MULTIMODAL LANGUAGE MODELS CANNOT SPOT SPATIAL INCONSISTENCIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial consistency is a fundamental property of the visual world and a critical requirement for models that aim to understand or generate realistic scenes. Yet, despite their impressive capabilities, today's multimodal large language models (MLLMs) often fail to reason about 3D geometry across views. We introduce a new task that directly tests this ability: given two views of the same scene, identify which object violates 3D consistency. To create data for this task at scale, we propose a simple, fully automatic method that constructs realistic pairs of inconsistent images from multi-view captures. The method uses object segmentation, inpainting, and cross-view replacement to introduce controlled geometric inconsistencies without manual annotation. Using this approach, we build a dataset and evaluate several state-of-the-art MLLMs, including GPT-5, Gemini 2.5 Pro, and Qwen3 VL 8B. Humans outperform all models by a large margin, revealing that current systems lack robust spatial reasoning. Moreover, fine-tuning an MLLM such as Qwen3 VL 4B on our task not only improves its accuracy and generalization but also enhances performance on other benchmarks like BLINK. Our findings underscore spatial consistency as a key frontier in multimodal reasoning and present a practical framework for advancing geometric understanding in next-generation MLLMs. Our code and benchmark will be available publicly.

1 INTRODUCTION

Spatial consistency is a defining property of the physical world. As a camera moves through a static scene, objects change appearance in predictable ways governed by geometry, depth, and viewpoint. Their shapes, scales, and occlusions vary smoothly and coherently across views. Humans are sensitive to violations of these relationships: any configuration that cannot physically exist immediately feels "off". For multimodal large language models (MLLMs) and generative systems that aim to understand or synthesize realistic scenes or videos, maintaining such consistency is essential.

However, even the most capable multimodal models, such as GPT-5, Gemini, and Qwen, struggle with reasoning about 3D structure and spatial coherence. While these models can describe individual views with impressive accuracy, they often fail to recognize when two views of a scene are mutually inconsistent. This gap highlights a broader issue: today's MLLMs demonstrate surface-level 3D visual understanding, but lack a robust internal model of spatial geometry. As generative models increasingly rely on MLLMs to judge realism, control synthesis, or enforce geometric consistency, this limitation becomes detrimental.

To study this problem systematically, we focus on a simple but revealing question: *can a model identify spatial inconsistency between two images of the same scene?* This task directly probes whether a model understands 3D structure rather than merely memorizing visual statistics. A key challenge, however, is obtaining high-quality pairs of images that differ only by a physically inconsistent object. Existing 3D rendering or view-synthesis methods, such as Zero123 (Liu et al., 2023) or SEVA (Zhou et al., 2025), can generate alternative viewpoints, but often produce artifacts or unrealistic geometry that confound evaluation. We instead develop a simple, lightweight, fully automatic procedure that uses real multi-view scenes to create natural yet inconsistent image pairs. The algorithm requires no manual annotation and operates in near real-time, serving purely as a scalable tool to enable this study.

Given three views of the same static scene, we select the first as a reference, segment an object from the second and third views, and perform three steps: (1) remove the selected object from the second view and inpaint the missing region using an off-the-shelf inpainting model, (2) copy and paste the segmented object from the third view into the second, and (3) form a pair consisting of the first (reference) and the second (modified) view. Because the camera poses differ across the three views, the pasted object naturally acquires a 3D pose inconsistent with the reference view. The resulting pair preserves photorealism while introducing a controlled geometric violation, precisely the type of inconsistency that humans can easily detect but confuses most recent MLLMs.

Using this method, we construct a dataset of spatially inconsistent image pairs. Each pair includes ground truth annotations specifying which object violates consistency. We formulate a simple evaluation task: given two images of a scene, identify the inconsistent object. This task serves as a direct test of spatial reasoning ability in both humans and MLLMs. We evaluate several state-of-the-art models, including GPT-5, Gemini, and Qwen3 VL, and find that human participants achieve significantly higher accuracy than all tested models (Section 3 and Appendix C.1). These results demonstrate that the generated image pairs provide a challenging and informative benchmark for diagnosing spatial reasoning gaps in current multimodal systems.

