MindBlock: Probing Spatial Assembly and Structure in Unified Multimodal Models

Anonymous CVPR submission

Paper ID

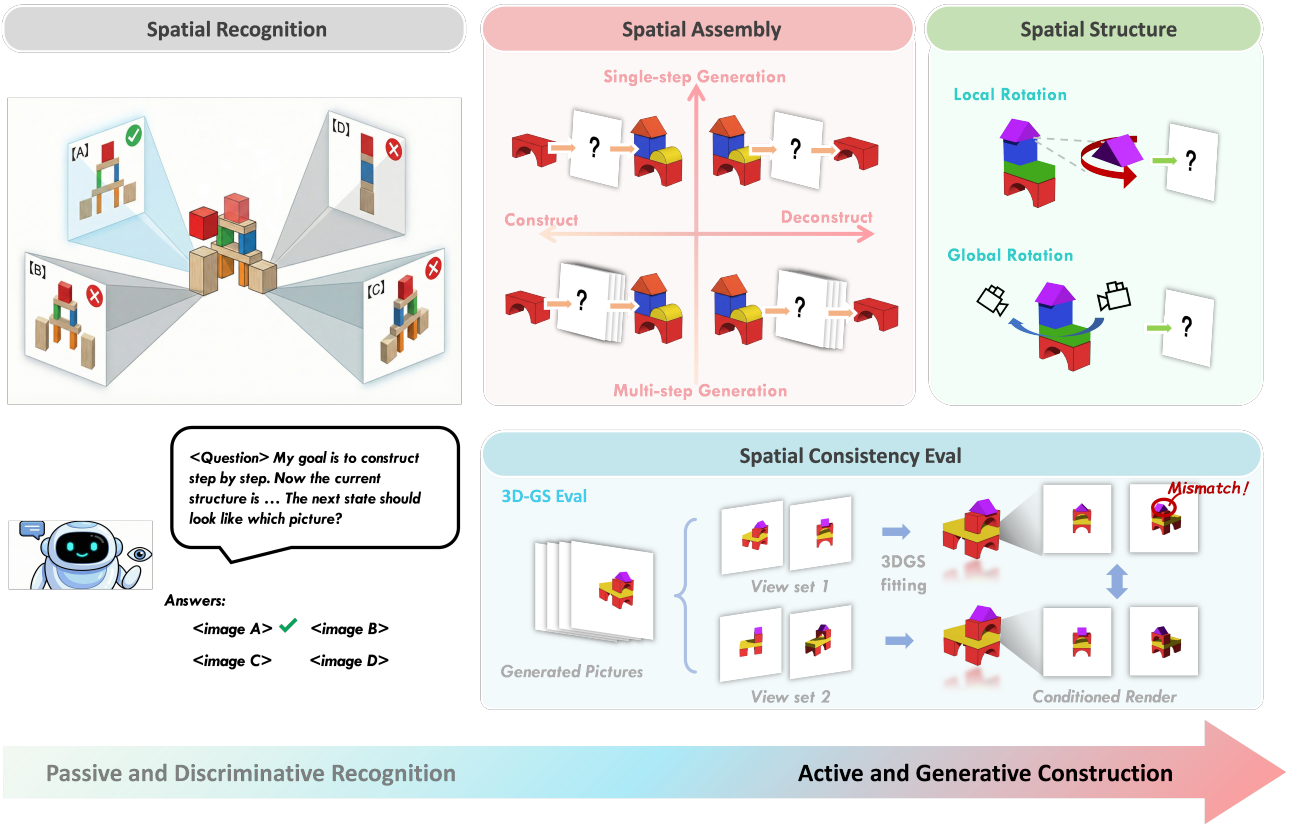

Figure 1. Children aged four to six begin constructing complex block structures from diagrams, demonstrate awareness of enclosure and decoration, and start to grasp three-dimensional spatial relationships. Can AI agents match this level of spatial intelligence? We transform visual-spatial reasoning from **passive selection (multi-choice question answering)** to **active construction (conditional image generation)**, rigorously examining its internal three-dimensional world modeling.

Abstract

While Unified Multimodal Models (UMMs) show remarkable reasoning capabilities, their spatial intelligence remains limited to passive 2D Question-Answering (QA). In this paper, we argue true spatial intelligence demands active construction: not only recognize a 3D structure in pixel space, but building and modifying it. We introduce ***MindBlock*** *to challenge models' active generative construction in pixel space across two primary axes:* ***Spatial Assembly*** *evaluating step-by-step compositional and causal reasoning, and* ***Spatial Structure*** *probing spatial equivariance through local sub-component rotation and global viewpoint transformation. To move beyond pixel-level metrics, we propose* ***3DGS-Eval****, a novel validation protocol using 3D Gaussian Splatting to reconstruct implicit scenes from model-generated multi-view images. This allows us to quantify structural consistency, verifying for the first*

time whether a model's generative output admits a coherent internal 3D world model. Furthermore, we conduct a deep-dive diagnostic analysis into the representational grounding of spatial logic, disentangling whether structural consistency relies on textual Chain-of-Thought (CoT) as a symbolic scaffold, or emerges as a native spatial intuition within the generative latent space. Our findings reveal a significant "perception-execution" gap: while current models correctly identify the intended spatial state yet fail in active construction. They struggle to maintain spatial equivariance without explicit symbolic scaffolding. MindBlock provides a rigorous foundation for the next generation of embodied, physically-grounded multimodal AI.

