbo Zhang, Junxiang Lei, Xiaoyu Chen, Yijia Fan, Qi Zhang, Yujiang Wang, Lili Qiu, Bo Li, Ziwei Liu, Caihua Shan, Yifan Yang, and Yifei Shen. UniG2U-Bench: Do unified models advance multimodal understanding? *arXiv preprint arXiv:2603.03241*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. DreamOmni2: Multimodal instruction-based editing and generation. *arXiv preprint arXiv:2510.06679*, 2025.

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 *International Conference on Learning Representations*, 2026.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. *arXiv preprint arXiv:2308.11417*, 2023.

Yanbing Zhang, Bo Wang, Jianhui Liu, Nan Jiang, Jiaxiu Jiang, Haoze Sun, Yijun Yang, Shenghe Zheng, Lin Song, Haoyang Huang, Nan Duan, and Wenbo Li. Thinking with Novel Views: A Systematic Analysis of Generative-Augmented Spatial Intelligence. *arXiv preprint arXiv:2605.10588*, 2026.

Chenming Zhu, Jingli Lin, Yilin Long, Peizhou Cao, Tai Wang, Jiangmiao Pang, and Xihui Liu. Thinking with imagination: Agentic visual spatial reasoning with world simulators. *arXiv preprint arXiv:2606.06476*, 2026a.

Muzhi Zhu, Shunyao Jiang, Huanyi Zheng, Zekai Luo, Hao Zhong, Anzhou Li, Kaijun Wang, Jintao Rong, Yang Liu, Hao Chen, Tao Lin, and Chunhua Shen. Exploring spatial intelligence from a generative perspective. *arXiv preprint arXiv:2604.20570*, 2026b.

A BENCHMARK CONSTRUCTION DETAILS

This section expands the five offline stages in Section 3.2. Task-specific constructors may interleave candidate generation, query construction, and validation. The generator receives none of these construction signals and sees only selected RGB views and the query.

A.1 DATA PREPARATION AND SPATIAL ANNOTATION

Data preparation. We apply the same blur-filtering procedure to ScanNet, ScanNet++, and ARKitScenes. We convert each frame to grayscale and resize its long side to 320 pixels. Its Tenengrad sharpness score is the mean squared gradient magnitude computed with 3×3 Sobel filters (Santos et al., 1997). Within each sequence, we retain frames at or above the median sharpness score and add the highest-scoring frame from each non-overlapping five-frame window. This union favors image clarity while preserving temporal coverage. For geometric checks, we use a centered square region of each retained image and adjust its camera intrinsics for resizing and cropping. Projection and visibility checks use these adjusted intrinsics and the corresponding image coordinates.

Spatial annotation. Recorded camera poses provide relative translations and heading changes. Horizontal displacements are expressed in a ground-plane frame aligned with the relevant camera heading. Instance labels and 3D geometry provide object identities, locations, extents, and spatial relationships, including proximity and vertical arrangement. We also derive object directions and distances relative to the observer. For each view, we project instance points using its pose and adjusted intrinsics, then compare their depths with recorded depth to estimate visibility within the effective image crop. The visible points support image-space locations and view-dependent object relationships. These annotations support query construction and target-view selection.

A.2 TASK-SPECIFIC CONSTRUCTION

Using these annotations, we construct each task around the information that specifies its target viewpoint. Geometry-Dominant Tasks convert the relative pose between recorded views into a motion query, whereas Semantics-Dominant Tasks use objects and spatial cues to identify valid target views. Every query has a real recorded target and uses an anchor only when it is defined relative to a starting view. Task-specific checks enforce geometric fidelity and, for semantic queries, reduce ambiguity across recorded views.

Geometry-Dominant Tasks. Translation encodes one or two ordered horizontal steps in the anchor camera frame without a commanded turn. Rotation changes the horizontal heading with only small camera-center drift. Composition supports either motion order. A displacement uses the anchor frame before a turn and the rotated camera frame after it. We retain pairs that satisfy task-specific position and heading constraints, together with height, pitch, and roll gates. Movement distances are rounded and turn angles are quantized for the query. We execute the discrete instruction geometrically and require the resulting camera center to remain close to the recorded target position. For Translation and Composition, available point clouds support collision checks with vertical body cylinders sampled along each translated segment. We reject a candidate if any sampled cylinder reaches the occupancy threshold, using these discrete checks to assess the traversability of the instructed path.

Relational Object Grounding. Without an anchor, the query uses one to three clear relational facts to identify an object. Its target is a real view in which that object is near the image center and all referenced objects are sufficiently visible. We remove views with a clear same-category distractor and reject a description that also matches another valid target candidate. Views supporting the same description are grouped when their camera poses and projected object layouts are equivalent. We select among these clusters by the horizontal clearance from the camera to the target object's footprint and reject cases with a near-tied alternative. ScanNet++ and ARKitScenes allow a single-relation fallback only if no multi-relation candidate exists and the scene records match the description to exactly one target instance and one compatible view cluster. This fallback occurs in 44 of 127 released ROG queries. The final instruction also specifies target distance and centering and requires all named objects to remain visible.

Target-Facing Reorientation. The task asks for an approximately in-place turn from Image 1 toward a uniquely described object. The object phrase must identify one scene instance, using either its category alone or that category with at most one nearby reference. We retain a real target view only when its camera faces the object and the object is sufficiently visible, fully inside the effective crop, and near the image center. A target is rejected when a similarly suitable view from approximately the same position points in a substantially different direction. We then select an anchor with a nearby camera center, similar height, a different heading, and consistent roll. To ensure that the turn reveals the requested object, we exclude anchors in which it passes a lenient depth-aware visibility test.

Object-Centric Repositioning. The task begins with an anchor that clearly shows the reference object near the image center, without another salient instance of the same category. We then select a target at the requested left or right angular offset, or on the side of the object opposite the anchor, with left and right defined in the anchor camera frame. The target must face and center the object from a similar distance while remaining consistent with the anchor in height, pitch, and roll. We discard queries when comparably suitable candidates imply distinct target poses. These checks validate the destination view, not a collision-free path around the object.

Multi-Cue Viewpoint Inference. Without an anchor, the query specifies the desired viewpoint through the observer's position relative to nearby objects and the image locations of visible objects. We match the complete cue set against every valid annotated frame in the scene and retain it only when all matches form one compact cluster in horizontal position and heading. A representative real frame from this cluster serves as the target. Within each scene, we remove repeated cue signatures and near-duplicate target poses. Because the search covers only recorded frames, it tests uniqueness among annotated observations rather than over continuous camera space.

Together, these procedures yield validated query-target pairs whose contexts are selected in the next stage.

A.3 Context Selection, Review, and Release

Offline context selection. The selector assembles six unique same-scene frames and orders them as a connected path, with any required anchor first. The exact target is excluded. Non-anchor candidates must satisfy temporal separation and are filtered when both horizontal position and viewing direction are too close to the target. Pairwise-overlap limits and coverage of target-visible content and the broader scene are enforced. For MCVI, we additionally exclude views identified as equivalent answers during query validation.

Human and LLM-assisted review. We first screened 2,227 candidate groups with GPT-5.6 for residual target ambiguity and limited context-view diversity. The model received the task type, query, six labeled context images, held-out target, and anchor indicator, and returned a binary decision. The prompt restricted target use to consistency checks, not ambiguity resolution. This stage retained 2,062 groups and rejected 165. The exact LLM screening prompt appears below.

Three experts in multimodal research then independently reviewed all 2,062 retained groups. All three were authors but had not participated in benchmark construction. They assigned initial accept-or-reject labels based on query-target consistency, target ambiguity, and context adequacy and diversity. Initial judgments yielded 1,598 unanimous acceptances and 297 unanimous rejections. Of the 167 disputed groups, 127 received two accept votes and 40 received two reject votes. Before discussion, three-way agreement was 91.90% (1,895/2,062), and nominal Krippendorff's α was 0.815. Disputed groups were adjudicated primarily through discussion. The final review accepted 1,692 groups and rejected 370.

Release. To improve the balance of task-specific displacement and rotation bins, we subsampled accepted groups in Translation, Rotation, and Composition, omitting 191 groups. One additional ROG group was excluded after its context images triggered a safety refusal from GPT-Image-2. The frozen benchmark contains 1,500 groups from 513 scenes across the three source datasets.

GPT-5.6 screening prompt.

You are screening one MindView candidate sample.

