Exploring Spatial Intelligence from a Generative Perspective

Supplementary Material

Appendix Overview

This supplementary material is organized as follows:

- **Section 1: Detailed Spatial Operation Definitions**
  Formal mathematical descriptions of the seven spatial operation categories used in GSI-Bench, including camera-relative translation, object-relative placement, rotation, receptacle placement, agent-camera control, spatial removal, and scaling.
- **Section 2: Evaluation Protocol**
  Definitions of all automatic evaluation metrics, including Instruction Compliance, Spatial Accuracy, Edit Locality, and Appearance Consistency, along with details on the 3D/2D detection pipeline and quality gating.
- **Section 3: Finetuning Details**
  Full training configurations for BAGEL and ablations such as classifier-free guidance (CFG).
- **Section 4: Human Study for Evaluating Spatial Consistency**
  Description of the human evaluation setup, mapping of metric scores (0–100 scale, normalized for comparison) to human-aligned discrete labels, and agreement analysis between model metrics and human judgments.
- **Section 5: Quantitative Analysis**
  Comprehensive comparison across open- and closed-source models on GSI-Bench, cross-domain generalization (GSI-Syn-Bathroom, GSI-Outdoor), additional baseline (Lumina-DiMOO), and per-operation performance breakdown.
- **Section 6: Qualitative Analysis**
  Additional visualizations showcasing model outputs, analysis of common failure modes, and qualitative comparisons across models.
- **Section 7: Limitations and Future Work**
  Discussion of current limitations in spatial reasoning, challenges in 3D geometric manipulation, sim-to-real generalization, data-centric considerations, and multi-turn instruction following.

1. Detailed Spatial Operation Definitions

This appendix provides formal mathematical definitions for the seven spatial operation categories used in GSI-Bench. Each operation is built upon a 3D scene representation $\mathcal{S} = \{\mathcal{O}_i\}_{i=1}^{N} \cup \{\mathcal{C}\}$, where $\mathcal{O}_i = (\mathbf{c}_i, \mathbf{s}_i, \mathbf{R}_i)$ represents the i-th object with center position $\mathbf{c}_i \in \mathbb{R}^3$, size $\mathbf{s}_i \in \mathbb{R}^3$, and rotation $\mathbf{R}_i \in SO(3)$, and $\mathcal{C} = (\mathbf{R}_c, \mathbf{t}_c, K)$ represents the camera parameters.

Each spatial instruction is structured as $\mathcal{T} = \langle \mathcal{R}, \mathcal{A}, \Phi_{\text{3D}} \rangle$, where $\mathcal{R}$ identifies the target object(s), $\mathcal{A}$ specifies the action, and $\Phi_{\text{3D}} : \mathcal{S}_{\text{src}} \rightarrow \mathcal{S}_{\text{dst}}$ defines the geometric transformation. The operations described below are quantitatively parameterized and physically executable within simulation, enabling unambiguous evaluation of generative spatial intelligence. This differs from open-world segmentation benchmarks that emphasize category-agnostic masks and prompts [9], where spatial relations are evaluated under recognition rather than executable 3D editing objectives.

(1) Camera-Relative Translation. The target object is translated along the camera's local coordinate axes:

$$\mathbf{c}_i' = \mathbf{c}_i + \Delta_{\text{cam}}, \quad \Delta_{\text{cam}} \in \{\text{left}, \text{right}, \text{forward}, \text{back}\}.$$

Example: "Move the apple 15 cm left relative to the camera." This requires the object to be visible and movable, with metric displacements consistent with the camera orientation.

(2) Object-Relative Placement. The target object is positioned relative to a reference (anchor) object:

$$\mathbf{c}'_{\text{target}} = \mathbf{c}_{\text{anchor}} + \Delta_{\text{rel}},$$
$$\Delta_{\text{rel}} \in \{\text{left of}, \text{right of}, \text{in front of}, \text{behind}\}.$$

Example: "Place the cup to the left of the plate, about 20 cm." This operation enforces pairwise spatial relations and includes collision and container-boundary checking.

(3) Object Rotation. An object is rotated about one or more axes:

$$\mathbf{R}_i' = \mathbf{R}_y(\theta_y)\, \mathbf{R}_x(\theta_x)\, \mathbf{R}_z(\theta_z)\, \mathbf{R}_i,$$

e.g., "Rotate the book 180° around the vertical axis." Rotation magnitude is discretized and constrained by object affordances and visibility.

(4) Receptacle Placement. An object is placed at a specified extremity or central location within a receptacle:

$$\mathbf{c}_i' = f_{\text{rec}}(\mathcal{O}_{\text{container}}, p),$$
$$p \in \{\text{center}, \text{leftmost}, \text{rightmost}, \text{nearest}, \text{farthest}\}.$$

Example: "Place the mug at the leftmost position on the shelf." This requires geometric partitioning of the receptacle and collision-free placement.

(5) Agent-Camera Control. The embodied agent's viewpoint is updated via discrete movement or rotation:

$$(\mathbf{R}'_c, \mathbf{t}'_c) = \Phi_{\text{agent}}(\mathbf{R}_c, \mathbf{t}_c),$$

where actions include `MoveAhead/Back/Left/Right`, `RotateLeft/Right`, and `LookUp/Down`. Example: "Rotate left 90°" or "Look down 30°." These changes alter the observed geometry and subsequent reasoning context.

(6) Spatial Removal. An object is deleted based on a spatial ranking criterion:

$$\mathcal{S}' = \mathcal{S} \setminus \{\mathcal{O}_k \mid \rho(\mathcal{O}_k) = \min_i \rho(\mathcal{O}_i)\},$$

where ρ ranks objects by attributes such as depth, height, or horizontal position. Example: "Remove the apple at the highest position." Ambiguity thresholds (e.g., 0.05 m) and visibility constraints ensure consistent selection.

(7) Object Scaling. An object's size is scaled while keeping its center fixed:

$$\mathbf{s}'_i = \gamma \cdot \mathbf{s}_i, \quad \gamma \in \{0.5, 0.75, 1.5, 2.0\}.$$

Example: "Enlarge the chair to twice its original size." Scaling updates both the `scale_factor` and `size` attributes, potentially affecting occlusion, collision, and spatial balance.

Together, these seven categories encompass a comprehensive spectrum of 3D manipulations— from egocentric camera motion to object-centric transformations and scene-level adjustments— providing a structured foundation for evaluating generative spatial intelligence under quantitative, controllable, and physically grounded conditions.

2. Evaluation Protocol

This section provides detailed definitions of the evaluation metrics used to assess generative spatial intelligence in GSI-Bench. Our multi-faceted protocol measures both geometric accuracy of spatial edits and the quality of generated imagery across four core metrics.