UMMs Spatial Intelligence Dataset and Benchmark

1. Introduction

The ability to assemble blocks into meaningful structures unfolds gradually during human development. Toddlers as young as two can stack multiple blocks, but purposeful construction typically emerges around age three. By ages four to five, children begin to replicate simple structures from observation and to grasp spatial relationships such as *next to* and *on top of* [8]. The skill of *spatial mental rotation* (imagining how a structure would look from a different viewpoint) develops around ages five to six [4]. This trajectory, from simple stacking to perspective-aware, multi-step assembly, raises a research question:

> *Can current AI models match the spatial intelligence of a six-year-old?*

Recent studies have begun to explore this question by evaluating and training foundation models in block-building tasks. `BrickGPT` [11] models the locations of LEGO bricks using spatial coordinates represented as text, employing large language models (**LLMs**) to generate physically stable structures. However, this approach operates entirely within text space, whereas humans build blocks using visual input. `PhyBlock` [10] evaluates 3D block assembly in vision-language models (**VLMs**) through visual question answering (VQA). Yet multiple-choice question answering can be hacked through statistical shortcuts: correct answers neither guarantee genuine understanding nor translate to robotic assembly. Notably, unified multimodal models (**UMMs**) [3, 15] have recently emerged to enable both multimodal inputs and outputs, typically in the form of images and text. This makes UMMs uniquely suited for perceiving visual instructions and generating step-by-step image sequences for planning [17].

In this work, we move beyond *passive, discriminative* VQA in spatial intelligence: rather than asking models to *select* the correct answer, we ask them to *build* the correct structure. Specifically, we introduce **MindBlock**, a benchmark that probes spatial intelligence through two complementary axes; see Fig. 1.

- **Spatial Assembly** tests whether models can actively construct block structures step by step. Given an image of a target structure, the model must generate an interleaved sequence of reasoning steps and intermediate images, each showing the scene after one block is placed. This demands *compositional reasoning* (assembling parts into a coherent whole) and *causal reasoning* (a block placed in mid-air will fall; a missing base block invalidates every step above it).
- **Spatial Structure** tests whether models maintain a consistent 3D understanding across viewpoints. Given the same structure, the model must generate what it looks like from a different angle, preserving every block's relative position and occlusion. We call this *spatial equivariance*: a model that truly understands a structure in 3D should produce a predictably transformed image when the viewpoint rotates.

Standard image-similarity metrics (*e.g.*, SSIM, LPIPS, CLIP-S) cannot capture a fundamental failure mode of modern Unified Multimodal Models (UMMs): a model may produce images that look plausible individually yet imply mutually inconsistent 3D structures. For instance, blocks may appear on different sides of the structure when rendered from different viewpoints, yielding images that are locally convincing but globally incompatible. To directly measure this phenomenon, we introduce **3DGS-Eval**. Given a set of model-generated multi-view images, we reconstruct a scene using 3D Gaussian Splatting (3DGS), a compact volumetric representation, and evaluate the geometric consistency of the reconstruction. If the generated views encode a coherent 3D structure, the reconstruction converges to a stable geometry; otherwise, contradictory views cause the volumetric fit to deteriorate, revealing hidden spatial inconsistencies that standard 2D metrics fail to detect.

Our evaluation experiments reveal that good performance under standard 2D metrics can be misleading. As shown in Table 1, several models achieve high perceptual similarity scores while producing multi-view generations that are geometrically inconsistent. Images that appear plausible in isolation often contradict each other when interpreted as a single 3D structure, exposing a fundamental *2D–3D consistency gap* in current UMMs.

To better understand the origin of these failures, we conduct a series of diagnostic analyses. First, we examine whether spatial reasoning benefits from explicit reasoning by comparing generations with and without Chain-of-Thought (CoT) explanations. CoT improves volumetric consistency under 3DGS-Eval, suggesting that language can act as a symbolic scaffold that stabilizes spatial rea-

soning during generation. We also extend the evaluation to multi-step construction tasks. While some models succeed in single-step edits, they often fail to maintain consistency across longer reasoning trajectories, exhibiting a phenomenon we term *spatial drift*, where small geometric errors accumulate over time.

Finally, by comparing generative and discriminative variants of the same tasks, we identify three distinct failure modes: perceptual failures (incorrectly interpreting the input structure), planning failures (producing an incorrect sequence of operations), and execution failures (failing to realize a correct plan in image generation). This analysis reveals a pronounced *perception–execution gap*: models frequently recognize correct spatial configurations more reliably than they can construct them.