Inputs: - TASK: {task_type} - QUERY: {query} - Six images labeled CONTEXT 1–6
- One held-out image labeled TARGET - ANCHOR: {1 or NONE}

Return PASS only if:

1. The QUERY and CONTEXT images uniquely constrain the target camera position and viewing direction. Return FAIL if objects or spatial cues are ambiguous, or if substantially different viewpoints could equally satisfy the QUERY or produce similarly plausible observations due to repeated structures, scene symmetry, or insufficient cross-view evidence.
2. The six CONTEXT images provide complementary scene information and are not dominated by duplicate or near-duplicate viewpoints.

Use TARGET only to check consistency, not to resolve ambiguity. Use only visible evidence. If uncertain, return FAIL. **Return exactly one word:** PASS or FAIL

A.4 Task Examples and Generation Prompts

Figures 7–13 show one released example for each task in the order introduced in Section 3.1. Each example contains six context images ordered from left to right, the held-out ground-truth target (GT), and the full generation prompt. A red border and "A" mark the first image when an anchor is required. The target is shown for reference and is not a generation input.

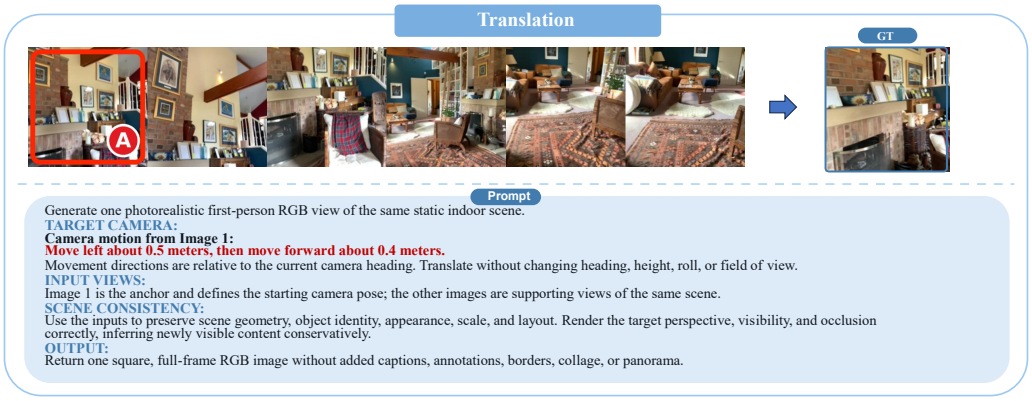


Figure 7: Translation example from ARKitScenes.

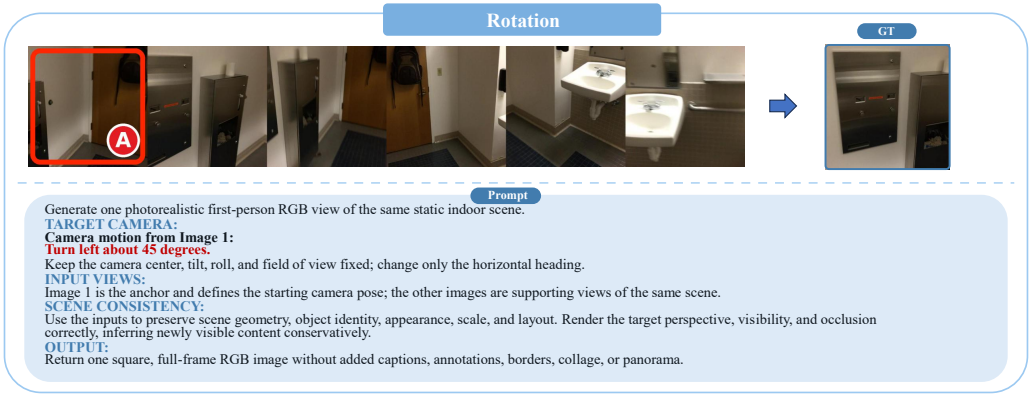


Figure 8: Rotation example from ScanNet.

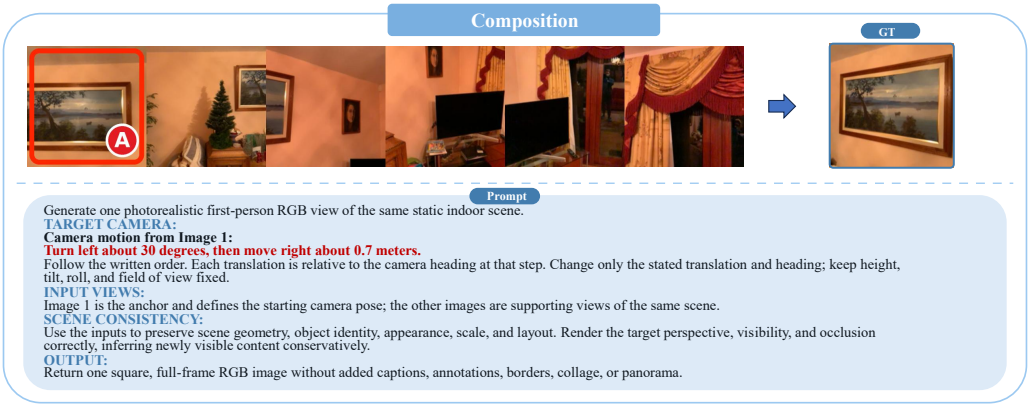


Figure 9: Composition example from ARKitScenes.

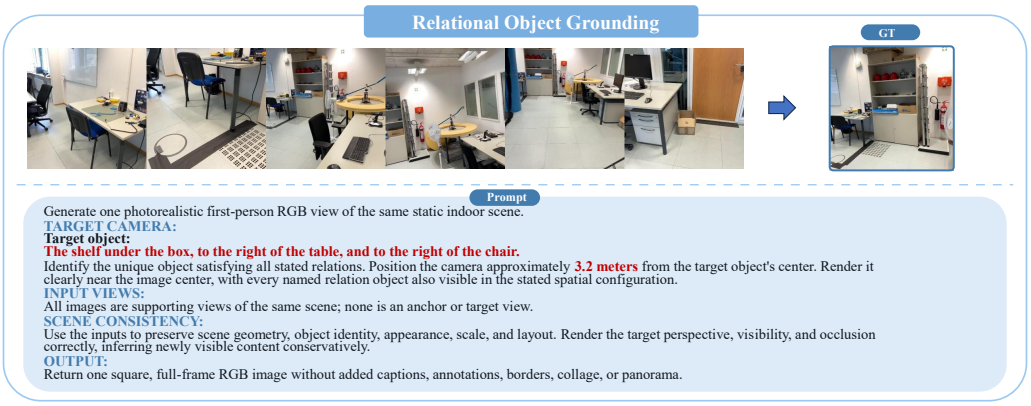


Figure 10: Relational Object Grounding example from ScanNet++.



Figure 11: Target-Facing Reorientation example from ScanNet.

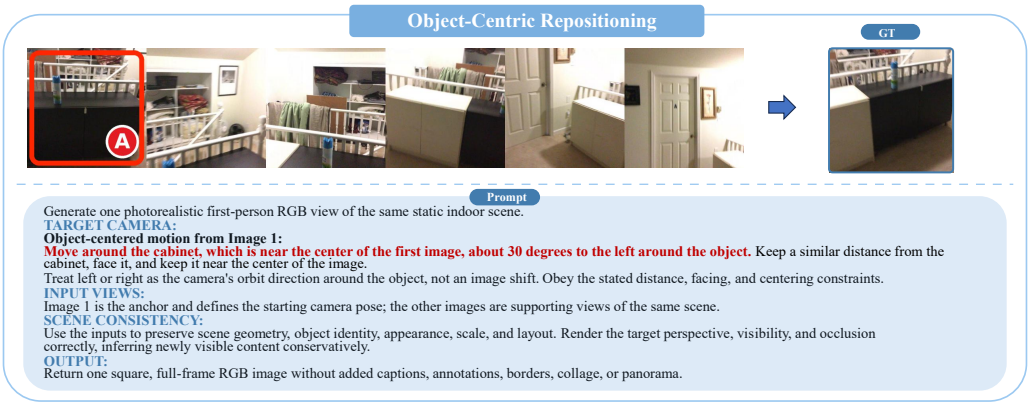


Figure 12: Object-Centric Repositioning example from ScanNet.