Final metric specification (canonical). All tables in the main paper and this appendix report scores on a **0–100 scale** (higher is better). **Instruction Compliance (IC)** is a **binary** metric (success/fail per sample); we report success rate. **Edit Locality (EL)** is computed as $100(1-\text{LPIPS})$ on **masked** non-target regions (projected 3D bounding box excludes the edited object); we additionally compute **masked SSIM** [17] on the same regions for quality gating. A **dataset-specific locality gate** applies before IC/SA scoring: on **GSI-Real**, a sample passes if SSIM $>$ 0.6 *or* LPIPS $<$ 0.3; on **GSI-Syn** subsets, a sample passes only if SSIM $>$ 0.6 *and* LPIPS $<$ 0.3. Samples that fail the gate receive zero for IC and SA; samples with zero IC also receive zero EL to avoid trivial high locality from no-edit outputs. **Spatial Accuracy (SA)** and **Appearance Consistency (AC)** are defined below; SA is set to zero when the locality gate fails.

For automatic evaluation, we employ DetAny3D [14], an open-vocabulary 3D object grounding model, to extract 3D bounding boxes and rotation matrices from edited images. This enables quantitative assessment of spatial transformations without manual annotation. The evaluation framework consists of four complementary metrics that target distinct aspects of the spatial editing task.

Instruction Compliance. This metric evaluates whether the edited scene satisfies the spatial operation specified in the textual instruction. For different operation types, we define success criteria as follows.

For **move** operations, we first verify that the target object is successfully detected at a new position distinct from all objects in the original image. We then check whether the predicted displacement magnitude d_{pred} is non-trivial ($d_{\text{pred}} \geq 0.05$ m) to filter out cases where the object barely moved. We compute a composite score combining 3D IoU and displacement accuracy: $\text{score}_{\text{move}} = 0.3 \cdot \text{IoU}_{\text{3D}} + 0.7 \cdot \exp(-\log^2(\text{ratio}_{\text{dist}}))$. Additionally, we verify object count consistency between the original and edited images; if extra objects appear or disappear, the composite score is halved. The operation is deemed successful if the score exceeds 0.1 when object counts match, or 0.05 otherwise.

For **rotate** operations, we check that the rotation matrix of the target object has changed from its original state (i.e., the predicted and original rotation matrices are not identical within numerical precision). The target object is first localized by finding the detection closest to the ground-truth destination within a 0.5 m tolerance.

For **remove** operations, an edit is deemed successful if (1) no object with 3D IoU > 0.1 relative to the original target bounding box is detected in the edited image, and (2) no other non-target objects from the original scene are incorrectly removed. We penalize the score by 0.3 for each incorrectly removed object, ensuring the final score is ≥ 0.5 for success.

For **resize** operations, we verify that the predicted scale change $\|\mathbf{s}_{\text{pred}} - \mathbf{1}\| > 0.1$, where $\mathbf{s}_{\text{pred}}$ is the ratio of the edited to original bounding box dimensions.

We report the overall **binary** success rate (pass/fail per sample) across all samples and per-operation success rates. The per-sample outcome is binary: the operation either satisfies the criteria above (success) or not (failure). Moreover, to enhance evaluation robustness, we incorporate 2D object detection as auxiliary verification when computing In-

struction Compliance and Spatial Accuracy. This addresses potential false positives or missed detections from the 3D grounding model, ensuring more reliable assessment even when 3D predictions are uncertain.

Spatial Accuracy. Beyond binary compliance, we measure the fine-grained geometric precision of spatial edits using operation-specific scoring functions. All scores are set to zero if the locality quality gate is not satisfied.

For **move** operations, we compute a weighted composite score: $\text{score}_{\text{move}} = 0.3 \cdot \text{IoU}_{\text{3D}} + 0.7 \cdot \exp(-\log^2(\text{ratio}_{\text{dist}}))$, where $\text{ratio}_{\text{dist}} = d_{\text{pred}}/d_{\text{gt}}$ measures the ratio of predicted to ground-truth displacement magnitudes, and the 3D IoU is computed only over detections with the correct object label. If the number of detected objects differs from the original image, the score is multiplied by 0.5 to reflect reduced reliability. When selecting the predicted bounding box, we choose the detection labeled as the target object that is closest to the ground-truth destination.

For **rotate** operations, we locate the target object by finding the detection with the correct label that is closest (within 0.5 m) to the ground-truth destination. We then compute the geodesic distance on SO(3) for both the predicted rotation $\theta_{\text{pred}} = \arccos\left(\frac{\text{tr}(\mathbf{R}_{\text{origin}}^{\top}\mathbf{R}_{\text{pred}})-1}{2}\right)$ and the ground-truth rotation $\theta_{\text{gt}} = \arccos\left(\frac{\text{tr}(\mathbf{R}_{\text{origin}}^{\top}\mathbf{R}_{\text{gt}})-1}{2}\right)$ relative to the original orientation, as well as the error between them: $\theta_{\text{diff}} = \arccos\left(\frac{\text{tr}(\mathbf{R}_{\text{gt}}^{\top}\mathbf{R}_{\text{pred}})-1}{2}\right)$. The final score combines angular magnitude matching with absolute accuracy: $\text{score}_{\text{rot}} = 0.5 \cdot \exp(-\log^2(\theta_{\text{pred}}/\theta_{\text{gt}})) + 0.5 \cdot \exp(-|\theta_{\text{diff}}|)$.

For **resize** operations, we identify the predicted bounding box as the detection closest to the ground-truth destination. We then measure scale accuracy using: $\text{score}_{\text{resize}} = \exp(-\text{mean}(\log^2(\mathbf{s}_{\text{pred}}/\mathbf{s}_{\text{gt}})))$, where $\mathbf{s}_{\text{pred}}$ and $\mathbf{s}_{\text{gt}}$ are the predicted and ground-truth scale factors along each dimension.

For **remove** operations, we assign a score of 1.0 if the target object is successfully removed (3D IoU ≤ 0.1 with original bounding box), then subtract 0.3 for each non-target object that was incorrectly removed, with a minimum score of 0.0.

We report the mean score per operation type and the overall mean score across all samples.

Edit Locality. A hallmark of spatial intelligence is the ability to perform localized edits that preserve unaffected regions. We assess edit locality by computing **masked LPIPS** [11] and **masked SSIM** [17] between the original and edited images, where the mask excludes the target object region. Specifically, we project the 3D bounding box of the target object (both in its original and, if applicable, destination pose) onto the image plane using the camera intrinsics, then mask out these regions before computing both metrics. Lower LPIPS (and higher SSIM) indicate better preservation of background content. We report the **Edit Locality (EL)** score as $100(1 - \text{LPIPS})$ on a 0–100 scale (higher is better). Critically, the same masked statistics control a **dataset-specific** quality gate before IC/SA are scored: on **GSI-Real**, a sample fails only if $\text{SSIM} \leq 0.6$ *and* $\text{LPIPS} \geq 0.3$; on **GSI-Syn**, a sample fails if $\text{SSIM} \leq 0.6$ *or* $\text{LPIPS} \geq 0.3$. Failing samples receive zero scores for Instruction Compliance and Spatial Accuracy, ensuring that models cannot achieve high performance through large-scale image regeneration.