Our contributions are three-fold:

- **New Tasks.** We evaluate spatial intelligence through *active construction*: models must generate correct visual structures and reasoning trajectories, rather than merely selecting answers. What a model builds provides a stronger probe of its internal 3D understanding than what it recognizes.
- **New Benchmark.** We introduce **MindBlock**, a spatial reasoning benchmark covering six tasks that span atomic operations (single-step assembly and deletion), multi-step structural reasoning (forward assembly and reverse deconstruction), and viewpoint consistency (local and global equivariance). Evaluation combines conventional 2D image similarity with **3DGS-Eval**, a volumetric consistency metric that exposes hidden geometric contradictions across views.
- **New Insights.** Our analysis reveals three systematic limitations of current UMMs: (1) a *2D–3D consistency gap*, where high perceptual similarity does not imply correct geometry; (2) *spatial drift* in multi-step reasoning, where structural errors accumulate over time; and (3) a pronounced *perception–execution gap*, where models recognize correct spatial configurations more reliably than they can generate them. Together, these findings suggest that while UMMs exhibit promising spatial intuition, their internal representations remain fragile and insufficient for robust 3D reasoning.

2. MindBlock: Probing Spatial Assembly and Structure

The MindBlock benchmark is designed to evaluate two fundamental dimensions of spatial intelligence: the capacity for *assembly*—the goal-directed construction and deconstruction of physical entities—and the mastery of *structure*—the geometric understanding of 3D space.

2.1. Evaluating Spatial Assembly

We define spatial assembly as the model's capacity to execute precise state transitions within a 3D coordinate frame. This is evaluated through two complementary single-step tasks and a long-horizon loop.

Single-Step Assembly We focus on the precision of atomic manipulations through two distinct paradigms:

- *Atomic Deletion:* **Targeted Deletion**. This task requires the model to precisely remove a specific target entity (e.g., "delete only the purple cube") from a given scene. It serves as a rigorous test for *structural persistence* and *occlusion recovery*: the model must preserve the exact coordinates of all remaining elements and the camera perspective while hallucinating the previously occluded background. This ensures the model treats the structure as a collection of independent 3D entities rather than a monolithic image.
- *Atomic Completion:* **Goal-Directed Completion**. Given a partial assembly and a target state image, the model must autonomously infer and generate the single most critical next step. This evaluates the model's proactive planning and its ability to bridge the gap between current and desired physical configurations.

Multi-Step Assembly-Deconstruction Loop To evaluate long-horizon consistency, we extend the aforementioned mappings into a bi-directional loop:

- *Iterative Completion:* **Forward Assembly**. The iterative execution of atomic completions, where the model must progressively build toward a target configuration across multiple steps.
- *Iterative Deletion:* **Reverse Deconstruction**. The continuous application of atomic deletions, requiring the model to sequentially dismantle a structure while maintaining spatial coherence at each step.

Unlike single-step tasks that probe atomic precision, the multi-step loop challenges the model to manage *structural evolution*. As the sequence progresses, the model must navigate dynamic occlusion and maintain the integrity of hierarchical support. Successfully deconstructing a complex structure requires an awareness of physical causalities—such as identifying load-bearing blocks before removal—that far exceeds the requirements of single-frame editing, providing a definitive test for the stability of the internal world model.

2.2. Evaluating Spatial Structure

To probe the model's internal 3D world model, we test for *geometric equivariance*—the principle that spatial representations should transform predictably under geometric operations.

Local Equivariance: **Local Structure Manipulation**: We task the model with rotating a specific sub-component (e.g., "Rotate the top triangular prism 90 degrees clock-