Beyond evaluation, we show that the same data can be used to improve MLLM performance. Fine-tuning an MLLM such as Qwen to identify inconsistent objects delivers strong gains on our dataset and also boosts zero-shot performance on other 3D tasks, including relative depth estimation, correspondence, and camera motion estimation (Appendix C.2). This suggests that learning to detect geometric inconsistencies encourages models to develop more structured internal representations of 3D scenes. Consequently, our data generation framework can serve as both a training signal and an evaluation resource for improving 3D reasoning in future multimodal language models.

Spatial inconsistency is also a persistent challenge in video generation. Many recent video synthesis models produce sequences that appear realistic frame by frame but subtly violate 3D geometry over time, leading to objects that drift, distort, or change shape across views. The principles underlying our task naturally extend to this temporal setting: similar notions of geometric plausibility and cross-view consistency apply across frames. Fine-tuned MLLMs that learn to detect spatial inconsistency could therefore act as automated consistency checkers for video generation pipelines, serving as part of a reward function or evaluation metric during training and inference. We leave the application of our approach to video generation models as an exciting direction for future work (Appendix C.3).

In summary, we present a simple, scalable algorithm to automatically generate spatially inconsistent image pairs, a dataset and evaluation task that reveal clear gaps between human and model reasoning, and initial evidence that training on such data improves general spatial understanding. Our findings highlight spatial consistency as a key frontier for multimodal reasoning and provide a practical path toward closing this gap in future models.

2 Synthesizing Spatial Inconsistencies

2.1 Problem Setting

Our goal is to construct pairs of views of a static scene in which a single object violates the spatial constraints implied by camera motion. Given multi-view imagery of a scene, we aim to generate two images that differ only in the 3D-consistency of one object. While one could synthesize such pairs using novel-view generation models (e.g., Zero123-style approaches (Liu et al., 2023; Zhou et al., 2025)), these methods are often slow, may introduce unintended artifacts, and may embed model-specific biases into the benchmark.

Instead, we adopt a lightweight cut-and-paste strategy that operates directly on observed views. This approach scales to large datasets, runs in almost 5 pairs per second on a single GPU, and preserves natural image statistics. Most importantly, by sourcing the inserted object from a true novel view, our method produces a controlled 3D violation that alters only the object's pose without unintentionally modifying its structure or texture. The resulting pairs are photorealistic and contain a single, well-defined geometric inconsistency.

