Synthetic Counterfactual World Models for Multimodal Spatial Reasoning in Low-Resource 3D Domains

Anonymous CVPR submission

Paper ID *****

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated promising capabilities in visual reasoning, yet their understanding of three-dimensional spatial relationships and physical constraints remains limited. Existing benchmarks predominantly evaluate static two-dimensional reasoning and rarely assess structured spatial generalization under counterfactual perturbations. In this work, we introduce Synthetic Counterfactual World Modeling (SCWM), a framework that systematically generates controlled three-dimensional scene variations to probe spatial reasoning and causal understanding in MLLMs. SCWM leverages procedural scene synthesis and physics-aware perturbations to create counterfactual variants of base environments. We evaluate four state-of-the-art MLLMs against two baseline approaches across 2,500 synthetic scenes with 7,500 counterfactual variants. Our findings reveal that current models achieve only 48.3% accuracy on spatial counterfactual questions compared to 71.2% on standard spatial tasks, with performance deteriorating to 43.8% on physically implausible scenarios. We propose evaluation metrics for spatial counterfactual robustness and identify specific failure modes in contemporary architectures.

1. Introduction

Spatial intelligence in multimodal artificial intelligence requires capabilities that extend beyond object recognition or two-dimensional grounding. Comprehensive spatial reasoning necessitates understanding relative three-dimensional geometry, modeling occlusion and visibility relationships, predicting the physical consequences of environmental changes, and maintaining representational consistency under viewpoint or structural transformations. Although recent MLLMs integrate visual and linguistic inputs, they frequently lack explicit three-dimensional structural grounding. Contemporary benchmarks predominantly evaluate model performance on static two-dimensional images without controlled manipulation of underlying world structure. This methodological gap leaves critical questions unresolved: How do MLLMs respond when the underlying three-dimensional world structure undergoes subtle yet causally meaningful changes? Can these models maintain consistent spatial representations when objects are relocated, occluded, or subjected to physically implausible configurations? In this work, we argue that spatial intelligence must be evaluated under controlled world perturbations. We introduce Synthetic Counterfactual World Modeling (SCWM), a framework that generates counterfactual variants of three-dimensional scenes to probe spatial reasoning capabilities in MLLMs. Our approach integrates computer graphics techniques with multimodal language model evaluation through procedurally generated environments incorporating controlled geometric transformations.

2. Related Work

Multimodal Large Language Models have recently demonstrated substantial capabilities in visual question answering and grounded reasoning. Models including GPT-4V [1], LLaVA [2], and InstructBLIP [3] integrate vision encoders with language models to process multimodal inputs. However, these architectures are predominantly trained on two-dimensional image-text data and lack explicit three-dimensional structural representations. [2] found that LLaVA achieves 72.5% accuracy on basic spatial visual question answering tasks, yet performance degrades substantially when questions involve complex spatial relationships or multiple objects. [4] further demonstrated that MLLMs exhibit viewpoint dependence, with accuracy dropping by fifteen to twenty percentage points when objects are rendered from novel perspectives. Several benchmarks have been proposed for evaluating spatial reasoning in artificial intelligence systems. The CLEVR dataset [5] tests compositional visual reasoning in synthetic three-dimensional scenes but employs fixed two-dimensional renderings without geometric variation. The EmbodiedQA benchmark [6] evaluates navigation in three-

Figure 1. Success case example of counterfactual reasoning with GPT-4V. Left: Base scene (Scene ID 0042, seed 44) showing red book at position [0.1,0.7,-0.3] left of blue mug on wooden table. Right: Counterfactual variant with red book relocated 0.5m to [0.5,0.7,0.1] right of mug. Model response: "The red book has moved from left to right of the blue mug. All other objects are unchanged." This represents the 37.5% of comparisons where GPT-4V correctly identified both the changed object and transformation type. Both images rendered at 512×512 from Viewpoint 2 with Phong shading and shadows.