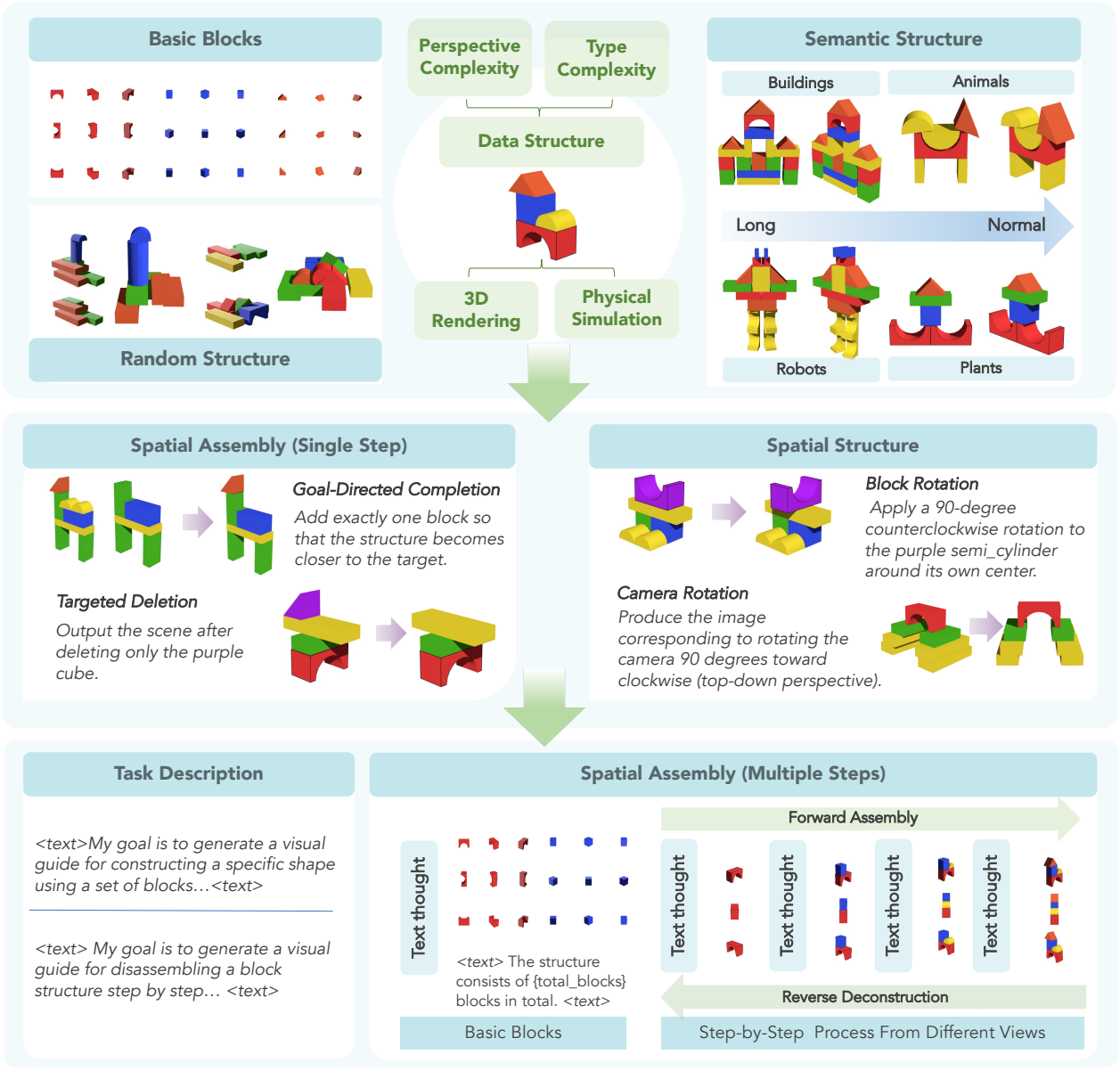

Figure 2. **Overview of the MindBlock benchmark and dataset.** The dataset is constructed from basic blocks via a 3D rendering and physical simulation pipeline (top). The benchmark evaluates spatial intelligence through two primary dimensions: **Spatial Assembly** (left and bottom), covering single-step and long-horizon construction/deconstruction, and **Spatial Structure** (right), testing geometric equivariance through local and global transformations.

wise") while preserving the rest of the structure. This evaluates the precision of the model's local coordinate system and its ability to disentangle individual parts within a structured whole.

Global Equivariance: **Global Structure Consistency**: We assess the model's ability to generate a globally transformed view (e.g., a 180-degree rotation) of the entire scene. This serves as a test for global 3D consistency; a model possessing a true structural understanding must maintain the relative positions and occlusions of all blocks across arbitrary viewpoint shifts, a capability we quantitatively verify using our **3DGS-Eval** protocol.

3. Dataset Construction and Evaluation

MindBlocks bridges the gap between 2D perception and 3D execution by grounding structural assembly in a rigorous physical and geometric protocol.

3.1. Physically-Grounded Data Synthesis

We utilize the **Genesis** physics engine [5] to generate a large-scale dataset of stable brick structures with precise control over geometry, lighting conditions, and camera placement. This simulation-based pipeline enables physically consistent rendering and fine-grained supervision signals for spatial reasoning tasks. Our structure library spans two complementary categories: (i) *semantic structures* (e.g., furniture, vehicles, animals), which leverage common-sense priors about real-world objects; (ii) *random structures*, which remove semantic cues to emphasize pure geometric reasoning and compositional spatial understanding. The overall distribution of categories and task complexity is illustrated in Fig. 3.

The construction of our dataset follows a rigorous simulation-in-the-loop pipeline. For the random structures, we employ a constrained random generation process combined with physical simulation to ensure the structural integrity of each step. For the semantic structures, we utilize a hybrid approach of manual modeling and physical simulation, encompassing a diverse set of scenes including buildings, furniture, and vehicles. A portion of the semantic models is adapted from PhyBlocks [10]. As shown in the block count distributions (Fig. 3, middle), the complexity of these structures is parameterized by their assembly sequence lengths, spanning from simple 5-block builds to complex "Long" sequences exceeding 40 blocks, thereby providing a comprehensive benchmark for evaluating model scalability in spatial reasoning.

For each structure, we generate a **bottom-up assembly sequence** that simulates the incremental construction process. Each construction trajectory consists of multiple discrete steps, where **exactly one block is added at each step**. This sequential formulation exposes the intermediate structural states of the scene and provides supervision for step-by-step reasoning about stability, placement, and support relationships.