Together, these metrics provide a holistic view of a model's spatial reasoning capabilities. For GSI-Syn, all metrics are computed automatically using ground-truth 3D annotations. For GSI-Real, we employ the same pipeline with DetAny3D for 3D reconstruction. We apply viewpoint consistency checks for rotation operations, filtering samples with significant camera motion. Excluding the edited object before computing perceptual scores is conceptually related to masking strategies explored in visual pre-training [8], except that our masks are obtained from projected 3D boxes rather than from a learned tokenizer.

Edit locality also serves as a bidirectional quality gate: (1) samples failing the dataset-specific masked SSIM/LPIPS criteria above cause IC and SA to be set to zero, and (2) samples with zero Instruction Compliance receive zero Edit Locality score to prevent models from trivially achieving high locality by avoiding edits altogether.

3. Finetuning Details

3.1. Training Settings

We fine-tune BAGEL-7B [1] on 8 NVIDIA GPUs for 10,000 steps using AdamW optimizer with learning rate 2×10^{-5}, batch size 1 per GPU, and gradient accumulation over 4 steps. We set maximum latent size to 64 and limit sequence length to 10,240 tokens per sample, with batch token usage capped at 11,520. The training data consists of 10,500 editing samples from GSI-Syn-Train (1,500 per operation type per environment), enabling the model to learn spatially grounded image editing through direct generation objectives. Checkpoints are saved every 1,000 steps, and we initialize from pre-trained BAGEL weights with distributed data parallel training via PyTorch.

3.2. Impact of Classifier-Free Guidance

To investigate the impact of Classifier-Free Guidance (CFG) during fine-tuning, we compare two training configurations: one with CFG enabled and one without. CFG is a widely adopted technique in diffusion-based generation

Table 1. **Impact of CFG on spatial understanding.** Accuracy (%).

Model	SAT-Real	OmniSpatial
BAGEL (step 10k, with CFG)	65.33	35.88
BAGEL (step 10k, w/o CFG)	**69.33**	**42.07**
Improvement	+4.00	+6.19

that improves sample quality by interpolating between conditional and unconditional predictions. Prior work on diffusion for dense prediction and few-shot segmentation likewise stresses how conditioning shapes locality and semantic fidelity [15, 18]. However, for unified MLLMs that jointly optimize perception and generation objectives, the unconditional generation component may interfere with instruction-conditioned spatial reasoning.

Table 1 shows the results on two spatial understanding benchmarks. Remarkably, training without CFG yields substantial improvements: +4.0% on SAT-Real [10] and +6.19% on OmniSpatial. This suggests that requiring the model to perform unconditional generation (as necessitated by CFG) disrupts its ability to ground spatial instructions accurately. During unconditional generation, the model must ignore textual instructions, which may degrade the tight coupling between language and spatial reasoning needed for instruction-following tasks. Since our primary goal is to enhance spatially grounded image editing—where precise instruction adherence is critical—we adopt the CFG-free training strategy for all subsequent experiments.

4. Human Study for Evaluating Spatial Consistency

To validate the reliability of our automatic evaluation metrics, we conducted a human study covering three perceptual dimensions of spatial correctness: **Instruction Compliance (IC)**, **Edit Locality (EL)**, and **Appearance Consistency (AC)**. For each sample, human raters were shown the original image and the predicted image side-by-side, and were asked to choose one of three categorical ratings for each dimension: *(1) completely consistent*, *(2) partially consistent*, or *(3) not at all consistent*. This yields a 3-level ordinal annotation for every dimension.

Mapping Automatic Metrics to Human-Aligned Categories. All tables in the main paper and this appendix report metrics on a **0–100 scale** (higher is better). For the human study comparison, we first **normalize** automatic scores to $[0, 1]$ (divide by 100), then discretize into the same three categories used in the human study. Specifically, for the normalized score $s \in [0, 1]$, we define:

$$\begin{aligned} &\text{completely consistent} \quad \text{if } s > 0.7, \\ &\text{partially consistent} \quad \text{if } s > 0.1, \\ &\text{not at all consistent} \quad \text{otherwise.} \end{aligned}$$

These thresholds reflect perceptually meaningful separation and enable a direct label-to-label comparison with human annotations.

Human–Metric Agreement. After discretization, we compute categorical agreement between human ratings and automatic predictions using sample-level classification accuracy (ACC). Table 2 summarizes the agreement for each dimension. The high consistency values demonstrate that the proposed metrics capture human-perceived spatial correctness effectively.

Table 2. Human–metric agreement (ACC) between discretized automatic predictions and human 3-class annotations. Higher is better.

Dimension	IC	EL	AC
ACC	0.7167	0.8056	0.8444

Summary. These results confirm that our evaluation pipeline provides a **robust and human-aligned** quantitative assessment of spatial reasoning. By normalizing automatic scores (0–100) to $[0, 1]$ for discretization and mapping to the same three-level scale used in the human study, we demonstrate strong agreement across all dimensions. Our automatic metrics thus serve as an effective proxy for human judgment and enable scalable evaluation without requiring manual inspection.

5. Quantitative Analysis

5.1. Analysis of Cross-View and Cross-Category Generalization

To strictly evaluate the model's capability to generalize across unseen scene categories and randomized camera viewpoints, we constructed the *GSI-Syn-Bathroom* dataset. As detailed in the main text, this dataset consists of bathroom scenes that are topologically distinct from the training environments (i.e., living rooms and tabletops) and features randomized camera poses to test robustness against viewpoint shifts. In-context adaptation for segmentation [7] suggests that a handful of exemplars can steer models under distribution shift; here we instead rely on fine-tuning on a larger synthetic corpus with fixed evaluation.

Table 3. **Performance comparison on the proposed GSI-Bench across the overall datasets** and four spatial reasoning dimensions: Instruction Compliance (IC), Spatial Accuracy (SA), Appearance Consistency (AC), and Edit Locality (EL). Higher is better.