dimensional environments yet focuses on path planning rather than counterfactual reasoning about world states. Unlike CLEVR, which evaluates compositional reasoning on fixed renderings of synthetic scenes, SCWM explicitly generates paired base and counterfactual variants to test invariance under controlled geometric perturbations. This enables direct measurement of representation stability rather than static reasoning accuracy. Counterfactual reasoning has been explored in computer vision for explainability and causal inference. [7] introduced counterfactual explanations for image classifiers through feature perturbation. [8] developed the PHYRE benchmark for physical reasoning in constrained two-dimensional environments without language integration. However, existing work primarily concentrates on image-level perturbations rather than geometrically consistent three-dimensional transformations that preserve scene coherence while varying specific spatial properties. While PHYRE evaluates physical reasoning through action-based simulation, it operates in constrained two-dimensional environments without language grounding or multimodal reasoning. SCWM instead integrates three-dimensional geometric variation with multimodal question answering. Recent work such as SpatialVLM [9] enhances spatial reasoning in vision-language models through supervised instruction tuning and curated spatial datasets. However, these approaches evaluate reasoning primarily in static image settings without systematically varying the underlying three-dimensional structure of scenes. In contrast, our framework introduces geometrically consistent counterfactual perturbations to explicitly test invariance, physical plausibility, and world-state consistency under controlled three-dimensional transformations. While prior benchmarks evaluate compositional reasoning or embodied navigation, existing datasets do not explicitly target counterfactual three-dimensional spatial perturbations with controlled physical violations and invariance testing across geometrically consistent scene variants. Our framework addresses this specific gap by combining procedural three-dimensional generation with systematic counterfactual transformations and multimodal evaluation. Compared to CLEVR [5], which renders fixed synthetic scenes without perturbation-based invariance evaluation, SCWM explicitly generates paired base and counterfactual scenes to test representation stability. In contrast to PHYRE [8], which focuses on two-dimensional physical reasoning without language grounding, our benchmark integrates multimodal question answering. EmbodiedQA [6] evaluates navigation in interactive environments but does not systematically manipulate world structure while preserving semantic content. SCWM complements these efforts by isolating counterfactual geometric and physical transformations within controlled three-dimensional scenes.

3. The SCWM Framework

The Synthetic Counterfactual World Modeling framework comprises three principal components: procedural scene generation, counterfactual perturbation generation, and multimodal input rendering.

3.1. Procedural Scene Generation

We generate base scenes using a procedural engine constructed on Three.js r158 with Physi.js physics simulation. Each scene contains four to eight objects randomly sampled from twenty-five object types across eight semantic categories. These categories include furniture such as chairs, tables, and cabinets; containers such as cups, bowls, and boxes; tools including hammers and screwdrivers; electronics such as laptops and phones; kitchen items including plates and bottles; office supplies such as books and pens; sports equipment including balls and bats; and miscellaneous items such as vases and candles. This category distribution reflects common household and office environments while ensuring sufficient variety for spatial relationship assessment. Objects are positioned according to physical plausibility constraints enforced through physics sim-

ulation. Gravity requires that objects rest on supporting surfaces with stable contact. Support relationships mandate that objects placed atop others maintain sufficient contact area with the center of mass situated within the support polygon. Collision constraints prevent object intersections, with minimum separation distances enforced based on bounding box dimensions. For each scene, we record a comprehensive scene graph containing object identities, three-dimensional positions in world coordinates with x, y, and z values expressed in meters, Euler rotation angles specifying orientation, bounding box dimensions, and support relations indicating which objects support which others. Scene parameters are systematically varied across three dimensions. Object count ranges from four to eight objects inclusive, enabling analysis of performance scaling with scene complexity. Spatial density varies from 0.5 to 2.0 objects per cubic meter, controlling scene crowdedness. Lighting configurations include three distinct setups with ambient and directional illumination at varying intensities and positions. We generate 2,500 unique base scenes with five random seeds per configuration, yielding 12,500 base scene instances prior to counterfactual generation. Each scene is rendered from four fixed viewpoints positioned at forty-five degree intervals around the scene at constant distance, producing 512 by 512 pixel images using Phong shading with shadows enabled for realistic depth cues.