At every step of the construction sequence, we render rich multimodal supervision signals:

- **Multi-view Observations:** We render images from multiple viewpoints covering a full 360-degree perspective. In practice, we sample eight predefined egocentric camera viewpoints around the structure, providing the multi-angle observations necessary for volumetric understanding and spatial disambiguation.
- **Structured Reasoning Traces:** We generate template-synthesized Chain-of-Thought (CoT) reasoning traces. These traces explicitly describe the intended assembly rationale, including block attributes, precise (x, y, z) coordinates, and hierarchical support relationships (e.g., "Place block A on top of block B").

3.2. 3DGS-Eval: Measuring Volumetric Consistency

To move beyond pixel-level similarity, we introduce **3DGS-Eval**, which employs **3D Gaussian Splatting (3D-GS)** to reconstruct a volumetric field from generated images. Following the pipeline in MVGBench [16], we initialize 3D Gaussians using generated views and their corresponding camera poses. Each primitive—parameterized by position, opacity, anisotropic covariance, and color—is optimized via **differentiable rasterization** to minimize photometric reconstruction error against the target images.

To capture fine geometric details, an **adaptive density control** mechanism iteratively refines the field by splitting, merging, or pruning Gaussians. This process forces the representation to converge into a coherent 3D entity. Consequently, 3DGS-Eval acts as a **"geometric polygraph"**: if the generated views lack underlying 3D consistency, the optimization fails to form a stable structure, leading to structural collapse.

Finally, the optimized 3D representation is re-rendered into 2D viewpoints. By comparing these renderings against the original generated images using **PSNR** and **SSIM**, we quantify the volumetric consistency and reconstruction fidelity of the multi-view content.

4. Experiments

4.1. Evaluation Setup

Models. We evaluate a diverse set of models across three primary categories to benchmark the current landscape of spatial intelligence:

- **General-purpose UMMs:** We include state-of-the-art autoregressive and diffusion-based unified multimodal models, namely **Emu3.5** [2], **Ovis-U1** [13], **ILLUME** [7], **BAGEL** [3] and **Omni-Gen** [14], which are capable of native image-text generation.
- **Specialized and Editing Models:** We assess **Blip3-o-next** [1] and **Step1X-Edit** [9], which are optimized for conditional image manipulation and structural preservation.

Metrics. We employ a dual-pronged evaluation strategy:

- **Image Similarity (2D):** We evaluate semantic alignment between generated and ground truth images using the CLIP Similarity score (*CLIP-S*), computed with a pre-trained CLIP encoder.
- **3DGS-Eval (3D):** Our core contribution. We use model-generated multi-view images to reconstruct a 3D scene via 3D Gaussian Splatting. Metrics include *PSNR*, *SSIM*, and *depth* for rendering quality, and Chamfer Distance (*CD*) for geometric fidelity against the ground truth.

Implementation Details. For 3DGS-Eval, we generate 8 views per structure. All baseline models are evaluated

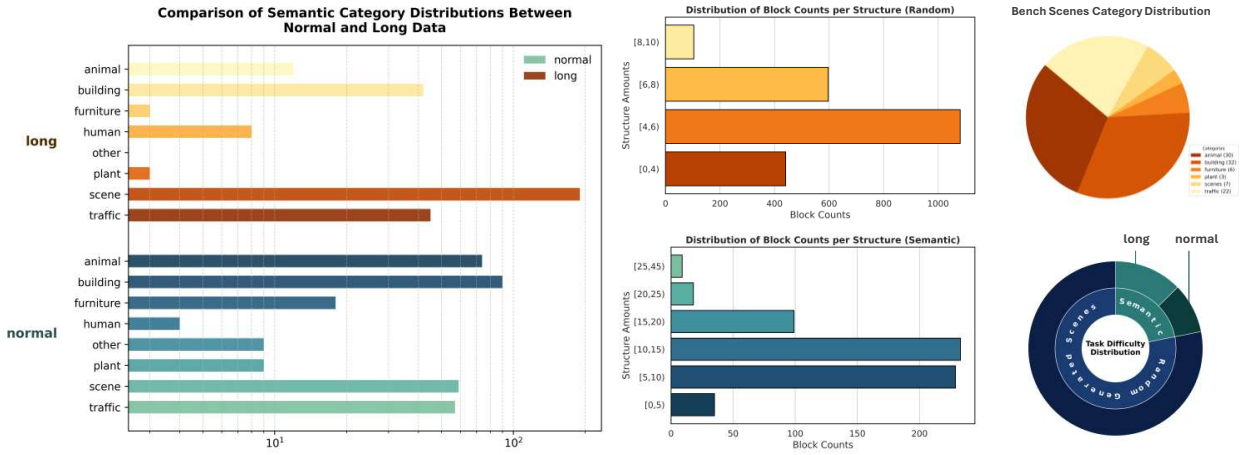

Figure 3. **Statistics of the synthesized dataset.** (Left) Comparison of semantic category distributions between "Normal" and "Long" data subsets. (Middle) Distribution of block counts (assembly steps) for Random and Semantic structures, illustrating the complexity gradient. (Right) Category distribution of bench scenes and the overall task difficulty ratio (Normal vs. Long) across Semantic and Random subsets.