	Evaluation Dimension	Closed-Source Models		Open-Source Models								
		Nano Banana	GPT img	Anyedit	Uniworld	Ultra	Qwen	Omnigen2	Emu3.5	BAGEL	BAGEL+GSI-Syn	$\Delta \uparrow$
GSI-real	IC	38.78	41.72	10.20	28.80	10.66	51.02	33.56	**51.70**	31.97	40.14	+8.16
	SA	21.60	28.04	8.37	18.36	5.70	**31.22**	19.62	29.51	22.07	27.76	+5.68
	AC	38.78	41.52	9.68	28.75	9.48	50.95	33.20	**51.70**	31.88	40.14	+8.25
	EL	34.92	27.52	8.75	18.51	8.97	40.55	29.82	**41.17**	27.89	37.11	+9.22
	Avg	33.52	34.70	9.25	23.61	8.70	43.44	29.05	**43.52**	28.46	36.28	+7.83
GSI-syn-table	IC	36.62	39.33	10.33	15.83	2.17	27.33	0.00	39.17	27.17	**50.67**	+23.50
	SA	38.96	26.16	22.84	30.33	3.09	25.52	0.00	24.09	26.52	**44.10**	+17.58
	AC	36.62	38.40	10.33	15.58	1.33	27.27	0.00	38.82	26.52	**50.67**	+24.15
	EL	35.91	31.98	9.52	14.43	1.93	25.51	0.00	34.91	26.17	**49.52**	+23.36
	Avg	37.03	33.97	13.26	19.04	2.13	26.41	0.00	34.25	26.59	**48.74**	+22.15
GSI-syn-bath	IC	**27.50**	22.96	6.00	12.00	1.50	14.00	0.00	22.00	12.50	26.00	+13.50
	SA	18.95	14.57	7.66	13.81	2.42	11.55	0.00	11.28	13.51	**23.66**	+10.16
	AC	**27.50**	22.76	6.00	12.00	1.15	14.00	0.00	21.80	12.50	26.00	+13.50
	EL	**26.95**	19.06	5.48	10.61	1.29	13.17	0.00	20.21	12.20	25.36	+13.16
	Avg	25.23	19.84	6.29	12.11	1.59	13.18	0.00	18.82	12.68	**25.25**	+12.58
GSI-syn-room	IC	20.65	8.05	7.00	12.69	2.20	20.40	18.71	20.70	16.11	**24.01**	+7.90
	SA	16.85	8.05	6.46	11.55	2.21	17.73	15.03	16.56	14.53	**19.41**	+4.88
	AC	28.01	16.69	11.85	20.40	3.46	28.67	25.94	26.98	24.00	**31.64**	+7.64
	EL	19.65	7.34	5.50	11.03	1.86	18.48	17.13	17.56	14.82	**22.61**	+7.79
	Avg	21.29	10.03	7.70	13.92	2.43	21.32	19.20	20.45	17.37	**24.42**	+7.05

Table 3 (row *GSI-syn-bath*) presents the quantitative results for this challenging setting. We observe the following key findings:

- **Effectiveness of GSI-Syn Fine-tuning:** Comparing **BAGEL** with **BAGEL+GSI-Syn**, the fine-tuning process yields substantial improvements across all four spatial reasoning dimensions. Notably, the Appearance Consistency (AC) score increases by $+$**13.50** (from 12.50 to 26.00), and the Edit Locality (EL) improves by $+$**13.16**. This indicates that the spatial priors learned from the training set (rooms and tabletops) effectively transfer to the unseen bathroom category, enabling the model to perform precise edits without degrading the scene's visual identity.
- **Superior Spatial Accuracy:** In terms of Spatial Accuracy (SA), which measures the precision of spatial operations, **BAGEL+GSI-Syn** achieves a score of **23.66**. This performance not only significantly surpasses the base model (+10.16) but also outperforms closed-source **GPT img** (14.57) and open-source **Emu3.5** (11.28). This suggests that our model has learned robust geometric representations that are invariant to scene category changes.
- **Robustness to Viewpoint Changes:** Despite the *GSI-Syn-Bathroom* dataset introducing randomized cross-view configurations not present in the standard training curriculum, our model maintains a balanced performance with an average score increase of $\Delta \uparrow$ **12.58**. This demonstrates that the model does not merely memorize canonical views but acquires a generalized understanding of 3D spatial relationships.

In conclusion, the performance on *GSI-Syn-Bathroom* provides strong empirical evidence that **BAGEL+GSI-Syn** possesses strong out-of-distribution generalization capabilities, effectively handling both novel semantic categories and diverse spatial perspectives.

5.2. Generalization to Outdoor Scenes (GSI-Outdoor)

To assess whether our findings generalize beyond indoor scenes, we collected an additional **31 outdoor samples** during the rebuttal phase, following the same data generation and evaluation pipeline as GSI-Bench. We refer to this subset as *GSI-Outdoor*; source images are drawn from existing open-source datasets (e.g., DL3DV). Tools for matching and one-shot segmentation [6] motivate studying heterogeneous appearance under fixed evaluation protocols; Table 4 reports results. Fine-tuning on GSI-Syn improves performance on outdoor scenes across AC, EL, IC, and average score, indicating that our conclusions extend to outdoor distributions.

Table 4. **Results on GSI-Outdoor.** BAGEL vs. BAGEL+GSI-Syn on 31 outdoor samples.

	AC	EL	IC	SA	Avg.
BAGEL	38.71	35.54	38.71	33.62	36.64
BAGEL+GSI-Syn	**41.61**	**37.86**	**41.94**	31.72	**38.28**

5.3. Additional Baseline: Lumina-DiMOO

To address the concern that our main findings rely solely on BAGEL, we include an additional baseline, **Lumina-DiMOO** [13], which uses a fundamentally different architecture (Discrete Diffusion LLM) and supports native unified training with an unfrozen backbone. We fine-tuned Lumina-DiMOO on GSI-Syn-train for one epoch under the same protocol as BAGEL. Table 5 summarizes the results.

Lumina-DiMOO exhibits **similar patterns to BAGEL**: strong sim-to-real transfer on GSI-Real (gains in IC, SA, AC, EL, and SAT-Real), and improved spatial understanding on OmniSpatial (+1.8% overall, with gains in Dynamic Reasoning, Spatial Interaction, and Perspective Taking). As with BAGEL, **Complex Logic** on OmniSpatial decreases after GSI-Syn fine-tuning; this is consistent with the training data lacking complex multi-step reasoning scenarios rather than a failure of the main conclusion. The fact that two architecturally distinct unified models (autoregressive vs. diffusion-based) show the same trends supports that our findings generalize beyond a single architecture.

Table 5. **Results on Lumina-DiMOO.** GSI-Real (left) and OmniSpatial (right). +GSI-Syn: fine-tuned on GSI-Syn.

GSI-Real	IC	SA	AC	EL	SAT-Real
Lumina-DiMOO	34.24	20.25	34.24	26.60	46.00
+GSI-Syn	**36.05**	**22.69**	**35.83**	**28.04**	**47.33**
OmniSpatial	Dynamic	Spatial	Logic	Persp.	Overall
Lumina-DiMOO	29.52	29.00	26.19	28.52	28.51
+GSI-Syn	**30.00**	**33.33**	24.60	**31.55**	**30.33**

5.4. Results for Each Operation

To better understand the model capabilities, we report the Spatial Accuracy (SA) scores decomposed by specific operation types in Table 6. The results reveal a distinct hierarchy in task difficulty and highlight the specific benefits of our fine-tuning strategy.

Table 6. **Breakdown of Spatial Accuracy (SA) across different editing operations.** BAGEL+ denotes our model fine-tuned on GSI-Syn.

Operation	**Nano Banana**	**GPT img**	**Emu3.5**	**BAGEL**	**BAGEL+**
Scale	71.09	45.27	49.28	62.05	**71.92**
Move	8.11	13.82	17.88	10.74	**19.16**
Rotate	**16.60**	7.09	3.42	1.60	10.07
Remove	41.51	27.42	51.91	**53.73**	48.10
View	55.21	46.27	54.89	48.62	**61.36**

Complexity Hierarchy. We observe a clear performance divide based on geometric complexity. Operations such as *Scale* and *Remove*, which can often be approximated by 2D affine transformations or in-painting, generally yield higher SA scores across all models (e.g., 71.92 on Scale and 48.10 on Remove for BAGEL+). In contrast, *Move* and *Rotate* impose stringent 3D consistency constraints, requiring the model to hallucinate unseen geometry and handle occlusions. Consequently, these operations result in significantly lower scores, identifying them as the primary bottleneck for current generative editing models.