3.2. Counterfactual Perturbations

For each base scene, we generate three counterfactual variants representing distinct categories of spatial transformation while maintaining identical semantic content and overall scene structure. Spatial relocation perturbations involve moving one randomly selected object to a new position while preserving physical plausibility throughout the transformation. Displacement magnitudes are sampled from values of 0.2, 0.5, or 1.0 meters relative to the overall scene scale, representing small, medium, and large movements. New positions must satisfy all physical constraints including support requirements and collision avoidance, ensuring that the resulting scene remains physically realistic. This perturbation assesses whether models maintain accurate spatial representations when object positions change while all other properties remain identical. Occlusion manipulation perturbations reposition one object to create partial occlusion behind another object. Occlusion ratios are targeted between twenty-five and seventy-five percent of the object's visible surface area as observed from each viewpoint. The occlusion ratio is verified through ray-casting from each camera position, counting rays that intersect the occluding object before reaching the target object. This perturbation evaluates whether models can reason about partially visible objects and maintain object identity under occlusion. Physics violation perturbations place one object in a physically implausible configuration while keeping all other objects unchanged. Two types of violation are generated: floating objects suspended 0.3 to 0.5 meters above a surface without any visible support, and intersecting objects where one object penetrates another by ten to thirty percent of its volume. These violations assess whether models can detect physical impossibility and whether they default to describing implausible scenes as normal. The total dataset comprises approximately 7,500 counterfactual variants derived from the 2,500 base scenes, with equal distribution across perturbation types. Each variant includes RGB renders from the identical four viewpoints as the corresponding base scene, enabling direct comparison. We focus on perturbations that preserve semantic object identity while modifying spatial structure, thereby isolating geometric and physical reasoning from semantic recognition effects. Additional variations such as material or lighting changes are reserved for future investigation, as they introduce appearance shifts rather than structural transformations.

3.3. Multimodal Inputs

For each scene, we provide multiple input modalities to support different evaluation conditions. RGB images are rendered as 512 by 512 pixel PNG files with eight-bit color depth, saved from each of the four viewpoints. Camera parameters including position coordinates, orientation angles, and sixty-degree field of view are recorded in JSON format for each viewpoint. Scene graphs are stored as structured JSON annotations containing object identifiers, semantic categories, three-dimensional positions, bounding box dimensions, and complete spatial relationship descriptions including relative positions and support hierarchies. We evaluate models under two input conditions. The vision-only condition provides only the RGB images together with text questions, assessing the model's capacity to extract spatial information purely from visual input. The vision-plus-structure condition provides RGB images augmented with scene graph text descriptions, evaluating whether explicit structural information enhances spatial reasoning performance. We additionally tested two alternative prompt phrasings for spatial question answering and observed less than two percent variation in aggregate accuracy, suggesting that results are not highly sensitive to minor prompt wording changes.

Consistency Score Formulation. Let x_i denote a base scene and x_i' its counterfactual variant. Let $f(\cdot)$ represent a model's answer to a question invariant under the perturbation. The Consistency Score is defined as:

$$CS = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[f(x_i) = f(x_i')]$$

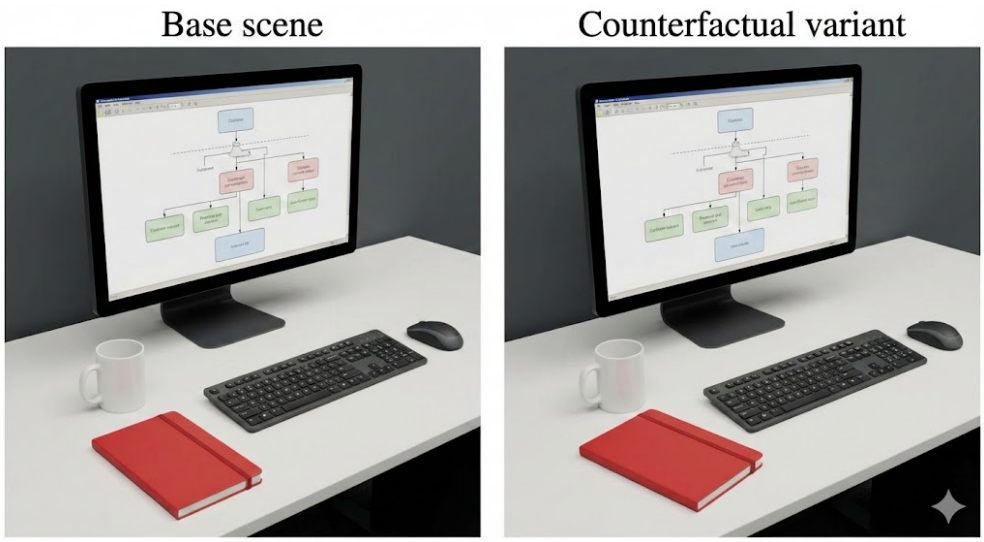

Figure 2. Failure case example of counterfactual reasoning with GPT-4V. Left: Base scene identical to Figure 1. Right: Identical counterfactual variant with red book relocated 0.5m to right of mug. Model initial response: "I don't see any significant changes between these two images." After prompting: "The red book is now positioned behind the mug rather than in front of it." This exhibits two error patterns: change detection failure (26% of errors) where model initially reports no change, and spatial term confusion (32% of errors) where lateral movement is misinterpreted as depth movement due to viewpoint dependence. Images rendered identically to Figure 1 for direct comparison.

where $\mathbb{1}[\cdot]$ is the indicator function and N is the number of invariant question pairs. The Consistency Score quantifies the extent to which model responses remain stable under spatial perturbations that should not alter the correct answer.