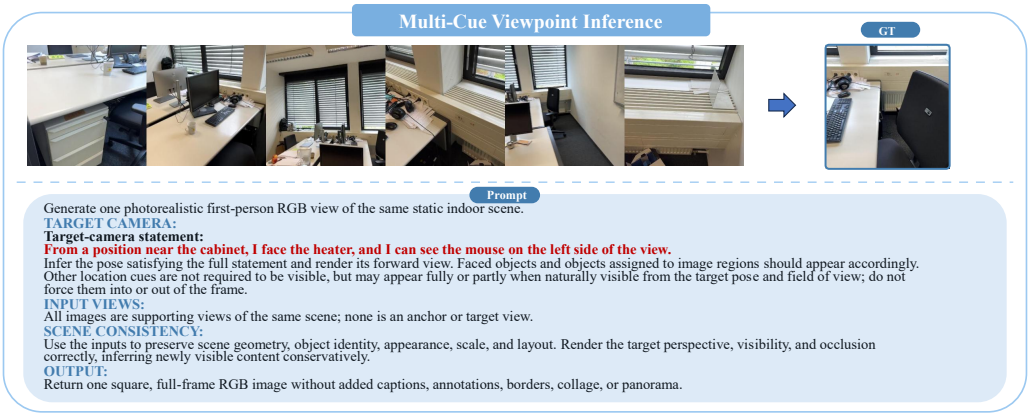


Figure 13: Multi-Cue Viewpoint Inference example from ScanNet++.

B Evaluation Protocol Details

This section details the six metrics in Section 3.3, which assess agreement with the real target and the generation's intrinsic quality.

B.1 Image-Feature Similarity

DINO-I compares the visual content of a generated image with its real target using the pretrained DINOv2-Large encoder (Oquab et al., 2023). Let $\mathbf{f}_i^{\text{gen}}$ and $\mathbf{f}_i^{\text{real}}$ denote their unit-normalized final-layer CLS embeddings. The score is their cosine similarity scaled by 100,

$$\text{DINO-I}_i = 100\,(\mathbf{f}_i^{\text{gen}})^\top \mathbf{f}_i^{\text{real}}. \tag{5}$$

No clipping or offset is applied, so negative similarities remain negative. Before feature extraction, both images are converted to RGB, center-cropped to a square, and resized to 512×512 with Lanczos interpolation. The encoder's standard preprocessing then produces normalized 224×224 inputs.

B.2 VLM-Based Assessment

We use Qwen3-VL-30B-A3B-Instruct (Qwen Team, 2025b) for two separate evaluations. The first receives the real target followed by the generation, labeled as the reference and candidate images, respectively. It returns separate Semantic and Spatial scores for scene-specific content and agreement in viewpoint and layout. The second receives only the generation and assigns a Quality score based on its internal coherence and visible defects.

Neither evaluation receives the query, context views, task type, camera metadata, or the other evaluation's response. Spatial therefore measures agreement with the target image, not direct compliance with the language query. The first evaluation uses the initial RGB conversion, square crop, and 512×512 resize described above. For Quality, the full RGB image is reduced to fit within 512×512 without upscaling, preserving its aspect ratio, borders, and peripheral defects. The judge's standard image preprocessing follows in both cases.

Each evaluation uses deterministic decoding with at most 256 output tokens. The three ratings must be integers from 0 to 10 and are multiplied by ten for reporting. Missing or invalid ratings constitute evaluation failures rather than zero scores. Appendix B.6 reproduces both prompts verbatim, apart from typesetting and line wrapping.

B.3 Geometric Consistency

We use the pretrained VGGT-1B model (Wang et al., 2025a) to measure agreement between the generated and real target viewpoints. For each sample, VGGT jointly processes ten ordered reference views, the real target, and the generation, in that order. The references are drawn from the sample's ten-context set and held fixed across generators and context budgets. Images undergo the same initial square preprocessing and are then resized to 518×518 for VGGT.

Camera-pose comparison. Joint inference places the reference views and both target images in a shared reconstructed coordinate system for each evaluation. For sample i, $k \in \{\text{gen}, \text{real}\}$ indexes the generated and real target images. Let $\mathbf{x}_w \in \mathbb{R}^3$ denote a point's coordinates in this world frame and $\mathbf{x}_c \in \mathbb{R}^3$ its coordinates in image k's camera frame. The estimated rotation $\hat{R}_{i,k} \in \mathrm{SO}(3)$ and translation $\hat{\mathbf{t}}_{i,k} \in \mathbb{R}^3$ map world coordinates to camera coordinates through $\mathbf{x}_c = \hat{R}_{i,k}\mathbf{x}_w + \hat{\mathbf{t}}_{i,k}$. Setting $\mathbf{x}_c = \mathbf{0}$ gives the camera center in the world frame, $\hat{\mathbf{c}}_{i,k} = -\hat{R}_{i,k}^{\top}\hat{\mathbf{t}}_{i,k}$. We compare camera centers, rather than extrinsic translation vectors, and compute the errors in Equation 2 as

$$
\begin{aligned}
e_{\text{rot},i} &= \frac{180}{\pi} \arccos\left(\text{clip}\left(\frac{\text{tr}(\hat{R}_{i,\text{gen}}\hat{R}_{i,\text{real}}^{\top}) - 1}{2}, -1, 1\right)\right), \\
e_{\text{trans},i} &= a_i \left\|\hat{\mathbf{c}}_{i,\text{gen}} - \hat{\mathbf{c}}_{i,\text{real}}\right\|_2 .
\end{aligned}
\tag{6}
$$

Here, tr denotes the matrix trace, $\|\cdot\|_2$ the Euclidean norm, and $\text{clip}(x, l, u)$ restricts x to $[l, u]$. The rotation error is in degrees. The positive scale a_i in Equation 7 converts reconstructed lengths to meters, so $e_{\text{trans},i}$ measures camera-center displacement in meters.

Depth-based scale calibration. We estimate a_i from sensor depths for the ten reference views and the real target, denoted collectively by $\mathcal{J}_i$. Measured depths are converted to meters, and invalid readings are excluded. Nearest-neighbor resampling aligns them with the predicted depth maps under the same centered image crop. For each real frame, we first retain pixels with finite measured and predicted depths, positive measured depth, and predicted depth above 10^{-8}. When at least 512 such pixels have finite depth confidence, we select those at or above their median confidence, provided that at least 256 remain. Otherwise, we retain the original valid set.

Let $\mathcal{V}_{if}$ be the retained pixels in frame f, with measured depth d_{ifp}^{sens} and predicted depth $\hat{d}_{ifp}$ at pixel p. We aggregate log-depth ratios first within each frame and then across frames,

$$
a_i = \exp\left[\text{median}_{f \in \mathcal{J}_i}\left\{\text{median}_{p \in \mathcal{V}_{if}}\left(\log d_{ifp}^{\text{sens}} - \log \hat{d}_{ifp}\right)\right\}\right]. \tag{7}
$$

Calibration requires at least 256 retained pixels per real frame and a finite, positive a_i. The generation participates in joint inference but contributes no sensor depth to calibration. The real target and sensor depths are used only for evaluation, and neither pose error requires ground-truth camera poses.

Geometric scores. We convert the errors to higher-is-better scores before averaging samples,

$$
\text{RScore}_i = 100\ \text{clip}\left(1 - \frac{e_{\text{rot},i}}{60^{\circ}}, 0, 1\right), \qquad \text{TScore}_i = 100\ \text{clip}\left(1 - \frac{e_{\text{trans},i}}{2\,\text{m}}, 0, 1\right). \tag{8}
$$

Each score decreases linearly from 100 at zero error to zero at its threshold and remains zero for larger errors.

B.4 Evaluator Reliability

Human–VLM ranking agreement. We assess whether the VLM judge reflects human preferences on a score-blind subset of 210 standard-six-view samples, with 30 samples from each task. The same three multimodal-research experts who conducted the benchmark review in Appendix A.3 independently rate outputs from the four models in Table 5 using the same Semantic, Spatial, and Quality criteria as the judge. The table reports the expert-averaged human scores and VLM scores on the same subset.

Table 5: **Human and VLM-judge component scores on the shared 210-sample subset.** Human scores are three-expert averages; all values are on a 0–100 scale, and higher is better.