Table 1. **Quantitative Comparison on MindBlock Benchmark.** We evaluate the single-step spatial intelligence of state-of-the-art Unified Multimodal Models (UMMs). The results are categorized into *Image Similarity* (2D heuristic) and our proposed *3D Gaussian Splatting* (volumetric consistency). ↑ denotes higher is better; ↓ denotes lower is better. Best results are **bolded**, and second best are underlined.

Task	Model	Image Similarity	3D Gaussian Splatting (3DGS-Eval)			
		CLIP-S↑	PSNR↑	SSIM↑	Depth Err.↓	CD↓
Spatial Assembly (Single-step Construction)						
	BAGEL [3]	0.77	13.83	0.73	75.97	11.72
	Blip3-o-next [1]	0.81	11.90	0.69	83.20	13.05
	ILLUME [7]	0.68	9.40	0.61	101.30	16.20
	Emu3.5 [2]	0.75	13.10	0.71	79.20	12.10
	Step1X-Edit [9]	**0.86**	**15.20**	**0.78**	**68.50**	**9.10**
	Ovis-U1 [13]	0.76	14.10	0.75	73.10	10.40
	Omni-Gen [14]	0.70	8.70	0.59	108.40	17.80
Spatial Structure (Geometric Equivariance)						
	BAGEL [3]	0.74	12.26	0.76	82.64	10.73
	Blip3-o-next [1]	0.79	10.80	0.70	90.50	12.90
	ILLUME [7]	0.66	8.30	0.60	110.20	17.40
	Emu3.5 [2]	0.72	11.50	0.73	86.70	11.80
	Step1X-Edit [9]	**0.83**	**13.80**	**0.80**	**74.20**	**9.30**
	Ovis-U1 [13]	0.73	12.90	0.77	79.60	10.20
	Omni-Gen [14]	0.68	7.90	0.57	118.60	19.30

using their default inference hyper-parameters with a temperature of 0.2.

4.2. Analysis of Single-step Assembly and Spatial Structure

As shown in Table 1, we observe several important findings regarding the spatial reasoning capabilities of current Unified Multimodal Models (UMMs).

The 2D–3D Consistency Gap. A clear discrepancy emerges between 2D perceptual similarity and volumetric consistency. For example, *Blip3-o-next* [1] achieves relatively strong 2D alignment with a CLIP-S score higher than *BAGEL* [3]. However, its 3D reconstruction metrics are substantially worse, exhibiting lower PSNR/SSIM and sig-

Table 2. Performance during forward assembly and reverse deconstruction phases. ↑ / ↓ indicates direction of improvement.

Model	Image Similarity	3DGS-Eval (3D)			
	CLIP-S↑	PSNR↑	SSIM↑	Depth Err.↓	CD↓
Phase I: Forward Assembly					
BAGEL-zebra-CoT [3]	0.78	13.5	0.74	71.3	11.8
ThinkMorph [6]	0.76	15.1	0.79	77.4	9.6
MindBlock-Random (Ours)	0.84	17.8	0.86	62.5	7.4
MindBlock-Semantic (Ours)	**0.90**	**20.4**	**0.91**	**55.8**	**5.9**
Phase II: Reverse Deconstruction					
BAGEL-zebra-CoT [3]	0.73	12.4	0.71	83.6	13.2
ThinkMorph [6]	0.75	14.0	0.76	76.8	10.7
MindBlock-Random (Ours)	0.82	16.3	0.83	67.9	8.5
MindBlock-Semantic (Ours)	**0.88**	**19.1**	**0.88**	**60.2**	**6.8**

nificantly higher depth and Chamfer errors. This indicates that although individual generated views appear semantically plausible, they fail to form a geometrically coherent 3D scene when aggregated across viewpoints. In contrast, models such as *Ovis-U1* [13] achieve competitive or even better 3D reconstruction quality despite having comparable or slightly lower CLIP-S scores. These results highlight that strong 2D image similarity does not necessarily imply correct spatial reasoning.

Spatial Assembly Remains Challenging but Partially Solvable. For the *Spatial Assembly* task, most models show moderate performance, with *Step1X-Edit* [9] achieving the best overall results across both 2D and 3D metrics. Models such as *Emu3.5* [2] and *Ovis-U1* [13] remain close to *BAGEL* [3], suggesting that current UMMs can partially reason about single-step object placement. However, large performance gaps remain between stronger models and weaker ones such as *ILLUME* [7] and *Omni-Gen* [14], which fail to produce geometrically stable structures.

Equivariance is a Major Bottleneck. All evaluated models show a consistent performance drop on the *Spatial Structure* task. Compared to *Spatial Assembly*, both 2D and 3D metrics deteriorate across nearly all models, indicating that reasoning about viewpoint transformations is substantially more difficult. Even strong models frequently produce inconsistent internal layouts when asked to imagine the same structure from a new viewpoint, leading to degraded reconstruction quality and larger geometric errors. This suggests that current UMMs still lack robust 3D-equivariant internal representations.