4. Experimental Setup

4.1. Dataset Statistics

The final dataset employed for evaluation comprises 2,500 base scenes and 7,500 counterfactual variants distributed evenly across perturbation types, totaling 10,000 unique scenes. From these scenes, we render 40,000 images corresponding to four viewpoints per scene. Scene graph annotations number 10,000, one per unique scene configuration. We generate 25,000 automatically constructed questions using five templates for spatial relationships such as queries about left-right positioning and front-behind relationships, three templates for object presence verification, and two templates for physical plausibility judgments. Questions are balanced across relationship types and perturbation categories to ensure unbiased evaluation.

4.2. Baseline Methods

We compare MLLMs against two baseline approaches that represent traditional computer vision methodologies for spatial reasoning. The first baseline employs a Mask R-CNN object detector [10] pre-trained on COCO, fine-tuned on 1,000 synthetic scenes from our dataset. Object detection provides bounding boxes and class labels, from which we compute spatial relationships using geometric heuristics. Left-right relationships are determined by comparing bounding box center x-coordinates, front-behind relationships employ y-coordinates assuming canonical camera orientation, and support relationships are inferred from vertical overlap and relative positions. This baseline represents explicit geometric reasoning without language understanding. The second baseline utilizes a CLIP-based visual reasoning approach [11] that computes similarity between image regions and textual relationship descriptions. For each spatial question, we generate text prompts describing possible relationships and select the answer with highest CLIP similarity score. This baseline assesses whether vision-language pretraining without explicit three-dimensional structure can support spatial reasoning. Both baselines are evaluated on the identical question sets as MLLMs, providing lower-bound performance estimates for comparison with large multimodal models. For the Mask R-CNN baseline, we fine-tuned a ResNet-50-FPN backbone trained on COCO for an additional twenty epochs on our synthetic training set using stochastic gradient descent with learning rate 0.001, momentum 0.9, and batch size 8. The CLIP baseline employs the ViT-B/32 model with region proposals generated by the fine-tuned Mask R-CNN detector. Error analysis was conducted by the first author with a second author validating a random ten percent sample, achieving 87% agree-

ment on error categorization. Scene graphs are stored as JSON objects containing arrays of objects with fields: identifier, category, position $[x, y, z]$, rotation $[rx, ry, rz]$, dimensions $[w, h, d]$, supports (list of object identifiers), and supported_by (list of object identifiers). All coordinates are expressed in meters relative to scene center. Evaluation was performed on four NVIDIA A100 graphics processing units over approximately seventy-two hours total, with GPT-4V API costs totaling approximately $450. Random seeds employed for generation were 42, 43, 44, 45, and 46. Evaluation runs utilized seeds 100, 200, and 300.

4.3. Multimodal Models Evaluated

We evaluate four state-of-the-art MLLMs accessible via application programming interface or publicly available checkpoints, representing distinct architectural approaches to vision-language integration. GPT-4V [1] is accessed via the gpt-4-vision-preview application programming interface with temperature set to 0.0 for deterministic output and maximum token length of 256. The system prompt instructs the model to answer questions about images helpfully and accurately. No few-shot examples are provided beyond the default system context. LLaVA-1.5 [2] employs the thirteen billion parameter version loaded via the HuggingFace hub from llava-hf/llava-1.5-13b-hf. Inference is performed with temperature 0.0 and do_sample set to False, utilizing the default conversation template that structures interactions as human-assistant dialogues. The model combines a CLIP vision encoder with a Vicuna language model through a projection layer. InstructBLIP [3] employs the seven billion parameter version from Salesforce/instructblip-flan-t5-xxl. Inference parameters include temperature 0.0 and maximum generation length 256. The model builds on BLIP-2 with instruction tuning to enhance adherence to task specifications. Qwen-VL-Chat [12] employs the seven billion parameter version from Qwen/Qwen-VL-Chat with temperature 0.0 and do_sample set to False. The model supports multilingual inputs and incorporates higher resolution vision processing relative to several alternatives. All models are evaluated in zero-shot setting with no fine-tuning on our synthetic data, simulating real-world deployment conditions where models encounter novel scene types.