Model	Semantic		Spatial		Quality	
	Human	VLM	Human	VLM	Human	VLM
GPT-Image-2	68.4	63.4	63.3	50.2	76.3	84.8
Nano Banana Pro	62.6	59.5	59.5	47.4	72.5	79.8
SenseNova-U1.5	57.6	54.1	51.5	43.5	66.8	70.3
BAGEL	45.5	45.4	43.0	33.4	44.1	48.6

For 194 of the 210 samples (92.4%), all three experts independently give the same four-model ranking, and this unanimous human ranking exactly matches the ranking produced by the VLM judge. Absolute scores nevertheless reveal a consistent calibration shift. For every model, the VLM judge assigns lower Semantic and Spatial scores and higher Quality scores than the experts. Despite this shift, all six aggregate pairwise model comparisons agree in direction for each component (18/18 overall), yielding the same complete ordering: GPT-Image-2, Nano Banana Pro, SenseNova-U1.5, and BAGEL. These results support the judge's comparative validity on the evaluated subset despite differences in absolute calibration.

Reliability of VGGT-based geometric scores. Joint reconstruction introduces a different concern: a degraded generation could perturb the real-view geometry against which it is scored. We evaluate this sensitivity quantitatively and inspect score behavior in a severe qualitative case. We sample 30 cases uniformly without replacement per task, covering 210 samples from 150 scenes. For each, we substitute the standard-six-view outputs of BAGEL, BAGEL-SFT, FLUX.2, and GPT-Image-2, fixing the ten reference images, real target (GT), input order, and preprocessing. Each reconstruction is independently calibrated to meters from the eleven real views' sensor depths (Equation 7). As a repeatability control, identical BAGEL inputs are rerun on three samples per task; all 21 reruns yield exactly identical saved poses, intrinsics, depths, confidence maps, and calibration scales.

We compare the fixed real views across generator conditions. For the GT, ΔR averages changes in its relative rotations to the ten reference cameras, and ΔC measures camera-center displacement after rigid alignment fitted only to those reference centers, without rescaling. For the references, ΔR and ΔL average relative-rotation changes and absolute camera-distance changes over all 45 camera pairs. These diagnostics remove global-coordinate ambiguity while retaining differences in calibrated scale. For each real image, depth change is $\Delta D = 100\ \mathrm{median}_p\big[2|d_p^{(a)} - d_p^{(b)}|/(d_p^{(a)} + d_p^{(b)})\big]$, where a, b index generator conditions, using calibrated depths on the full 518×518 grid and a fixed valid-sensor-depth mask across conditions. We report GT depth change separately and average the ten per-image changes for the references. Table 6 reports medians across samples after averaging each diagnostic over the six unordered generator pairs within each sample; the identical-input reruns are excluded. The pooled median changes are 0.71°, 2.75 cm, and 0.76% for GT rotation, position, and depth; reference changes are 0.75°, 2.63 cm in pairwise distance, and 0.87% in depth. These small median changes indicate that outputs from different generators typically have little effect on the reconstructed geometry of the fixed real views, supporting the stability of VGGT-based geometric evaluation.

Table 6: **VGGT reference stability across generators.** Only the generation changes; ten contexts and the real target stay fixed. Entries are sample medians after averaging six generator pairs per sample. ΔR: relative-rotation change; ΔC: aligned GT-center displacement; ΔL: context-pair distance change; ΔD: symmetric relative depth change. Distances are depth-calibrated; lower values indicate greater stability.

Task	N	Real target (GT)			Contexts		
		ΔR (°)	ΔC (cm)	ΔD (%)	ΔR (°)	ΔL (cm)	ΔD (%)
Translation	30	0.85	4.96	1.09	1.09	6.02	1.39
Rotation	30	0.52	1.92	0.58	0.65	2.14	0.79
Composition	30	0.96	4.60	0.95	0.88	4.71	1.18
ROG	30	0.82	3.25	0.99	0.93	2.62	0.97
TFR	30	0.63	1.60	0.55	0.70	1.88	0.80
OCR	30	0.50	1.96	0.48	0.54	2.04	0.67
MCVI	30	0.74	2.84	0.75	0.77	2.33	0.77
All sampled cases	210	0.71	2.75	0.76	0.75	2.63	0.87

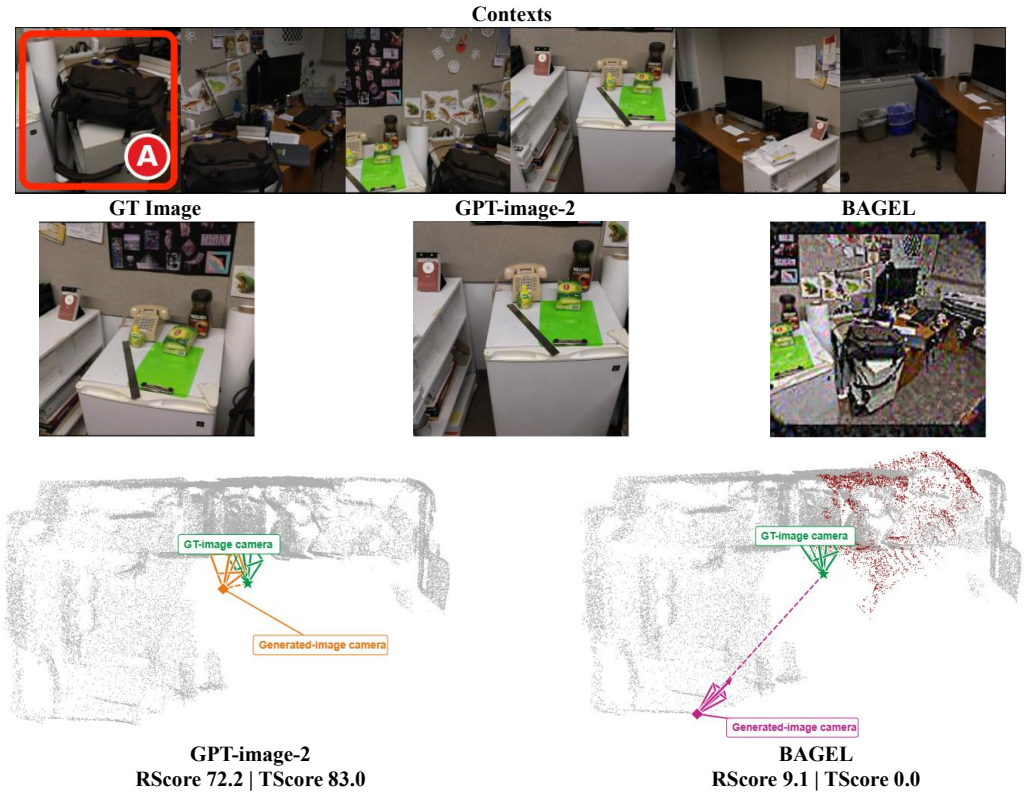


Figure 14: **VGGT evaluation of GPT-Image-2 and BAGEL outputs.** For the ScanNet TFR query "Turn to face the telephone," the top row shows six generation contexts (**A**: anchor); evaluation uses ten fixed references, the real target, and one generation. Reconstructions are independently depth-calibrated and rigidly aligned using only the ten reference camera centers, without scale fitting. All cameras are VGGT estimates; GT denotes the real target image. Dark red marks points contributed by BAGEL's generation; all other scene points are gray.

Figure 14 complements the aggregate analysis with a visibly degraded generation. GPT-Image-2 preserves the target-facing scene content, whereas BAGEL introduces severe appearance and layout distortions. Their RScore/TScore values are 72.2/83.0 and 9.1/0.0, respectively. The reference scene and estimated GT camera remain visually similar after alignment, while BAGEL's estimated camera

lies much farther from the estimated target camera. This case illustrates score behavior consistent with the visible target-view discrepancy despite the degraded generation.

Together, the quantitative and qualitative results support RScore and TScore as meaningful measures of target-view geometric consistency. Reference geometry exhibits small median changes across model outputs. In the illustrated case, it remains visually stable despite severe generation artifacts, while the scores reflect the visible viewpoint discrepancy.

B.5 Score Aggregation

For each generator and context setting, let s_{itj} be sample i's score on metric j for task t after the conversions above, and let n_t be the number of samples in that task. We first average samples within each task, then give equal weight to the six components and the seven tasks,

$$\bar{s}_{tj} = \frac{1}{n_t} \sum_{i=1}^{n_t} s_{itj}, \qquad S_t = \frac{1}{6} \sum_{j=1}^{6} \bar{s}_{tj}, \qquad \text{Overall} = \frac{1}{7} \sum_{t=1}^{7} S_t. \tag{9}$$

Each benchmark-level component score is likewise averaged equally across the seven tasks, so tasks with more samples do not receive greater weight.