Impact of GSI-Syn Fine-tuning. Comparing **BAGEL** with **BAGEL+**, the fine-tuning leads to consistent improvements across the most challenging dimensions.

- **Mastering Rotation:** The most notable improvement is in the *Rotate* operation. While strong baselines like **Emu3.5** struggle significantly with rotation (scoring only 3.42), **BAGEL+** achieves a score of **10.07**, nearly tripling the performance of Emu3.5 and surpassing GPT img (7.09). This suggests that our dataset effectively teaches the model 3D object orientation priors that are absent in general pre-training.
- **Precision in Placement:** For the *Move* operation, BAGEL+ improves from 10.74 to 19.16, exceeding the top-performing open-source baseline Emu3.5 (17.88).
- **Scaling Robustness:** In *Scale*, BAGEL+ (71.92) leads among compared models and substantially outperforms GPT img (45.27), demonstrating strong size control after fine-tuning.

Trade-offs. We note a slight regression in the *Remove* operation for BAGEL+ compared to the base model (53.73 $\rightarrow$ 48.10). This is likely because our fine-tuning curriculum (GSI-Syn) emphasizes object manipulation (adding, moving, rotating) to enforce geometric understanding, which may slightly deemphasize the "erasing" capability—a trade-off reminiscent of curricula that widen instance distributions with generative data in segmentation [2]. However, the performance remains robust and competitive.

In summary, while fundamental challenges remain in perfect 3D rotation, **BAGEL+** demonstrates that targeted training on synthetic spatial data significantly enhances the model's ability to perform complex, geometry-aware edits compared to general-purpose baselines.

6. Qualitative Analysis

6.1. More Visualization Examples

Figures 1 and 2 clearly show that BAGEL+ delivers more coherent and semantically faithful editing outcomes than its predecessor, BAGEL. The visual comparisons simultaneously highlight the strong generative and spatial reasoning abilities of GPT img and Emu3.5.

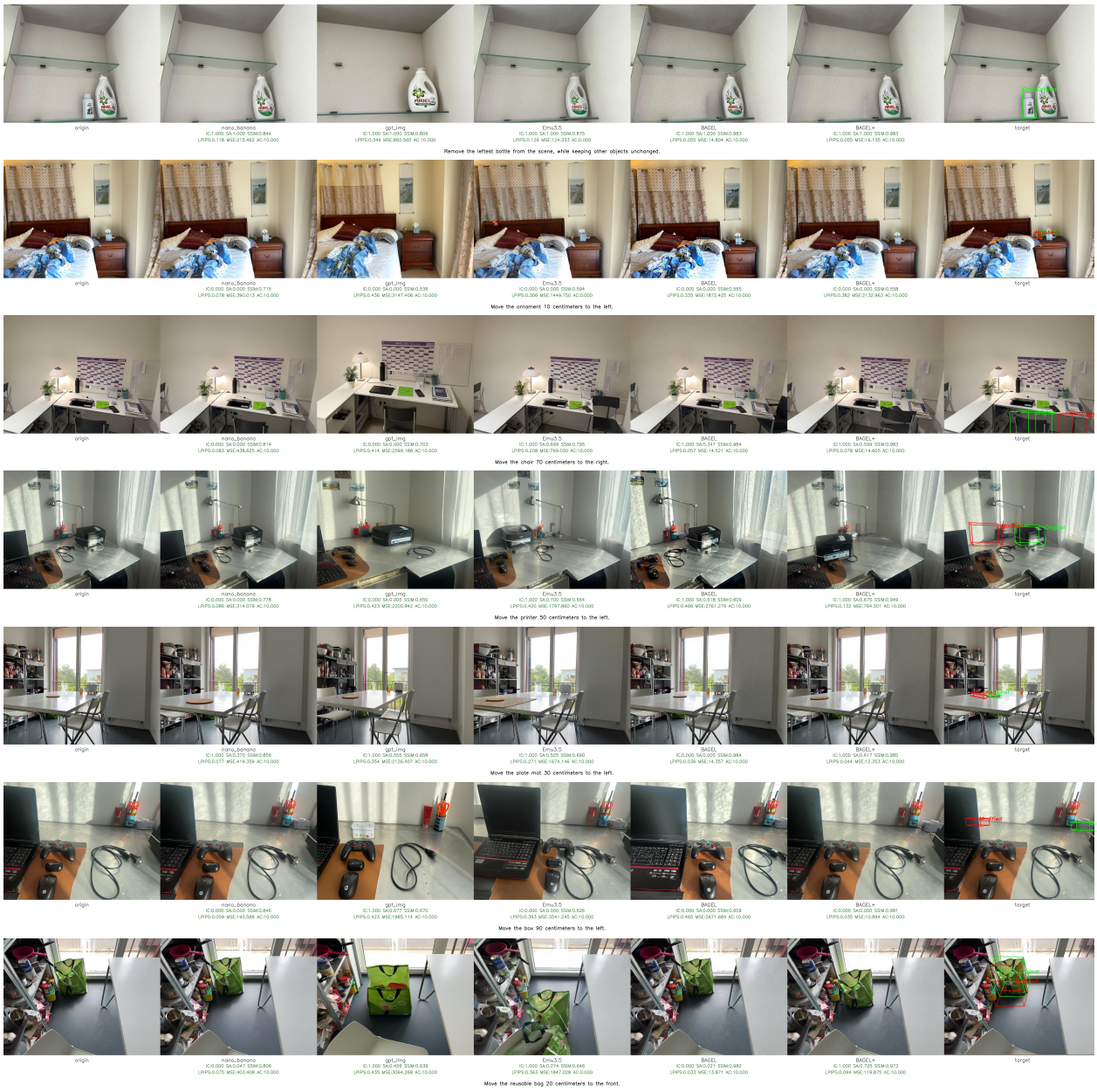


Figure 1. **Examples on GSI-Real.** Representative samples from the real-world subset of the GSI dataset.

6.2. Some Analysis on Failure Cases

As shown in Figure 3, current models still exhibit several consistent failure patterns across different instruction types. One major issue is insufficient editing strength, where the applied operation is too weak to meaningfully affect the target region. In these cases, objects meant to be removed, translated, or rotated remain nearly identical to the input, suggesting that the model either fails to fully parse the instruction or lacks confidence to modify the scene aggressively. A second frequent error arises when the model misinterprets object-level spatial transformations as camera motion. Instead of moving or rotating the specified object, the model synthesizes a global viewpoint change, producing results that resemble a shifted perspective rather than the intended geometric manipulation. This often indicates that the model entangles object geometry with scene-level structure.