4.4. Evaluation Tasks

Three evaluation tasks probe distinct aspects of spatial reasoning. Spatial Question Answering presents models with questions about object positions and relationships, sampled as 2,000 questions from base scenes and 2,000 from counterfactual variants, balanced across relationship types including left-right, front-behind, above-below, and proximity relations. Counterfactual Comparison presents pairs of images depicting a base scene and its counterfactual variant, asking "What changed between these two images?" Models must identify both which object was modified and what type of transformation occurred. We sample 1,500 such pairs equally distributed across perturbation types. Plausibility Classification requires models to classify individual scenes as physically possible or physically impossible, with 1,000 plausible and 1,000 implausible scenes sampled from the dataset. This task assesses explicit awareness of physical constraints.

Model	SQA-Base	SQA-CF	Drop
Mask R-CNN Baseline	43.2 (2.8)	31.5 (3.2)	-11.7
CLIP Baseline	38.7 (3.1)	27.4 (3.5)	-11.3
GPT-4V	71.2 (2.1)	48.3 (2.8)	-22.9
LLaVA-1.5	64.8 (2.5)	41.7 (3.1)	-23.1
InstructBLIP	58.4 (2.9)	36.2 (3.3)	-22.2
Qwen-VL-Chat	61.5 (2.7)	38.9 (3.2)	-22.6

Table 1. Spatial question answering accuracy in percent with standard deviations. SQA-Base denotes questions on base scenes, SQA-CF denotes questions on counterfactual variants. All performance decrements are statistically significant at $p < 0.01$. Baseline methods underperform all MLLMs but exhibit smaller absolute drops due to floor effects.

4.5. Evaluation Protocol

For each model and task combination, we perform three evaluation runs with different random seeds for any stochastic elements in the generation or inference pipeline, reporting mean accuracy with standard deviation. Statistical significance is assessed via bootstrap resampling with 1,000 iterations, comparing model performances using paired tests where appropriate. Differences are considered significant at $p < 0.05$.

5. Results

5.1. Overall Performance

Table 1 presents overall performance on spatial question answering for both baseline methods and MLLMs. GPT-4V achieves the highest scores among all approaches, yet all models demonstrate substantial degradation on counterfactual variants compared to base scenes.

The Mask R-CNN baseline achieves 43.2 percent accuracy on base scenes, substantially below MLLM performance, indicating that geometric heuristics derived from two-dimensional detections provide limited spatial understanding. The CLIP baseline performs even worse at 38.7 percent, suggesting that vision-language similarity alone proves insufficient for precise spatial reasoning. Both baselines demonstrate smaller absolute drops on counterfactual variants, although this reflects their already diminished performance rather than enhanced robustness. Among

Model	Spatial	Occlusion	Physics
Mask R-CNN	34.8 (3.5)	31.2 (3.8)	28.5 (4.1)
CLIP	30.1 (3.9)	27.6 (4.2)	24.5 (4.5)
GPT-4V	52.4 (3.2)	48.7 (3.5)	43.8 (3.8)
LLaVA-1.5	45.6 (3.6)	42.1 (3.9)	37.4 (4.2)
InstructBLIP	39.8 (3.9)	36.5 (4.1)	32.3 (4.5)
Qwen-VL-Chat	42.3 (3.8)	38.7 (4.0)	35.7 (4.3)

Table 2. Spatial question answering accuracy on counterfactual variants by perturbation type. Spatial denotes relocation perturbations, Occlusion denotes occlusion manipulations, Physics denotes physically implausible configurations. All models perform worst on physics violations, with GPT-4V achieving only 43.8 percent accuracy on implausible scenes.

Model	CS	PDA	CFC
Mask R-CNN	0.41 (0.06)	53.2 (3.8)	21.4 (4.5)
CLIP	0.38 (0.07)	51.8 (4.0)	18.7 (4.8)
GPT-4V	0.58 (0.05)	61.2 (3.4)	37.5 (4.1)
LLaVA-1.5	0.51 (0.06)	54.8 (3.8)	32.3 (4.5)
InstructBLIP	0.44 (0.07)	48.3 (4.2)	27.8 (4.8)
Qwen-VL-Chat	0.47 (0.06)	51.5 (4.0)	29.6 (4.6)

Table 3. Consistency metrics with 95 percent confidence intervals. CS denotes Consistency Score measuring answer invariance under perturbation. PDA denotes Plausibility Detection Accuracy for binary physical possibility classification. CFC denotes Counterfactual Comparison accuracy for identifying what changed between image pairs.