Aggregation requires valid scores for every benchmark sample. Failed or missing evaluations are neither assigned zero nor silently omitted. Scores and gains are displayed to one decimal place, while differences and rankings are computed before rounding. Overall is a composite score, not a calibrated accuracy. Comparisons are descriptive, without significance tests.

B.6 Evaluator Prompts

Semantic and Spatial prompt.

You are a strict, task-independent evaluator of image fidelity. The labeled REFERENCE IMAGE is the ground-truth target. The labeled CANDIDATE IMAGE is the model prediction.

Important rules: - The two images may or may not depict the same scene. Do not assume that they match. - Inspect each image independently before comparing them. - Use only visible pixel evidence. - Never attribute an object, region, or property to one image merely because it appears in the other. - Do not use or infer any hidden prompt, context images, task type, or camera metadata. - Evaluate Semantic Fidelity and Spatial Fidelity independently. - Do not judge aesthetic preference or technical image quality in this evaluation.

1. Semantic Fidelity: WHAT is present

Evaluate: - scene identity and scene category; - main objects, structures, and recognizable content; - approximate object count and identity-defining attributes; - missing, extra, substituted, hallucinated, or unrecognizable content.

Do NOT evaluate: - camera viewpoint, orientation, framing, crop, or visible side by itself; - object position, scale, depth, occlusion, or relative layout; - global color tone, lighting, rendering style, sharpness, or visual polish.

A spatially displaced object can still be semantically correct. A viewpoint difference alone must not reduce Semantic Fidelity if the same important content remains visible and recognizable. However, important content that is visibly missing or hallucinated is a semantic error, regardless of its cause.

Semantic Fidelity scoring: 10 = The same scene-specific content is preserved almost completely. All major objects and structures are present and recognizable, with no meaningful missing, extra, substituted, or hallucinated content. 8 = The same scene and all major content are preserved, with only minor errors involving secondary objects or identity-defining attributes. 6 = The core scene is clearly recognizable, but one major object or several secondary elements are missing, substituted, hallucinated, or unrecognizable. 4 = Only partial scene-specific content is preserved. Multiple major objects or structures are incorrect, although some meaningful correspondence remains. 2 = The images share only a generic scene category or a few isolated object categories; scene-specific correspondence is very weak. 0 = No meaningful semantic correspondence, or the candidate contains no interpretable scene content.

Semantic score constraints: - Generic category similarity alone, such as both images being bedrooms or offices, cannot receive a score above 3. - If one major or scene-defining object is missing, replaced, or hallucinated, the score cannot exceed 6. - If multiple major objects are incorrect, the score cannot exceed 4. - A score of 9 or 10 requires nearly complete scene-specific correspondence, not merely a similar category or overall impression.

2. Spatial Fidelity: WHERE things are and FROM WHERE the scene is viewed

Evaluate: - camera viewpoint and orientation; - framing, crop, and visible scene regions or surfaces; - relative object positions, scale, depth ordering, and occlusion; - scene layout, geometry, and composition.

Semantic and Spatial prompt (continued).

Judge spatial relations using scene elements that can be meaningfully corresponded between the two images.

Do NOT evaluate: - object category or semantic identity by itself; - color, lighting, texture, rendering style, sharpness, or visual attractiveness; - isolated generation artifacts unless they make spatial relationships impossible to assess.

Do not reward a generic similar composition unless corresponding scene elements support it. If meaningful spatial correspondence cannot be established, use a score of 0, 1, or 2.

Spatial Fidelity scoring: 10 = Viewpoint, orientation, framing, visible regions, and spatial layout essentially match, with only negligible offsets. 8 = The intended view and layout are correct, with only minor shifts, crop or scale differences, or small geometric errors. 6 = The overall view is similar, but noticeable differences exist in camera pose, framing, visibility, scale, or relative placement. 4 = Some spatial correspondence exists, but the viewpoint, visible region, or layout is clearly incorrect. 2 = Very little spatial correspondence remains; there is a major camera-pose, framing, or layout mismatch. 0 = No usable spatial correspondence, or spatial structure cannot be assessed.

Spatial score constraints: - A clearly incorrect viewing direction or target-facing direction cannot receive a score above 4. - A major mismatch in the visible scene region or room side cannot receive a score above 4. - Generic centered composition without scene-specific correspondence cannot receive a score above 2. - A score of 9 or 10 requires both camera view and scene layout to match closely.

Use integers from 0 to 10. Integers between the stated anchors may be used when justified. Apply all score constraints before producing the final scores.

Each reason must identify the strongest visible evidence and explicitly distinguish the Reference from the Candidate. Keep each reason to one short sentence.

Return ONLY valid JSON in exactly this format: `{ "semantic_score": <integer 0-10>, "spatial_score": <integer 0-10>, "semantic_reason": "<short evidence-based reason>", "spatial_reason": "<short evidence-based reason>" }`

Quality prompt.

You are a strict evaluator of the intrinsic visual quality of a generated image. Only the labeled CANDIDATE IMAGE is provided. Evaluate whether the image is visually clean, structurally coherent, and free of generation defects. Use only visible pixel evidence. Judge the candidate image independently, without assuming any unseen reference, prompt, context image, task, or intended camera pose.

Evaluate: - malformed, melted, stretched, fused, or broken objects and structures; - internally inconsistent room, object, perspective, or surface geometry; - duplicated or unnaturally repeated elements; - seams, corrupted regions, incomplete rendering, and white or black borders; - unintended text, watermarks, blank regions, or padding; - unintended blur, pixelation, noise, aliasing, or compression artifacts; - implausible edges, lighting discontinuities, and inconsistent shadows.

Do NOT evaluate: - whether the image matches an unseen reference, prompt, task, or intended viewpoint; - whether its objects or layout are correct for an external target; - semantic fidelity or spatial fidelity; - aesthetic preference or similarity to a particular rendering style.

Do not reward an image merely because it looks attractive, polished, colorful, or cinematic. Do not penalize unusual content when it is internally coherent. Judge coherence within the image's own visual domain.

Visual Quality scoring: 10 = Technically clean and internally coherent. No meaningful deformation, broken geometry, duplication, border, seam, corruption, or unintended artifact is visible. 8 = High visual quality. Only minor and localized defects are visible, without affecting important objects or structures. 6 = Generally usable, but one prominent defect or several noticeable local artifacts are present. 4 = Multiple prominent defects affect important regions, such as malformed objects, inconsistent geometry, repeated structures, seams, blur, or borders. 2 = Severe and widespread distortion, incomplete rendering, broken perspective, large blank regions, or obvious corruption makes the image difficult to use. 0 = Blank, invalid, unreadable, almost completely corrupted, or a failed generation.

Visual Quality score constraints: - One obvious structural deformation in a salient region prevents a score above 7. - Multiple visible deformations, repetitions, seams, or broken structures prevent a score above 5. - A large white or black border, incomplete rendering, or corrupted region prevents a score above 4. - A score of 9 or 10 requires careful inspection to reveal no meaningful defect. - Attractive lighting, color, or visual polish alone must not increase the score.

Use an integer from 0 to 10. Integers between the stated anchors may be used when justified. Apply all score constraints before producing the final score.

The reason must identify the most important visible defect or, when no meaningful defect exists, explicitly state that the image is visually coherent. Keep the reason to one short sentence.

Return ONLY valid JSON in exactly this format: `{ "visual_quality_score": <integer 0-10>, "visual_quality_reason": "<short evidence-based reason>" }`

C Training and Inference Details

C.1 Training Resource

The released training resource contains 20,000 examples from 3,406 scenes, each with six context images, a language query, and a real target image. Training examples are constructed from the training splits of all three source datasets, while the benchmark uses their validation splits for held-out evaluation. The training resource and benchmark share no scenes within any source dataset.

C.2 Released Training Configuration

The release configuration specifies 10,000 optimizer updates with eight distributed processes and no gradient accumulation. It uses AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-5}, cosine decay to 10^{-6}, 200 warmup steps, gradient clipping at 0.5, and EMA decay 0.9995. The VAE and vision transformer are frozen, while the language model and understanding branch remain trainable. Conditioning dropout is disabled. Images are center-cropped and resized to 512×512 for the VAE and 518×518 for the vision transformer, with training timestep shift 4. These settings describe the released training recipe.

C.3 Generation Configuration

The saved BAGEL-SFT inference configuration uses 50 sampling steps, text and image guidance scales of 4 and 2, timestep shift 3, and no-thinking inference. Input and saved output images are 512×512. Local generators share a deterministic per-sample seed across the supplementary context settings. Closed-source APIs receive no seed. In our setup, directly supplying six separate images to Qwen-Image-Edit-2511 and Emu3.5 led to inference failures or severely degraded outputs. We therefore pack supporting views into a contact sheet and provide the anchor separately when present. For anchor-free tasks, all context views are packed into a single contact sheet.