Less commonly, the edited content suffers from appearance degradation, such as blurriness, loss of texture fi-

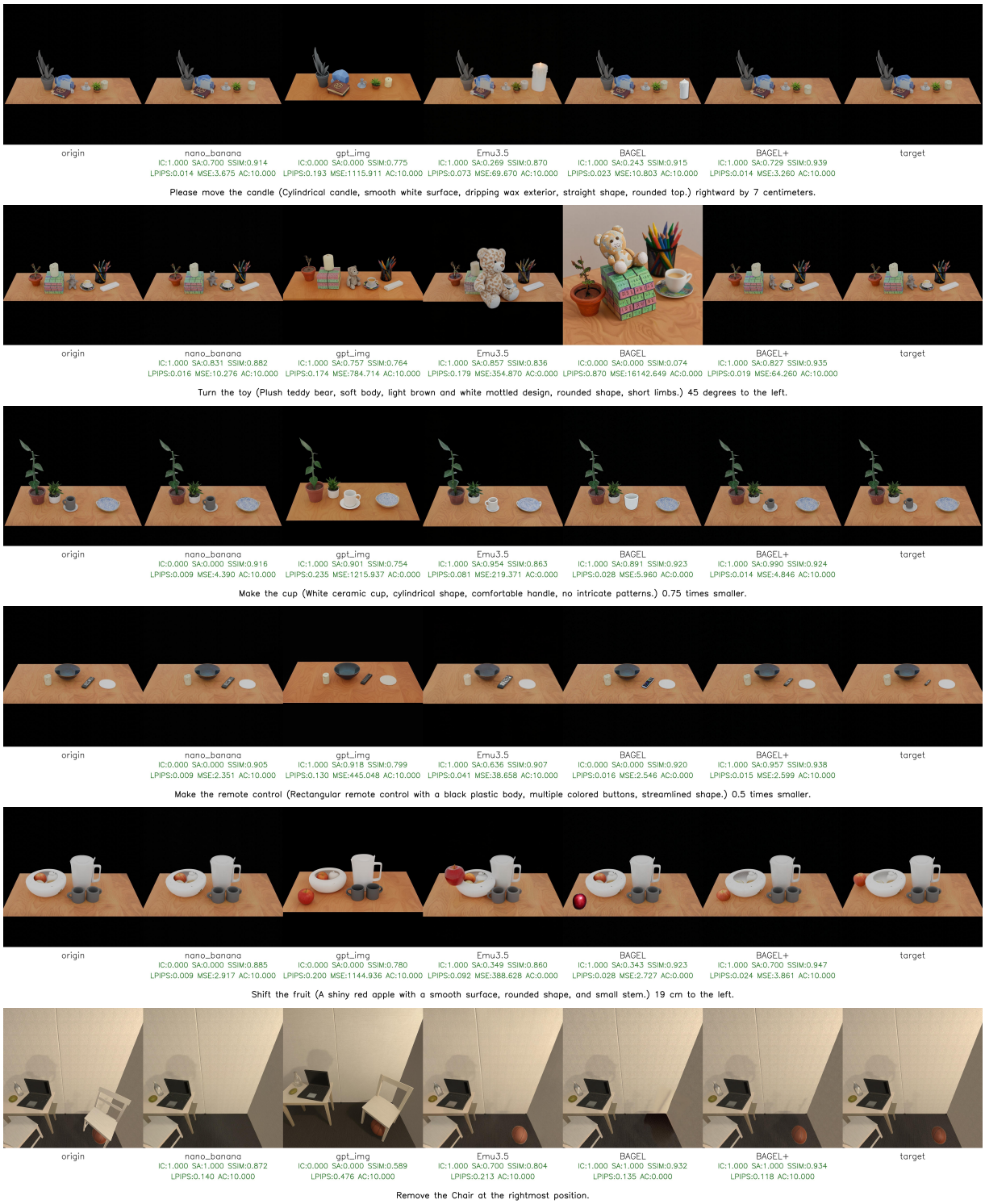


Figure 2. **Examples on GSI-Table and GSI-Room.** The figure illustrates representative editing results on two subsets of our GSI benchmark