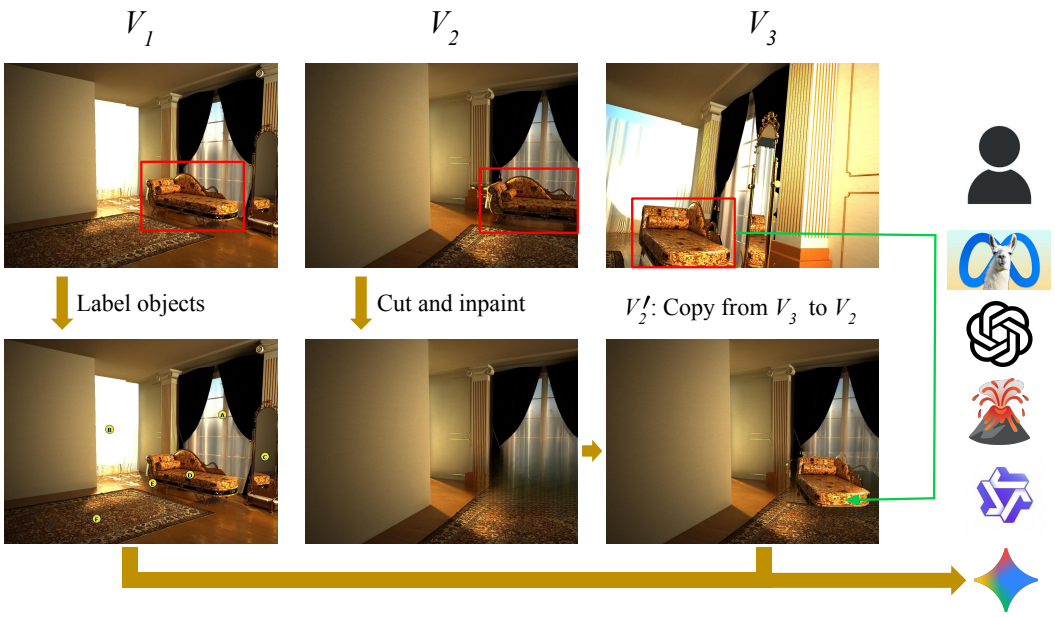

Figure 1: **Synthesizing spatially inconsistent image pairs from multi-view data.** Given three views (V_1, V_2, V_3) of the same static scene, we (1) select an object O visible in all views, (2) Erase O in view V_2 and inpaint to obtain a clean background, and (3) paste the instance of O from view V_3 back into V_2 at its original footprint. Because V_3 is captured from a different camera pose, the pasted object has an appearance that is incompatible with the 3D geometry implied by $V_1 \rightarrow V_2$, while all other objects remain consistent. This yields realistic image pairs with controlled spatial violations and no manual annotation. We label objects in V_1 for our forced-choice evaluation.

2.2 Source Data and Triplet Selection

We build on the Hypersim dataset (Roberts et al., 2021), which provides multi-view indoor scenes with tracked instance-level masks. For each scene, we sample three frames (V_1, V_2, V_3) and select an object O visible in all three views. To avoid degenerate cases where the views are nearly identical, we impose that at most 75% of the objects overlap between (V_1, V_2) and between (V_2, V_3). We further restrict selection of O to objects occupying between 5% and 10% of the image area to ensure sufficient visual detail while avoiding large objects that lead to inpainting artifacts or small objects that lack textural cues. Additionally, we enforce the pixels of O in V_3 must project to at 40% of the area of O in V_2, ensuring that V_2 and V_3 have enough overlap.

2.3 Violating Spatial Consistency

Given a selected triplet (V_1, V_2, V_3, O), we construct an edited version V_2' in which only object O violates the geometric constraints consistent with the transition of camera view point $V_1 \rightarrow V_2$. The process consists of two steps:

1. **Removing O in V_2.** We erase O from view V_2 by inpainting its bounding box (enlarged by 20% to include context) using the LaMa inpainting model (Suvorov et al., 2021). This produces an object-free background that remains visually coherent. We inpaint the bounding box instead of the segmentation mask to ensure contour details of the object are also erased.
2. **Reinserting O from V_3.** We extract O from V_3 using its instance segmentation mask and paste it into the inpainted region of V_2'. The pasted object preserves its aspect ratio and is centered to match the footprint of O in V_2. All pixels belonging to other object instances within the inpainted area are restored to maintain continuity and correct occlusion ordering.

Because the camera poses differ across (V_1, V_2, V_3), the appearance of O as seen from V_3 will generally not be geometrically compatible with its appearance in V_1. This edit therefore introduces

a precise 3D inconsistency while leaving all remaining objects unchanged. Models with strong 3D geometric understanding must detect mismatches in scale, orientation, shading, correspondence, and occlusion patterns across the two views.

Model	Reasoning	Overall	Model	Reasoning	Overall
Random Chance		7.9	Random Chance		7.9
Human		84.8	Human		84.8
Gemma 3 12B		8.5	Llama 3.2 Multimodal 11B		23.4
Idefics3 8B		8.9	Qwen3 VL 4B		24.7
Idefics2 8B		11.9	Qwen3 VL 8B Thinking	✓	25.2
GPT-5 Nano		15.3	Qwen3 VL 8B Instruct		27.6
Qwen2.5 VL 7B		15.9	Gemini 2.5 Pro (HR)	✓	28.9
InternVL 3.5 8B		16.1	Gemini 2.5 Pro (LR)	✓	29.4
LLaVA OneVision 1.5 8B		16.6	Gemini 2.5 Pro (MR)	✓	29.4
Gemini 2.5 Flash		17.6	GPT-5 (HR)	✓	30.2
SpaceQwen2.5 VL 3B		18.2	GPT-5 (MR)	✓	31.4
GPT-4o		19.0	GPT-5 (LR)	✓	**34.2**

Table 1: **Overall accuracy (%) on the spatial inconsistency identification task.** We report overall accuracy for a random-choice baseline, humans, and a range of multimodal language models (sorted by increasing overall accuracy). A checkmark indicates runs using an explicit reasoning budget (HR/MR/LR denote high/medium/low reasoning budget, respectively). Consistent with the benchmark findings, humans substantially outperform all tested models, and higher reasoning budget does not necessarily yield higher accuracy.