D Spatial-Understanding Evaluation Details

All results in Table 3 are obtained in our own evaluations. SAT-v2 uses 150 questions evaluated in two answer orders. Micro accuracy counts correct predictions among all 300 responses, while both-order accuracy counts questions answered correctly in both orders among the 150 pairs. BAGEL and BAGEL-SFT use identical inputs and greedy, non-thinking decoding. Their correct counts are 174 and 188 responses, respectively, and 41 and 62 complete pairs. BAGEL-SFT answers 322 of 1,000 MMSI-Bench questions correctly under the Vanilla non-circular protocol and 383 of 1,050 MindCube-Tiny questions under raw-QA evaluation. Our reproduced BAGEL accuracies are 31.0% and 34.7%, respectively, matching SenseNova-SI (Cai et al., 2026) at the reported precision. Eval. Size in Table 3 denotes the number of evaluated responses or question pairs under each protocol.

E MindCue Implementation

MindCue uses a target-view hypothesis to identify useful additional evidence while preserving the complete core. Following Section 4.5, $\mathcal{I}_i$ indexes the ordered input pool, $\mathcal{F}_i$ the core, and $\mathcal{D}_i$ the additional views. The selector uses only input RGB images, the query, task label, and core and anchor metadata, with no access to held-out targets, sensor depth, ground-truth poses, generated images, or evaluation scores.

E.1 Scene Evidence

Frozen VGGT jointly predicts camera poses, intrinsics, depth, and confidence, placing the input views in a shared reconstruction. For semantics-dominant tasks, a fixed parser extracts queried entities and frozen SAM 3 segments them. After confidence filtering and duplicate suppression, masks are lifted into the reconstruction and grouped into object clusters. Each cluster must include a core observation, while additional views can refine its geometry. Geometry-dominant tasks use the predicted geometry without object masks.

Table 7: Task-specific target-view hypotheses in MindCue.

Task	Target-view hypothesis
Geometry-Dominant Tasks	
Trans.	Move from the reference position along its horizontal camera axes by the requested displacements, keeping its orientation and intrinsics.
Rot.	Keep the reference position and intrinsics. Rotate about world-up by the requested yaw, with left positive and right negative. A turn-around is 180°.
Comp.	Apply translations and rotations in query order from the reference pose. Each displacement uses the current horizontal camera axes, accounting for preceding rotations. Keep the reference intrinsics.
Semantics-Dominant Tasks	
ROG	For each core-supported target centroid, place cameras at the requested distance along rays toward input-camera centers. Face the centroid and inherit source intrinsics; select the highest-scoring object–camera pair.
TFR	Keep the reference center, roll, and intrinsics, and face the core-supported target centroid.
OCR	Orbit the reference camera around the target centroid about world-up by the requested angle, preserving distance, roll, and intrinsics; face the target.
MCVI	For each facing-object candidate, reorient cameras where it is detected toward its centroid, retaining centers and intrinsics; select the highest-scoring pair.

E.2 Target-View Hypothesis

The query and scene evidence define a hypothesized target camera $\hat{\mathcal{C}}_i^\star$ with pose and intrinsics. This camera guides evidence selection; it is not supplied to the generator as a pose condition. For tasks with an explicit anchor, MindCue updates the first input camera according to the requested motion or object reference. ROG and MCVI instead compare candidates derived from predicted objects and input-camera positions, as detailed in Table 7. World-up is derived from predicted camera poses.

ROG ranks candidates by a fixed weighted combination of relation consistency, target confidence, and proximity to the source input camera. MCVI instead combines facing-object centering, satisfaction of available constraints, and observed semantic support. ROG requires evidence for every parsed relation object, whereas MCVI omits unsupported auxiliary constraints and treats generic *between* cues as proximity to both objects.

E.3 Reliable Target Coverage

To measure target coverage, we construct sparse per-view point sets $\hat{\mathcal{P}}_{iv}$ through depth sampling, confidence filtering, and unprojection. Projecting their union through $\hat{\mathcal{C}}_i^\star$ retains the nearest depth in each occupied grid cell, forming the virtual surface $\mathcal{S}_i$ in Equation 3. View v supports a cell when one of its projected points agrees with this surface within a depth tolerance. Thus, support requires depth agreement, not projection overlap alone.

Let $\mathcal{V}_{iv} \subseteq \mathcal{S}_i$ be the supported cells and $\mathcal{V}_{i,-v} = \bigcup_{u \in \mathcal{I}_i \setminus \{v\}} \mathcal{V}_{iu}$ the support from all other input views. For a nonempty surface, reliable coverage is

$$R_i(v) = \frac{|\mathcal{V}_{iv} \cap \mathcal{V}_{i,-v}|}{|\mathcal{S}_i|} + \min\left\{ \frac{|\mathcal{V}_{iv} \setminus \mathcal{V}_{i,-v}|}{|\mathcal{S}_i|}, \eta \right\}. \tag{10}$$

The first term credits cross-view support; the second caps the total contribution of evidence supplied only by the candidate. We use $\eta = 0.05$ and retain every additional view whose reliable coverage reaches $\tau = 0.20$. Cells are counted equally, and support is computed once from the original pool, including views subsequently discarded. Unlike diagnostic Target Coverage in Section 4.3, this score uses only predicted input geometry and a hypothesized camera.

E.4 Connectivity and Final Selection

MindCue constructs the retained additional-view set A_i while preserving the complete core $\mathcal{F}_i$. If the virtual surface $\mathcal{S}_i$ is empty, $A_i = \varnothing$ and the generator receives only the core. Otherwise, it first retains every additional view $v \in \mathcal{D}_i$ with reliable target coverage $R_i(v) \geq \tau = 0.20$.

For tasks with an explicit anchor, one below-threshold view may also be retained to connect the core to the target hypothesis. Candidates must meet minimum coverage and connection-score requirements and pass nonredundancy checks against the core and coverage-selected views. We select the eligible candidate maximizing

$$Q_i(v) = \sqrt{R_i(v) \max_{u \in \mathcal{F}_i} \delta_i(v, u)}. \tag{11}$$

Here $\delta_i(v, u)$ is the larger of the two directed overlaps: the fraction of source points reprojecting into the other view with consistent depth, or zero for an empty source. The selected view forms X_i; without an anchor or an eligible candidate, $X_i = \varnothing$.

For a nonempty surface, the final additional-view set is $A_i = \{v \in \mathcal{D}_i : R_i(v) \geq \tau\} \cup X_i$, as defined in Equation 4. All coverage-selected additions are retained; connectivity adds at most one further view. The generator receives the original query and the pool views indexed by $\mathcal{F}_i \cup A_i$, preserving their original order and every core view.

E.5 Experimental Controls

Our experiments use $C_{10,i}$ as the pool and $C_{6,i}$ as the core. RandDrop randomly selects additional views to match MindCue's per-sample count, whereas CounterSel retains the additions rejected by MindCue and is not count-matched. Both preserve the core and input order, with generator checkpoints and evaluation references fixed across policies. MindCue uses the same selection for BAGEL and BAGEL-SFT.

E.6 Runtime

On a single NVIDIA A100-SXM4-80GB, MindCue adds 1.44 s per sample on average, raising the summed latency from 42.48 s to 43.91 s with identical selected images and order (Table 8). The 3.4% overhead is the ratio of unrounded weighted mean MindCue and BAGEL generation times. VGGT reconstruction and SAM 3 segmentation with mask lifting contribute 0.57 s and 0.82 s on average, respectively. The SAM 3 mean includes zero time for geometry-dominant tasks.

We use 63 samples stratified by task and dataset, with three repetitions. Dataset means are weighted by their release frequencies within each task, then averaged equally across the seven tasks. Models are preloaded and timed separately with GPU synchronization. These stage-wise sums exclude RGB reading and preprocessing, model loading and warmup, inter-stage handoff, output saving, and evaluation.

Table 8: Runtime on a single A100 80GB. View counts refer to generator inputs, with identical images and order in the selected-view rows. Means are weighted and rounded independently.