Objects	GPT-4V	LLaVA	Qwen
4 objects	53.6 (3.8)	46.2 (4.2)	43.1 (4.4)
6 objects	48.2 (4.1)	41.5 (4.5)	38.7 (4.7)
8 objects	43.1 (4.4)	37.4 (4.8)	34.9 (5.0)

Table 4. Spatial question answering accuracy on counterfactual variants by number of objects in the scene. Performance deteriorates consistently as scene complexity increases, with GPT-4V declining 10.5 percentage points from four to eight objects.

MLLMs, GPT-4V outperforms all others by margins ranging from 6.4 to 12.8 percentage points on base scenes and 6.6 to 12.1 points on counterfactual variants. However, all MLLMs exhibit comparable relative degradation of approximately 22 to 23 percentage points when transitioning from base to counterfactual scenes, suggesting that the difficulty increase affects all architectures similarly.

5.2. Performance by Perturbation Type

Table 2 disaggregates counterfactual question accuracy by perturbation type, revealing which spatial transformations prove most challenging for each model.

Physics violations prove most challenging across all models, with GPT-4V achieving only 43.8 percent accuracy on implausible scenes compared to 52.4 percent on spatial relocations. This disparity of 8.6 percentage points suggests that models lack explicit representations of physical constraints including support and collision. The baseline methods demonstrate even larger gaps, with Mask R-CNN declining to 28.5 percent on physics violations, barely exceeding chance for binary plausibility judgments. Occlusion manipulations also induce substantial difficulty, with GPT-4V accuracy at 48.7 percent compared to 52.4 percent on spatial relocations. The 3.7 point gap indicates that partial visibility disrupts models' capacity to maintain accurate spatial representations, even when the underlying scene configuration remains largely unchanged.

5.3. Consistency and Plausibility Detection

Table 3 presents consistency scores measuring whether models provide identical answers to invariant questions across base and counterfactual variants, together with explicit plausibility detection accuracy and counterfactual comparison performance.

Consistency scores below 0.6 for all models indicate that MLLMs frequently alter their responses when they should remain invariant. For example, when queried whether a cup rests on a table, relocating an unrelated object elsewhere in the scene should not affect the answer, yet models modify their response in over forty percent of cases. This finding suggests that spatial representations are not robustly maintained across scene variations. Plausibility detection accuracy ranges from 48.3 percent for InstructBLIP to 61.2 percent for GPT-4V, with chance performance at 50 percent. Only GPT-4V significantly exceeds chance, and even then by a modest margin, indicating that explicit physical reasoning remains a substantial challenge. Baseline methods perform near chance, confirming that simple visual cues prove insufficient for detecting physical implausibility. Counterfactual comparison proves exceptionally difficult, with GPT-4V correctly identifying changed objects and transformation types only 37.5 percent of the time. Models frequently report no change or identify the incorrect object, suggesting that they lack mechanisms for detecting and localizing spatial transformations.

5.4. Impact of Scene Complexity

Table 4 analyzes performance scaling with object count in the scene, providing insight into how models accommodate increasing visual and relational complexity.

All models demonstrate consistent degradation with increasing object count. GPT-4V accuracy declines from 53.6 percent on scenes with four objects to 43.1 percent on scenes with eight objects, a relative decrease of 19.6 per-

Model	Vision	+Structure	Gain
GPT-4V	48.3 (2.8)	52.1 (2.9)	+3.8
LLaVA-1.5	41.7 (3.1)	44.8 (3.2)	+3.1
InstructBLIP	36.2 (3.3)	38.9 (3.4)	+2.7
Qwen-VL-Chat	38.9 (3.2)	41.5 (3.3)	+2.6

Table 5. Spatial question answering accuracy on counterfactual variants under vision-only versus vision-plus-structure conditions. Structural information provides modest yet consistent improvements across all models, with gains ranging from 2.6 to 3.8 percentage points.

cent. This finding suggests that current architectures struggle to maintain accurate spatial representations as relational complexity increases, consistent with limitations in attention mechanisms when numerous objects compete for representational capacity.

5.5. Input Modality Ablation

Table 5 compares model performance under vision-only versus vision-plus-structure conditions, assessing whether explicit scene graph information enhances spatial reasoning.