3D Evaluation Reveals Hidden Failures. Importantly, several models that appear competitive under 2D metrics reveal substantial failures under our 3DGS-Eval protocol. This demonstrates that traditional image-based evaluation may significantly overestimate spatial reasoning ability. By explicitly reconstructing a volumetric scene from model-generated views, our evaluation exposes inconsistencies that remain invisible under standard 2D metrics.

4.3. Multi-step Assembly: The Challenge of Interleaved Reasoning

Multi-step assembly is a more rigorous probe of spatial intelligence, as it requires the model to generate an interleaved sequence of text (reasoning/instructions) and images (intermediate states). Most contemporary UMMs are limited to single-turn image generation and fail this task entirely. We evaluate the few models capable of long-context interleaved output: *BAGEL-zebra-CoT*, *Thinkmorph*, and our trained *MindBlock* variants.

Model Specialization for Interleaved Reasoning. Since the ability to generate long-context, interleaved text-image sequences is not yet a standard feature in most off-the-shelf UMMs, we perform continual training on the *BAGEL-zebra-CoT* [3] backbone to establish a competitive baseline for multi-step evaluation. We develop two specialized variants: **MindBlock-Random** and **MindBlock-Semantic**, fine-tuned on the respective partitions of our dataset2. During the training phase, the models are jointly supervised on two core assembly operators: *forward assembly* and *reverse deconstruction*. To enforce robust 3D structural awareness, each state transition is supervised across four views, requiring the model to maintain viewpoint-invariant information while progressing through the interleaved reasoning trajectory. This specialization ensures that the performance gains reported in Table 2 reflect the models' internalized spatial logic rather than a mere failure to adhere to the interleaved format.

The results in Table 2 underscore the difficulty of maintaining structural integrity over multiple steps. General-purpose models like *Thinkmorph* suffer from "spatial drift," where small errors in early steps accumulate, leading to physically impossible final structures. In contrast, our *MindBlock-Semantic* model, which benefits from a two-stage curriculum training on meaningful structures, demonstrates superior stability, effectively learning a native spatial intuition that goes beyond simple symbolic matching.

5. Further Analyses and Discussions

5.1. Is Spatial Reasoning Carried by CoT or Intuition?

A fundamental question in unified multimodal modeling is whether spatial structural awareness is an emergent property of generative "intuition" (pixel-level patterns) or if it relies on explicit symbolic reasoning. We investigate this by comparing UniCoT [12]—which generates a textual Chain-of-Thought (CoT) prior to the image—against the base BAGEL model [3] which performs direct image synthesis.

The Divergence Metric. We formalize the impact of reasoning through the *Spatial Reasoning Gain* (Δ_{SR}), measuring the improvement in 3D structural integrity when CoT

Table 3. Comparison of Spatial Intelligence between Direct Generation (BAGEL) and CoT-Augmented Generation (UniCoT)[12]. 3DGS-Eval metrics represent the quality of a 3D reconstruction from 8 generated views. CoT-based approach significantly reduces geometric errors (Chamfer Distance) while maintaining high semantic alignment.

Model	CLIP-S ↑	3DGS-PSNR ↑	3DGS-SSIM ↑	Chamfer Dist. ↓
BAGEL (Direct)	0.782	13.42	0.712	12.4
UniCoT[12]	**0.814**	**20.88**	**0.785**	**6.5**
Gain (Δ_{SR})	+0.032	+7.46	+0.073	−47.6%

is enabled. Let $\mathcal{G}_{cot}$ and $\mathcal{G}_{dir}$ denote the image generation processes for UniCoT and the baseline, respectively. The divergence is defined as:

$$\Delta_{SR} = \mathbb{E}_{s\sim\mathcal{S}}[\Phi(\mathcal{G}_{cot}(I, T|\text{CoT})) - \Phi(\mathcal{G}_{dir}(I, T))] \quad (1)$$

where Φ represents the 3DGS-Eval score, and $\mathcal{S}$ is the set of single step assembly instructions in MindBlock.

Quantitative Analysis: CoT as a Symbolic Scaffold. As shown in Table 3, direct generation (BAGEL) achieves reasonable 2D similarity (CLIP-S) but fails under 3DGS-Eval. Without the "textual anchor" of a CoT, the model often neglects occluded block faces or misinterprets the depth of "next-to" relations. In contrast, UniCoT shows a improvement in 3D Reconstructibility. This suggests that the CoT acts as a **symbolic scaffold**: by explicitly predicting the coordinates and support relations (e.g., *"Block A is placed at (x,y,z) on top of B"*), the model constrains the latent diffusion process to respect 3D geometry rather than merely hallucinating a 2D projection.

5.2. The Consistency Gap: Perception vs. Execution

To further point out whether the failure of spatial reasoning originates from perceptual misunderstanding or generative incapacity, we conduct a focused diagnostic analysis on the **BAGEL** model. Specifically, we decouple the *Spatial Assembly* task into two paradigms: (1) **Discriminative Recognition**, where BAGEL must select the correct next-step image from a 4-way Multiple-Choice Question (MCQ), and (2) **Constructive Execution**, where the model generates the next-step image in pixel space.