Configuration	Mean views	Mean (s)
BAGEL, C_6	6	38.528
BAGEL, C_{10}	10	50.625
BAGEL, selected views	7.30	42.476
MindCue + BAGEL, same views	7.30	43.913
MindCue component		**Mean (s)**
VGGT reconstruction		0.572
SAM 3 segmentation and mask lifting		0.819
Query parsing, camera planning, and selection		0.046
Total MindCue overhead (+3.4%)		**1.438**

Including RGB reading and initial preprocessing in both MindCue and BAGEL timings gives an additional 1.86 s, or 4.4%. These results indicate that MindCue adds only modest overhead relative to generation under this profiling setup.

F SUPPLEMENTARY RESULTS

This section complements the main experiments with component, context-scaling, context-selection, and qualitative results on the same frozen 1,500 samples and under the same evaluation protocols.

F.1 TASK-LEVEL COMPONENT SCORES ON MINDVIEW

Tables 9–12 report all six component scores for each task under the standard six-view input, complementing the task composites in Table 2. The comparison includes eleven off-the-shelf generators and BAGEL-SFT. BAGEL-SFT scores higher than BAGEL on all 42 task–component pairs and ranks first among the twelve models on 31 of them. It leads Semantic, Spatial, and TScore on all seven tasks. Its Quality score also exceeds BAGEL on every task, although GPT-Image-2 remains highest on that component throughout. Task abbreviations follow Section 3.1. Bold and underline indicate the highest and second-highest scores across all twelve models within each task. Sem., Spa., and Qual. denote Semantic, Spatial, and Quality. † denotes tiled context views, as in Table 1.

Table 9: Component scores on MindView for Translation and Rotation.

Model	Trans.						Rot.					
	DINO-I	Sem.	Spa.	Qual.	RScore	TScore	DINO-I	Sem.	Spa.	Qual.	RScore	TScore
BAGEL	47.7	44.2	34.5	48.6	47.8	27.5	45.8	41.4	30.4	46.4	33.7	35.1
FLUX.2	53.2	49.8	40.9	85.3	59.0	29.7	46.7	42.4	33.9	78.5	31.9	48.7
Qwen-Image-Edit-2511†	57.3	51.8	41.0	67.3	67.8	35.4	43.1	36.9	29.3	69.8	32.2	63.8
Emu3.5†	52.0	46.0	38.0	69.6	**80.1**	48.4	38.7	32.0	24.5	63.3	30.3	72.5
DreamOmni2	56.3	51.3	41.8	61.8	74.2	41.1	39.8	31.8	25.5	60.8	23.7	63.7
SenseNova-U1	51.4	45.5	37.9	59.0	46.9	24.5	48.7	41.5	34.0	59.7	20.4	36.3
SenseNova-U1.5	57.8	50.7	41.8	71.3	62.9	37.2	48.9	43.4	35.2	65.7	38.0	49.4
Nano Banana	52.4	50.5	41.0	74.1	47.8	29.9	53.3	48.4	37.3	64.5	38.2	46.8
Nano Banana Pro	55.5	55.9	44.4	82.4	53.0	35.9	59.3	54.9	43.5	74.6	44.8	45.1
GPT-Image-1.5	52.8	52.5	42.4	85.4	52.8	26.8	52.2	50.2	39.7	81.1	34.9	29.1
GPT-Image-2	59.1	55.0	45.1	**87.1**	63.6	32.1	57.1	52.9	42.1	**81.8**	50.0	41.3
BAGEL-SFT	**67.1**	**70.4**	**59.0**	63.9	78.4	**64.0**	**72.2**	**70.3**	**56.2**	58.0	**69.1**	**83.5**

Table 10: Component scores on MindView for Composition and Relational Object Grounding.

Model	Comp.						ROG					
	DINO-I	Sem.	Spa.	Qual.	RScore	TScore	DINO-I	Sem.	Spa.	Qual.	RScore	TScore
BAGEL	42.7	34.9	28.0	45.7	29.8	35.5	50.9	48.8	36.2	45.2	42.5	35.4
FLUX.2	43.6	40.0	33.5	80.8	32.9	33.6	54.1	52.0	40.5	81.3	33.2	42.2
Qwen-Image-Edit-2511†	43.7	38.3	30.7	66.5	31.6	39.0	55.2	54.6	42.0	69.3	38.4	39.8
Emu3.5†	38.0	31.3	26.9	63.8	26.8	57.3	49.0	45.5	32.8	58.5	29.9	39.1
DreamOmni2	36.0	28.5	24.4	64.0	19.5	42.6	35.4	29.4	23.6	70.5	11.2	32.3
SenseNova-U1	44.7	37.6	32.5	57.5	24.8	30.2	61.1	55.7	42.5	66.5	30.3	40.1
SenseNova-U1.5	49.4	40.2	34.1	68.2	35.4	31.9	62.3	59.2	47.7	69.3	50.9	51.0
Nano Banana	47.1	44.9	36.5	67.5	39.3	39.5	60.5	59.4	43.8	64.6	45.3	46.9
Nano Banana Pro	46.8	46.1	37.6	79.0	39.5	34.9	65.6	61.1	48.8	85.2	35.7	41.5
GPT-Image-1.5	47.1	45.1	37.3	82.3	40.4	25.5	66.9	66.9	53.1	82.8	45.7	39.6
GPT-Image-2	49.7	46.1	38.1	**85.2**	44.6	25.3	67.1	63.6	50.2	**85.5**	**55.7**	34.4
BAGEL-SFT	**57.4**	**54.3**	**44.8**	60.2	**56.3**	**64.8**	**68.5**	**71.2**	**57.3**	60.9	54.6	**63.4**

Table 11: Component scores on MindView for Target-Facing Reorientation and Object-Centric Repositioning.

Model	TFR						OCR					
	DINO-I	Sem.	Spa.	Qual.	RScore	TScore	DINO-I	Sem.	Spa.	Qual.	RScore	TScore
BAGEL	43.4	42.0	30.9	45.0	32.4	38.3	61.7	59.2	44.1	50.8	27.0	22.0
FLUX.2	50.5	48.5	38.3	76.1	32.7	53.4	70.0	68.4	52.9	77.6	28.4	25.8
Qwen-Image-Edit-2511†	48.8	45.6	35.3	75.6	31.0	43.5	74.2	79.4	60.6	61.9	31.1	34.8
Emu3.5†	40.6	40.8	29.8	73.3	25.7	59.4	62.7	61.2	43.8	68.5	37.7	40.0
DreamOmni2	31.0	25.6	19.2	67.7	13.2	53.4	62.7	60.2	45.6	63.4	23.8	31.5
SenseNova-U1	56.7	54.4	40.5	67.0	18.8	43.3	60.9	57.2	42.7	68.5	15.7	22.5
SenseNova-U1.5	59.3	55.2	43.5	70.3	33.9	58.3	70.2	70.0	53.6	70.5	32.5	31.3
Nano Banana	64.7	64.1	48.2	60.2	43.3	53.6	62.5	59.4	45.7	65.7	26.1	30.3
Nano Banana Pro	67.2	64.5	51.2	76.2	41.1	51.3	68.3	67.7	51.9	76.9	32.2	30.9
GPT-Image-1.5	66.8	68.0	52.9	79.4	42.4	46.8	67.9	72.6	54.0	83.2	23.1	24.7
GPT-Image-2	**71.5**	74.5	58.9	**81.4**	55.6	58.7	74.6	78.2	61.2	**85.2**	37.5	37.8
BAGEL-SFT	70.5	**80.3**	**63.3**	60.3	**59.2**	**65.3**	**78.9**	**86.7**	**69.1**	63.7	**70.0**	**72.2**

Table 12: Component scores on MindView for Multi-Cue Viewpoint Inference.

Model	MCVI					
	DINO-I	Sem.	Spa.	Qual.	RScore	TScore
BAGEL	45.8	41.2	31.4	47.9	30.0	29.3
FLUX.2	47.5	45.8	36.0	82.9	30.4	36.7
Qwen-Image-Edit-2511†	44.6	37.9	31.4	76.2	28.8	25.5
Emu3.5†	32.9	30.6	22.6	73.0	21.4	32.3
DreamOmni2	31.0	24.9	21.2	70.1	11.3	34.5
SenseNova-U1	51.7	45.8	37.8	66.3	22.5	35.6
SenseNova-U1.5	58.1	51.2	42.1	72.3	38.4	43.3
Nano Banana	55.0	51.4	39.6	67.2	33.0	37.7
Nano Banana Pro	57.1	50.2	40.5	85.7	25.4	33.5
GPT-Image-1.5	58.2	56.7	44.9	81.2	39.7	41.8
GPT-Image-2	**61.5**	60.1	47.7	**86.2**	42.6	44.0
BAGEL-SFT	61.0	**63.8**	**50.2**	65.8	**44.0**	**52.5**

F.2 Context-Scaling Component Scores

Table 13 reports task-macro component scores under the reduced C_3 and expanded C_{10} settings. The corresponding standard C_6 scores appear in Table 1. From C_6 to C_{10}, Semantic and Spatial decrease for all eleven off-the-shelf generators, whereas Quality increases for nine. Over the same change, DINO-I and RScore decrease for ten generators, and TScore decreases for nine. For the off-the-shelf generators, the common Overall decline therefore does not imply that every component worsens as context expands. Bold and underline mark the highest and second-highest scores within each context setting, and † again denotes tiled context views.