Providing explicit scene graph information enhances performance across all models, with gains ranging from 2.6 to 3.8 percentage points. GPT-4V exhibits the largest improvement, attaining 52.1 percent accuracy with structural information. However, even with explicit coordinates and relationships provided in text, models fail to achieve robust performance, suggesting that the core limitation resides in reasoning about spatial information rather than perception alone. Notably, performance gains from structural input are modest despite providing exact object coordinates and support relations. This finding suggests that models may not fully exploit explicit geometric descriptions, instead treating them as additional text rather than structured constraints. Future analysis could probe whether attention weights align with graph-provided spatial information or whether models largely ignore structural annotations.

5.6. Statistical Significance

Bootstrap tests with 1,000 resamples confirm that differences between models are statistically significant. Comparing GPT-4V to LLaVA on counterfactual question accuracy, the mean difference of 6.6 percentage points exhibits a 95 percent confidence interval from 3.8 to 9.4 percentage points with $p < 0.001$. All pairwise comparisons between MLLMs are significant at $p < 0.05$ except the comparison between InstructBLIP and Qwen-VL-Chat where $p = 0.12$, indicating comparable performance between these two models. Comparisons between MLLMs and baselines are all significant at $p < 0.001$.

5.7. Error Analysis

We conduct manual analysis of 200 error cases per model, totaling 800 errors examined, to identify systematic failure patterns across architectures. Spatial term confusion accounts for 32 percent of errors, with models frequently confusing left and right when objects are viewed from different perspectives. For example, when the camera rotates by ninety degrees, models often maintain egocentric spatial judgments rather than allocentric reasoning about world coordinates, responding based on image coordinates rather than true spatial relationships. Occlusion blindness appears in 28 percent of errors involving partially visible objects. When objects are partially occluded, models report them as absent in 24 percent of cases and misidentify their category in 18 percent of cases. This finding suggests that models lack robust amodal completion capabilities and rely substantially on visible surface features for object recognition. Physics violation insensitivity proves particularly striking, with models describing floating or intersecting objects as normal in 59 percent of implausible scenes. Even when explicitly queried about physical possibility, models frequently fail to detect violations, suggesting that training data rarely includes negative examples of physical implausibility. Change detection failure in counterfactual comparison tasks occurs in 37 percent of cases, with models either reporting no change or identifying the incorrect object. This finding indicates that models lack explicit mechanisms for comparing visual inputs and detecting spatial differences. Counting errors account for 23 percent of mistakes, with object counting accuracy declining from 84 percent in base scenes to 61 percent in counterfactual variants even when object counts remain constant. Spatial perturbations appear to disrupt basic perceptual grouping and enumeration. Within counterfactual comparison errors, 41 percent correspond to incorrect object identification, 33 percent correspond to correct object identification but incorrect transformation type, and 26 percent correspond to reporting no change despite a perturbation. This distribution indicates that models struggle both with localizing the modified object and with categorizing the transformation nature, suggesting absence of explicit scene-differencing mechanisms.

5.8. Why Do MLLMs Fail on Counterfactual Spatial Reasoning?

The consistent degradation across perturbation types suggests that current models lack stable geometric representations that persist across controlled transformations. We hypothesize three contributing factors. First, token-based multimodal architectures discretize spatial information into visual tokens optimized for semantic recognition rather than continuous geometric reasoning. While such representations capture object identity and coarse relations, they do not explicitly encode metric structure, support relations, or

rigid transformations in a geometrically consistent manner. Second, current MLLMs lack inductive biases for three-dimensional equivariance. Rotations, occlusions, and object relocations substantially alter pixel configurations, yet the underlying world structure may remain partially invariant. Without architectural mechanisms enforcing viewpoint or transformation consistency, models rely on appearance-level correlations rather than stable three-dimensional relational representations. Third, large-scale training corpora contain abundant descriptions of typical physical scenes but relatively few examples of counterfactual or physically implausible configurations. Consequently, models acquire strong priors over common spatial patterns but are not explicitly trained to detect violations of physical constraints or to maintain consistent world-state tracking under controlled perturbations. Together, these observations suggest that enhancing spatial intelligence in MLLMs may require architectural integration of structured three-dimensional representations, counterfactual training signals, or persistent world models rather than purely scaling data and parameters. These limitations may reflect a representational bottleneck in current token-based multimodal architectures. Visual encoders convert continuous spatial geometry into discrete token embeddings optimized for semantic discrimination rather than metric reasoning. As scene complexity increases, attention mechanisms must distribute capacity across more object tokens, potentially degrading relational precision. Without explicit three-dimensional inductive biases or equivariant constraints, learned representations may fail to preserve geometric structure under spatial perturbations. This finding suggests that architectural changes—not merely larger datasets—may prove necessary for robust spatial world modeling.