To objectively evaluate the semantic fidelity of the generated images, we introduce a **GPT-Matching** protocol. A GPT-5 judge is presented with the model-generated image alongside the four ground-truth and distractor options from the MCQ. The judge identifies which of the four options the generated image most closely resembles. If the judge selects the ground-truth option, the execution is deemed successful.

Table 4. **Diagnostic Analysis of BAGEL: The Knowledge-Action Gap.** We contrast the model's accuracy in selecting the correct spatial step against its ability to generate a semantically aligned image of that same step.

Task Setting	Perception Acc.↑	Execution Match↑	The Gap (Δ)↓
Single-step Assembly	52.5%	28.2%	**24.3%**

The Generative Bottleneck. As shown in Table 4, a discrepancy exists between BAGEL's internal spatial knowledge and its constructive output. While the model correctly identifies the intended spatial state in nearly half of the cases (52.5%), it successfully executes that state in pixel space only 28.2% of the time. This 24.3% gap reveals that spatial intelligence in current UMMs is often *superficial*: models may possess the discriminative "intuition" to recognize correct 3D relationships, yet they lack the **generative grounding** required to translate this symbolic understanding into consistent, physically-plausible visual structures. Our GPT-Matching results further indicate that failed generations frequently drift toward visual distractors, even when the model's discriminative head had correctly rejected them during the MCQ phase.

6. Conclusion

In this work, we introduced **MindBlock**, a diagnostic benchmark for evaluating spatial intelligence in unified multimodal models through *active construction*. Moving beyond passive visual question answering, MindBlock requires models to generate step-by-step visual assemblies and maintain structural consistency under spatial transformations. To rigorously evaluate geometric coherence, we further proposed **3DGS-Eval**, a volumetric consistency protocol that reconstructs a 3D representation from model-generated multi-view images.

Our key findings are three-fold. First, a *2D–3D consistency gap* pervades current UMMs: models that achieve high perceptual similarity scores often produce multi-view generations that are geometrically incoherent, a failure revealed only through our **3DGS-Eval** protocol. Second, geometric equivariance remains a critical bottleneck, with all evaluated models degrading substantially when asked to reason about the same structure from a novel viewpoint. Third, a pronounced *perception–execution gap* (24.3% for BAGEL) demonstrates that models frequently understand what to build yet fail to build it, suggesting that discriminative spatial knowledge does not automatically transfer to generative spatial grounding. Explicit Chain-of-Thought reasoning partially bridges this gap, acting as a symbolic scaffold that stabilizes 3D structure during generation.

References

[1] Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Lifu Huang, Tianyi Zhou, Junnan Li, Silvio Savarese, Caiming Xiong, and Ran Xu. Blip3o-next: Next frontier of native image generation, 2025. 5, 6

[2] 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, 2025. 5, 6, 7

[3] 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. 2, 5, 6, 7

[4] Andrea Frick, Katrina Ferrara, and Nora S Newcombe. Using a touch screen paradigm to assess the development of mental rotation between $31/2$ and $51/2$ years of age. *Cognitive processing*, 14(2):117–127, 2013. 2

[5] Genesis Authors. Genesis: A universal and generative physics engine for robotics and beyond. https://github.com/Genesis-Embodied-AI/Genesis, 2024. 5

[6] Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning, 2026. 7

[7] Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, and Hang Xu. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement, 2025. 5, 6, 7

[8] Barbara Landau, E Emory Davis, Cathryn S Cortesa, Zihan Wang, Jonathan D Jones, and Amy L Shelton. Young children's copying of block constructions: Significant constraints in a highly complex task. *Cognitive Development*, 71:101463, 2024. 2

[9] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing, 2025. 5, 6, 7

[10] Liang Ma, Jiajun Wen, Min Lin, Rongtao Xu, Xiwen Liang, Bingqian Lin, Jun Ma, Yongxin Wang, Ziming Wei, Haokun Lin, Mingfei Han, Meng Cao, Bokui Chen, Ivan Laptev, and Xiaodan Liang. Phyblock: A progressive benchmark for physical understanding and planning via 3d block assembly, 2025. 2, 5

[11] Ava Pun, Kangle Deng, Ruixuan Liu, Deva Ramanan, Changliu Liu, and Jun-Yan Zhu. Generating physically stable and buildable brick structures from text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14798–14809, 2025. 2

[12] Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision, 2026. 7, 8

[13] Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, Yang Li, and Qing-Guo Chen. Ovis-u1 technical report, 2025. 5, 6, 7

[14] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation, 2024. 5, 6, 7

[15] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 2

[16] Xianghui Xie, Chuhang Zou, Meher Gitika Karumuri, Jan Eric Lenssen, and Gerard Pons-Moll. Mvgbench: Comprehensive benchmark for multi-view generation models, 2025. 5

[17] Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiang Zhu, and Jianyu Chen. Up-vla: A unified understanding and prediction model for embodied agent. In *International Conference on Machine Learning*, pages 74911–74922. PMLR, 2025. 2