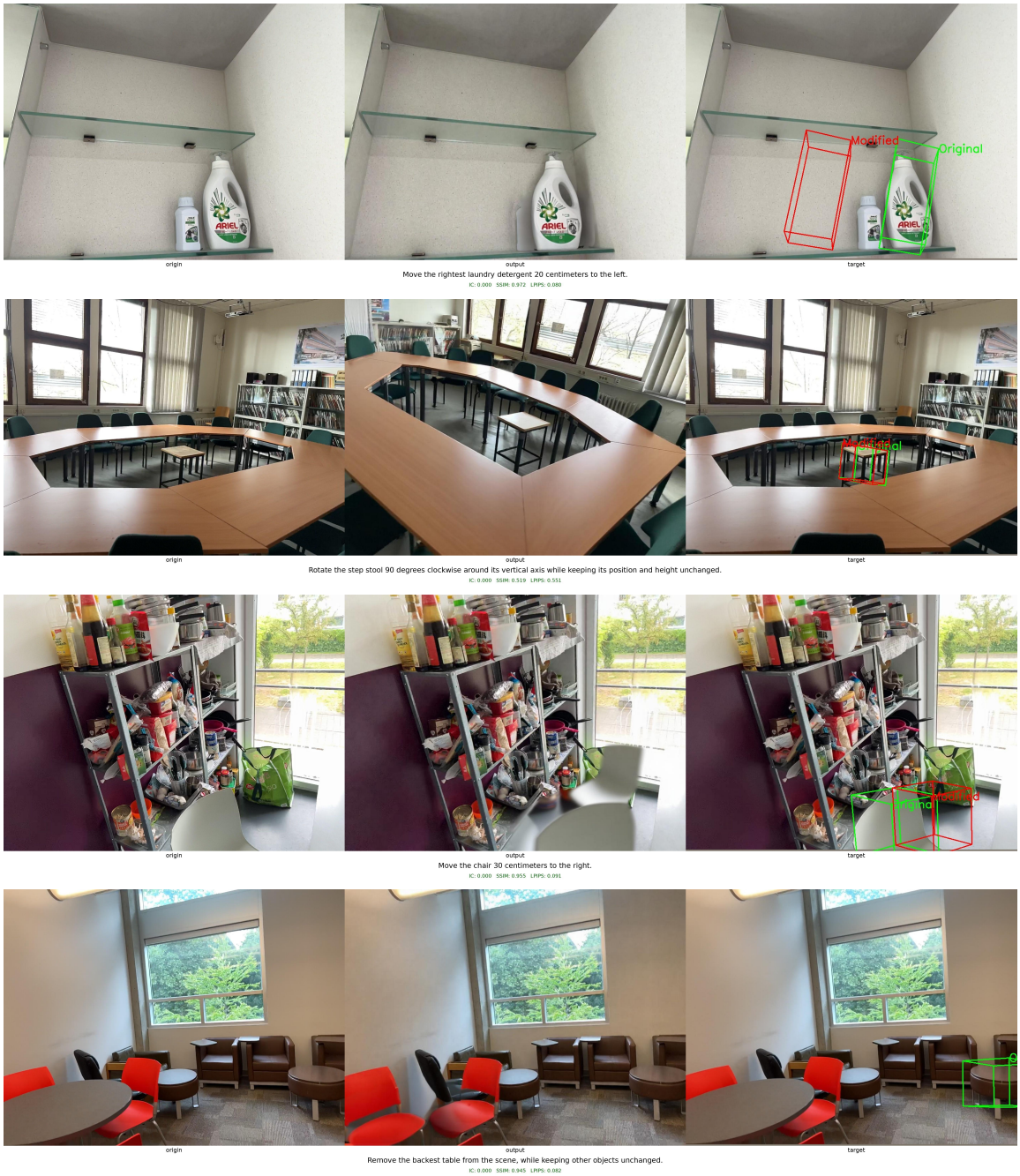
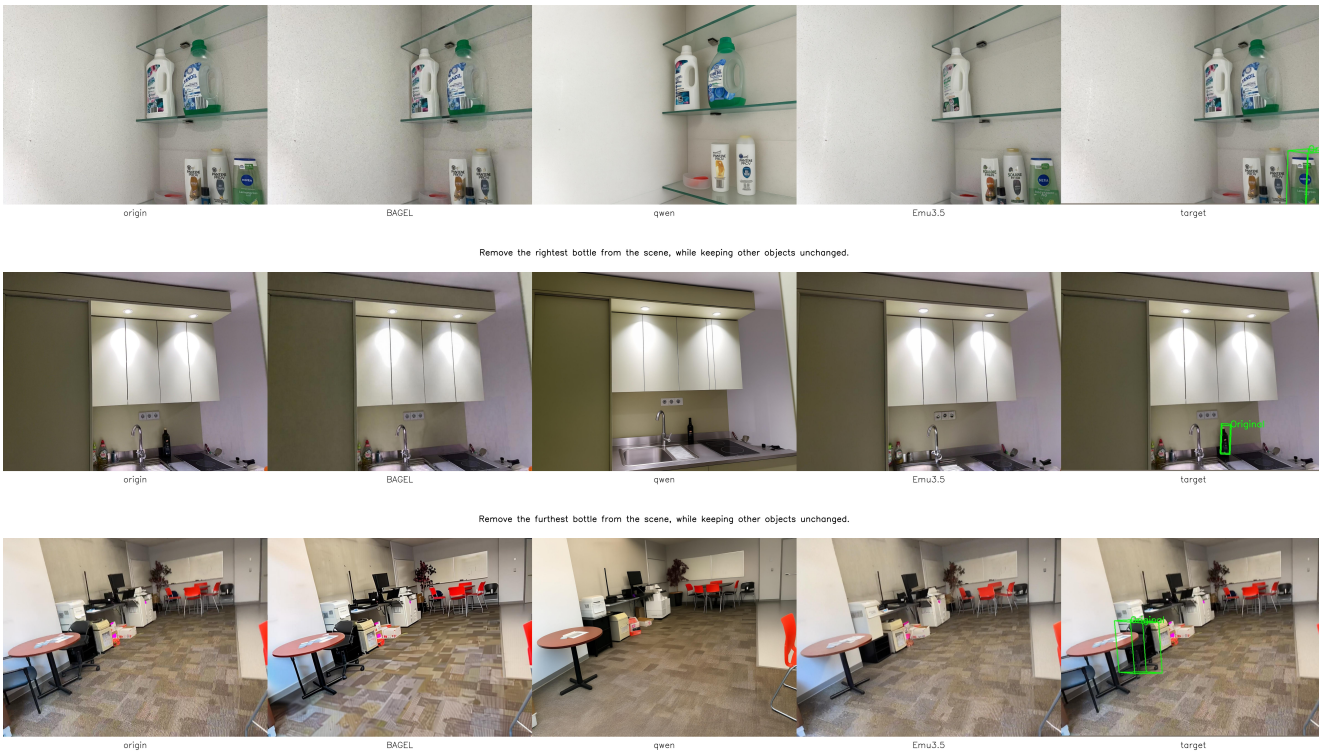


Figure 3. **Some Failure Examples**

delity, or inconsistent shading. These artifacts typically occur when the model attempts to hallucinate unseen object surfaces or reconstruct fine details after transformation. Another relatively rare but notable failure involves incorrect grounding of the target object. The model occasionally applies the edit to the wrong instance—either a spatially nearby object or one with semantically similar features—revealing limitations in referential understanding and precise localization, consistent with open challenges in pixel-level understanding for multimodal models [19].

Overall, the first two categories—weak or absent edits and confusion between object motion and camera mo-

(a) **Failure examples of AnyEdit and Ultra.** These cases show that both models may substantially distort the appearance of the edited object.



(b) **Failure examples of BAGEL, Qwen, and Emu3.5.** The models occasionally perform unintended removals, altering regions beyond the user-specified scope.

Figure 4. **Combined failure cases from multiple editing models.**

Move the bed 90 centimeters to the left.

Remove the physically highest office chair from the scene, while keeping other objects unchanged.

Move the tissue box 20 centimeters to the right.

Move the monitor 30 centimeters to the left.

Figure 5. **Additional failure cases across different models.** Examples include unintended viewpoint shifts, severe removal artifacts, and issues with content preservation.

tion—constitute the majority of observed failures. The latter two—appearance distortion and incorrect target selection—occur less frequently but highlight remaining challenges in reliable spatial manipulation and fine-grained grounding.

6.3. More Examples on Different Models

As illustrated in Fig. 4a, edits produced by AnyEdit and Ultra often **result in substantial changes to the object's appearance**. These qualitative results indicate that both models struggle to preserve essential visual characteristics, leading to noticeable deviations from the original content.

As shown in Fig. 4b, the results generated by BAGEL, Qwen, and Emu3.5 sometimes **exhibit unintended removal beyond the scope of the user instruction**. These instances suggest that these models face difficulties in strictly constraining edits locally, resulting in excessive or undesired content deletion.

We present additional examples across different models in Fig. 5. The first row illustrates that BAGEL may **inadvertently alter the camera viewpoint**. The second row demonstrates a case where BAGEL **introduces severe artifacts following removal**. The last two rows provide further evidence that GPT img **suffers from poor content preservation**.

7. Limitations and Future Work

While GSI-Bench and BAGEL represent a significant step towards spatially intelligent generative models, several limitations remain, pointing towards promising directions for future research.

Scope: Rigid vs. Non-Rigid Edits. The current GSI-Bench focuses on **rigid spatial edits**: move, rotate, scale, and camera motion. Deformable structure priors have been explored in domains where topology varies smoothly [5]; we do not cover non-rigid operations (e.g., folding, pouring, deforming) for three reasons. First, rigid edits are already highly challenging for current models and suffice to study generative spatial intelligence under explicit 3D constraints. Second, rigid transformations admit cleaner quantitative evaluation: they are uniquely defined given the instruction and scene state, so ground-truth targets and metrics can be standardized. Third, non-rigid edits typically require physics-based simulation, involve multiple plausible outcomes, and lack a single canonical ground truth, making fair benchmarking and automated evaluation difficult. Extending GSI-Bench to deformable objects and physical interactions (e.g., contact, gravity, compliance) is a natural and important direction for future work.