6. Discussion

The experimental results reveal several important findings regarding current MLLM capabilities for spatial reasoning. The substantial performance disparity between base and counterfactual scenes indicates that models appear to rely on appearance-level correlations rather than maintaining transformation-consistent geometric representations. When spatial configurations change in ways that should not affect certain judgments, models frequently alter their responses, suggesting that internal representations lack invariance under relevant transformations. The particularly diminished performance on physics violations suggests that current training regimes inadequately expose models to physically implausible configurations. Models learn to describe typical scenes but fail to detect when configurations violate basic physical constraints. This limitation carries significant implications for deploying these models in embodied artificial intelligence applications where detecting physical impossibility proves crucial for safe planning and execution. Comparison with baseline methods confirms that MLLMs substantially outperform traditional computer vision approaches on spatial reasoning, yet the performance gap narrows on counterfactual tasks. The Mask R-CNN baseline, despite its simplicity, achieves 31.5 percent accuracy on counterfactual scenes, demonstrating that basic geometric heuristics provide a foundation upon which MLLMs build but do not fundamentally transcend. We did not include a human baseline in this study. Informal inspection suggests that most counterfactual and plausibility questions are straightforward for human observers; however, systematic human evaluation would provide a valuable upper-bound reference for future investigation. The modest improvements from explicit scene graph information suggest that the core limitation resides in reasoning about spatial relationships rather than perceiving them. Even when provided with exact coordinates and relations in text, models struggle to maintain consistent representations and detect implausible configurations. Several limitations of this study warrant acknowledgment. The use of synthetic scenes, while enabling controlled perturbation, may not fully capture the complexity of real-world environments with natural textures, lighting variations, and object diversity. Performance on real images could differ substantially from our synthetic evaluation. The static nature of our scenes precludes evaluation of dynamic spatial reasoning involving object motion or agent interaction. Our perturbation taxonomy, while systematic, may not encompass all forms of counterfactual variation relevant to spatial intelligence, such as material property changes or lighting transformations. Future work would be to extend evaluation to dynamic scenes with object motion, real-world images with known three-dimensional ground truth captured via depth sensors or multi-view reconstruction, additional perturbation types including material and lighting variations, fine-tuned models trained on counterfactual data to assess whether these limitations can be addressed through targeted training, and embodied agents that must reason about counterfactuals during planning and execution.

7. Conclusion

We introduced SCWM, a framework for evaluating multimodal spatial reasoning under controlled 3D perturbations. Evaluating four MLLMs and two baselines on 2,500 base scenes and 7,500 counterfactual variants reveals significant limitations: performance drops 22.9 percentage points from standard to counterfactual tasks, physics violation accuracy reaches only 43.8% for the best model, consistency scores remain below 0.6, and counterfactual comparison accuracy falls below 40%. These findings demonstrate that scaling multimodal architectures alone is insufficient explicit structural grounding, counterfactual supervision, or integration of geometric world models may be necessary.

References

[1] OpenAI, G. P. T. (2023). 4V (ision) System Card https://cdn. openai. com/papers. GPTV_System_Card. pdf. 1, 5

[2] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in neural information processing systems, 36, 34892-34916. 1, 5

[3] Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., ... & Hoi, S. (2023). Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems, 36, 49250-49267. 1, 5

[4] Xue, Q., Liu, W., Wang, S., Wang, H., Wu, Y., & Gao, W. (2025). Reasoning Path and Latent State Analysis for Multi-view Visual Spatial Reasoning: A Cognitive Science Perspective. arXiv preprint arXiv:2512.02340. 1

[5] Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2901-2910). 1, 2

[6] Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018). Embodied question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-10). 1, 2

[7] Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019, May). Counterfactual visual explanations. In International Conference on Machine Learning (pp. 2376-2384). PMLR. 2

[8] Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L., & Girshick, R. (2019). Phyre: A new benchmark for physical reasoning. Advances in Neural Information Processing Systems, 32. 2

[9] Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., & Xia, F. (2024). Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14455-14465). 2

[10] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969). 4

[11] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR. 4

[12] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., ... & Zhu, T. (2023). Qwen technical report. arXiv preprint arXiv:2309.16609. 5