Table 13: Task-macro component scores under reduced and expanded contexts.

Model	C_3						C_{10}					
	DINO-I	Sem.	Spa.	Qual.	RScore	TScore	DINO-I	Sem.	Spa.	Qual.	RScore	TScore
BAGEL	46.7	43.2	32.4	47.3	33.0	34.0	45.9	40.8	31.0	48.5	32.4	35.5
FLUX.2	49.6	46.4	37.7	**85.9**	33.7	41.5	51.5	48.7	38.1	69.3	32.5	38.0
Qwen-Image-Edit-2511†	50.2	47.1	37.2	69.8	33.8	38.1	51.6	48.1	37.9	71.1	36.2	38.3
Emu3.5†	44.5	38.9	28.6	59.5	34.8	46.8	44.9	39.9	30.6	68.3	33.3	49.1
DreamOmni2	42.7	37.6	30.8	76.9	25.1	36.1	32.2	27.7	22.1	59.0	25.9	51.8
SenseNova-U1	50.3	46.1	36.8	60.1	27.4	35.1	51.6	45.3	36.5	66.4	22.3	32.2
SenseNova-U1.5	54.8	49.1	39.5	69.8	38.8	45.2	57.3	51.8	41.6	72.1	40.6	43.0
Nano Banana	55.1	51.6	39.7	62.8	38.4	44.2	54.5	51.1	39.7	68.9	37.1	39.3
Nano Banana Pro	57.2	54.3	43.4	79.3	37.0	40.9	58.2	54.9	43.5	80.8	36.6	38.6
GPT-Image-1.5	57.6	55.9	44.5	82.1	38.0	37.7	57.2	56.7	44.6	83.2	35.5	31.9
GPT-Image-2	59.0	57.9	46.3	84.8	45.2	41.6	61.2	59.8	47.9	**85.1**	48.5	37.5
BAGEL-SFT	**63.7**	**62.5**	**50.4**	61.4	**59.4**	**65.8**	**67.4**	**69.5**	**55.8**	59.6	**61.7**	**64.9**

F.3 Detailed Results for MindCue

Table 14 decomposes the Overall results in Table 4 into six task-macro components for the same five context policies on BAGEL and BAGEL-SFT. Fixed C_6 and Fixed C_{10} are fixed-context baselines. RandDrop is a count-matched random control evaluated with one selection seed. It retains the same number of additions as MindCue for each sample. CounterSel instead retains the candidate additions rejected by MindCue and is not count-matched. MindCue uses the query and predicted scene evidence to select additions without updating either generator. Bold and underline mark the highest and second-highest policy scores for each model and component. Against RandDrop, MindCue is higher on five of six components for BAGEL and all six for BAGEL-SFT. Quality is the only BAGEL exception, by 0.3 points. The largest gains over RandDrop are in RScore, at 2.6 points for BAGEL and 1.6 points for BAGEL-SFT. Relative to Fixed C_6, MindCue is higher on all six components for BAGEL and five for BAGEL-SFT, with only a 0.3-point decrease in Quality for BAGEL-SFT.

Table 14: Task-macro component scores for the five context policies in Table 4.

Model	Policy	DINO-I	Sem.	Spa.	Qual.	RScore	TScore
BAGEL	Fixed C_6	<u>48.3</u>	<u>44.5</u>	<u>33.6</u>	47.1	<u>34.7</u>	31.9
	Fixed C_{10}	45.9	40.8	31.0	**48.5**	32.4	**35.5**
	CounterSel	43.7	38.4	29.3	<u>48.4</u>	29.5	33.0
	RandDrop	47.3	42.8	32.0	47.9	34.0	33.3
	MindCue	**49.2**	**44.9**	**33.9**	47.6	**36.6**	<u>34.7</u>
BAGEL-SFT	Fixed C_6	<u>68.0</u>	<u>71.0</u>	<u>57.1</u>	**61.8**	61.7	<u>66.5</u>
	Fixed C_{10}	67.4	69.5	55.8	59.6	61.7	64.9
	CounterSel	66.2	67.7	54.5	60.2	61.8	65.4
	RandDrop	67.8	70.4	56.6	61.3	<u>62.0</u>	66.2
	MindCue	**68.9**	**71.8**	**57.7**	<u>61.5</u>	**63.6**	**67.1**

F.4 Qualitative Comparisons

Figure 15 extends Figure 4 from two tasks to all seven. The examples support separate inspection of scene preservation and target-view agreement across task types, complementing the component scores above.

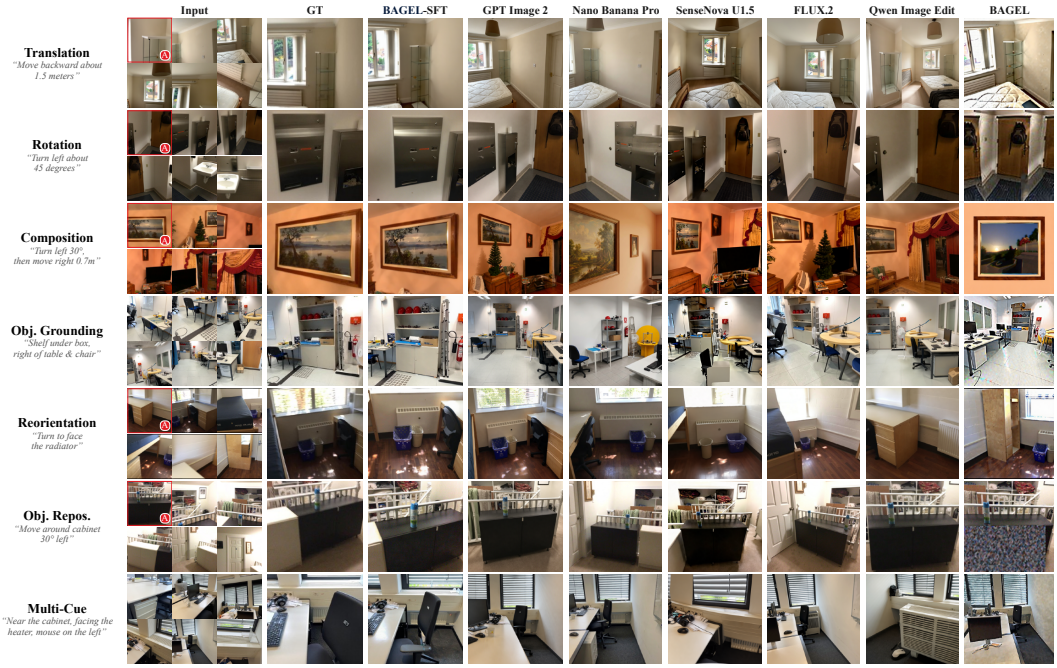


Figure 15: Qualitative comparison across all seven MindView tasks. Each row shows a query summary, six input contexts, the held-out real target (GT), and outputs from the selected generators named above the columns. A red border and **A** identify the anchor when one is required.

G LIMITATIONS

MindView uses one real target image per query to maintain a simple, uniform evaluation protocol across all seven tasks. This is a trade-off for ROG and MCVI, whose semantic constraints may admit nearby valid viewpoints. We reduce ambiguity through object-reference and visibility checks, bounded view clustering, checks for competing viewpoints, and LLM-assisted and independent expert review (Appendix A). For example, the recorded matching sets of the 243 released MCVI samples have median horizontal-position and yaw diameters of only $1.7\,\text{cm}$ and $0.93°$. Feature similarity, graded VLM ratings, and linearly decreasing geometric scores allow partial credit for modest deviations (Appendix B), so valid views with similar content and layout within the same compact cluster are expected to receive similar scores. Taken together, these ambiguity controls and graded metrics support using a single representative GT as a reasonable practical compromise for consistent evaluation across tasks.