Challenges in 3D Geometric Transformations. As evidenced in Table 6, operations involving rigid body transformations—specifically *Move* and *Rotate*—remain significantly more challenging than 2D-approximated tasks like *Scale* or *Remove*. Although our fine-tuned model outperforms baselines (e.g., BAGEL+ reaches 10.07 on Rotate vs. Emu3.5's 3.42), the absolute accuracy scores for *Rotate* (10.07) and *Move* (19.16) indicate ample room for improvement. Unlike scaling or removal, these operations impose stringent requirements: the model must not only maintain the object's 3D structure under a new pose but also plausibly generate disoccluded background regions that were previously hidden. Physics-aware predictors in other domains show that even coarse physical priors can regularize ill-posed regression [4]; analogous inductive biases may eventually help 3D-consistent editing. Future work could explore incorporating explicit 3D control signals (e.g., depth or normal maps) or leveraging 3D-native diffusion backbones to better handle these complex spatial constraints.

Exploration of Unified Architectures. In this work, we primarily utilized BAGEL to validate our benchmark and training strategy. However, GSI-Bench serves as a broad evaluation protocol that can be used to investigate the impact of different architectural designs on spatial intelligence. Future work can leverage our suite to benchmark diverse unified model structures—examining how variations in visual encoders, LLM parameter sizes, and connector strategies influence a model's intrinsic capacity for spatial reasoning, alongside emerging omnimodal and reinforcement-style training recipes [16, 20].

Data-Centric Investigations. Our current fine-tuning utilizes a fixed distribution of synthetic data. A critical open question concerns the specific influence of different data sources (e.g., room-scale vs. tabletop) and operation types on downstream capabilities. We plan to conduct fine-grained ablation studies to quantify these data contributions. Denoising-based pre-training has been shown to accelerate structured discovery in scientific settings [12]; similar ideas might inform curricula that prioritize informative spatial trajectories. Furthermore, investigating advanced data filtering algorithms—to automatically identify samples with the highest "spatial instructional value"—could enable more data-efficient training curricula and reduce the reliance on large-scale datasets.

Sim-to-Real Gap. Our fine-tuning relies heavily on the synthetic GSI-Syn dataset. While our results on GSI-Real demonstrate strong generalization, a performance gap persists between synthetic and real-world scenarios. Real-world environments often present complex lighting effects (e.g., shadows, reflections), diverse textures, and cluttered

backgrounds that are difficult to simulate perfectly. Bridging this gap via unsupervised domain adaptation or collecting larger-scale real-world interaction data remains a critical next step.

Compound and Multi-turn Instructions. Currently, our evaluation focuses on atomic spatial operations to isolate specific capabilities. However, practical applications often require executing compound instructions (e.g., "Move the cup to the left *and* rotate it 90 degrees") or sequential multi-turn edits. Investigating how models maintain spatial memory and geometric consistency across long-horizon editing sessions is an important avenue for future exploration.

Deformable and Physical Interactions. As noted above, the present benchmark is limited to rigid edits. Future work may extend GSI-Bench to deformable object manipulation and physical interactions (e.g., cloth folding, liquid pouring, compliant contact), once standardized data generation and evaluation protocols for such settings become feasible. In other generative modeling communities, constrained diffusion with anchor or motif structure has been used to compose complex spatial arrangements [3]; analogous compositional control may eventually transfer to 3D-aware image editing.

References

[1] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 3

[2] Chengxiang Fan, Muzhi Zhu, Hao Chen, Yang Liu, Weijia Wu, Huaqi Zhang, and Chunhua Shen. Divergen: Improving instance segmentation by learning wider data distribution with more diverse generative data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3986–3995, 2024. 6

[3] Ke Liu, Weian Mao, Shuaike Shen, Xiaoran Jiao, Zheng Sun, Hao Chen, and Chunhua Shen. Floating anchor diffusion model for multi-motif scaffolding. *arXiv preprint arXiv:2406.03141*, 2024. 13

[4] Ke Liu, Hao Chen, and Chunhua Shen. Physics aware neural networks for unsupervised binding energy prediction. In *Forty-second International Conference on Machine Learning*, 2025. 12

[5] Ke Liu, Shangde Gao, Yichao Fu, and Shangqi Gao. Towards generalizable retina vessel segmentation with deformable graph priors. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 12

[6] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*, 2023. 5

[7] Yang Liu, Chenchen Jing, Hengtao Li, Muzhi Zhu, Hao Chen, Xinlong Wang, and Chunhua Shen. A simple image segmentation framework via in-context examples. *Advances in Neural Information Processing Systems*, 37:25095–25119, 2024. 4

[8] Yang Liu, Xinlong Wang, Muzhi Zhu, Yue Cao, Tiejun Huang, and Chunhua Shen. Masked channel modeling for bootstrapping visual pre-training. *International Journal of Computer Vision*, 133(2):760–780, 2025. 3

[9] Yang Liu, Yufei Yin, Chenchen Jing, Muzhi Zhu, Hao Chen, Yuling Xi, Bo Feng, Hao Wang, Shiyu Li, and Chunhua Shen. Unified open-world segmentation with multi-modal prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21557–21567, 2025. 1

[10] Arijit Ray, Jiafei Duan, Ellis Brown, 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. *arXiv preprint arXiv:2412.07755*, 2024. 4

[11] Zhang Richard, Isola Phillip, Alexei A. Efros, Shechtman Eli, and Wang Oliver. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[12] Shuaike Shen, Ke Liu, Muzhi Zhu, and Hao Chen. A denoising pre-training framework for accelerating novel material discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 28368–28376, 2025. 12

[13] Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025. 6

[14] Hanxue Zhang, Haoran Jiang, Qingsong Yao, Yanan Sun, Renrui Zhang, Hao Zhao, Hongyang Li, Hongzi Zhu, and Zetong Yang. Detect anything 3d in the wild. *arXiv preprint arXiv:2504.07958*, 2025. 2

[15] Canyu Zhao, Yanlong Sun, Mingyu Liu, Huanyi Zheng, Muzhi Zhu, Zhiyue Zhao, Hao Chen, Tong He, and Chunhua Shen. Diception: A generalist diffusion model for visual perceptual tasks. *arXiv preprint arXiv:2502.17157*, 2025. 4

[16] Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *arXiv preprint arXiv:2505.20256*, 2025. 12

[17] Wang Zhou, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 2, 3

[18] Muzhi Zhu, Yang Liu, Zekai Luo, Chenchen Jing, Hao Chen, Guangkai Xu, Xinlong Wang, and Chunhua Shen. Unleashing the potential of the diffusion model in few-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 37:42672–42695, 2024. 4

[19] Muzhi Zhu, Yuzhuo Tian, Hao Chen, Chunluan Zhou, Qingpei Guo, Yang Liu, Ming Yang, and Chunhua Shen. Segagent: Exploring pixel understanding capabilities in mllms by imitating human annotator trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3686–3696, 2025. 9

[20] Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Zheng Huang, Mingyu Liu, Hao Chen, Cheng Zou, Jingdong Chen, Ming Yang, et al. Active-o3: Empowering multimodal large language models with active perception via grpo. *arXiv preprint arXiv:2505.21457*, 2025. 12