3 How Many Spatial Inconsistencies Can MLLMs Catch?

As shown in Table 1, humans can easily spot spatial inconsistencies with relative ease at 84.8% accuracy. However, most models struggle with our task, with the very best (GPT-5 with "low" effort reasoning) achieves only 34.1%. This suggests that learning to spot inconsistencies still remains a perceptual gap in multimodal language models, even as accuracy in many 3D tasks such as depth or correspondence approaches human capabilities (Bai et al., 2025b). Surprisingly, models such as SpaceQwen, which have been finetuned to excel at spatial reasoning tasks, still fail to spot inconsistencies. This may suggest that existing methods to teach models spatial understanding still leave large gaps in their 3D reasoning capabilities. Additionally, as is shown by the result corresponding to Qwen3-VL, Gemini 2.5 Pro, and GPT-5, increasing test-time compute *does not* lead to improved performance. In fact, we consistently find that increasing reasoning effort often degrades model performance or leads to no further improvement. This finding is counterintuitive to the promise of visual reasoning and merits further investigation. We also explore the effects of several factors such as depth, lighting, and physical plausibility in the appendix.

4 Conclusion

We introduced a new task for evaluating 3D spatial reasoning in multimodal large language models (MLLMs), together with a simple and scalable method for synthesizing cross-view inconsistencies. Our results demonstrate that state-of-the-art MLLMs struggle to detect even single-object geometric violations that humans find obvious, highlighting a persistent gap in 3D understanding. Fine-tuning on our automatically generated data improves performance on both our task and related 3D tasks we provide no additional supervision for. This indicates that inconsistency detection serves as an effective proxy objective for learning 3D priors. We believe that our task and data generation pipeline offer a practical approach for advancing spatial reasoning in future multimodal systems and for enabling applications such as robotics, scene understanding, and video generation to identify failures in simulating or generating the physical world. We hope our work inspires others to look beyond teaching multimodal models to simply describe the visual world, and to build a visual understanding that fundamentally understands the physical and spatial constraints of our world.

REFERENCES

Anonymous. Controlling video generation with vision language models. In *Submitted to The Fourteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6SC61wyq8w. under review.

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

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

Anand Bhattad, Konpat Preechakul, and Alexei A Efros. Visual jenga: Discovering object dependencies via counterfactual inpainting. *arXiv preprint arXiv:2503.21770*, 2025.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024. URL https://arxiv.org/abs/2401.12168.

Ziming Cheng, Binrui Xu, Lisheng Gong, Zuhe Song, Tianshuo Zhou, Shiqi Zhong, Siyu Ren, Mingxiang Chen, Xiangchao Meng, Yuxin Zhang, et al. Evaluating mllms with multimodal multi-image reasoning benchmark. *arXiv preprint arXiv:2506.04280*, 2025.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

Erik Daxberger, Nina Wenzel*, David Griffiths*, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and Peter Grasch. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *ICCV*, 2025. URL https://arxiv.org/abs/2503.13111.

Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 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, 2017.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. URL https://arxiv.org/abs/2405.02246.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation. 2024a.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024b.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024c.

Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. *arXiv preprint arXiv:2505.01481*, 2025.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023.

Grace Luo, Jonathan Granskog, Aleksander Holynski, and Trevor Darrell. Dual-process image generation. In *ICCV*, 2025.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Josh Tenenbaum, Dan Yamins, Judith Fan, and Kevin Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *Advances in Neural Information Processing Systems*, 36:67048–67068, 2023.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11998–12008, 2023.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Jianrui Zhang, Cai Mu, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos.

Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint*, 2025.

APPENDIX

A QUALITATIVE EXAMPLES

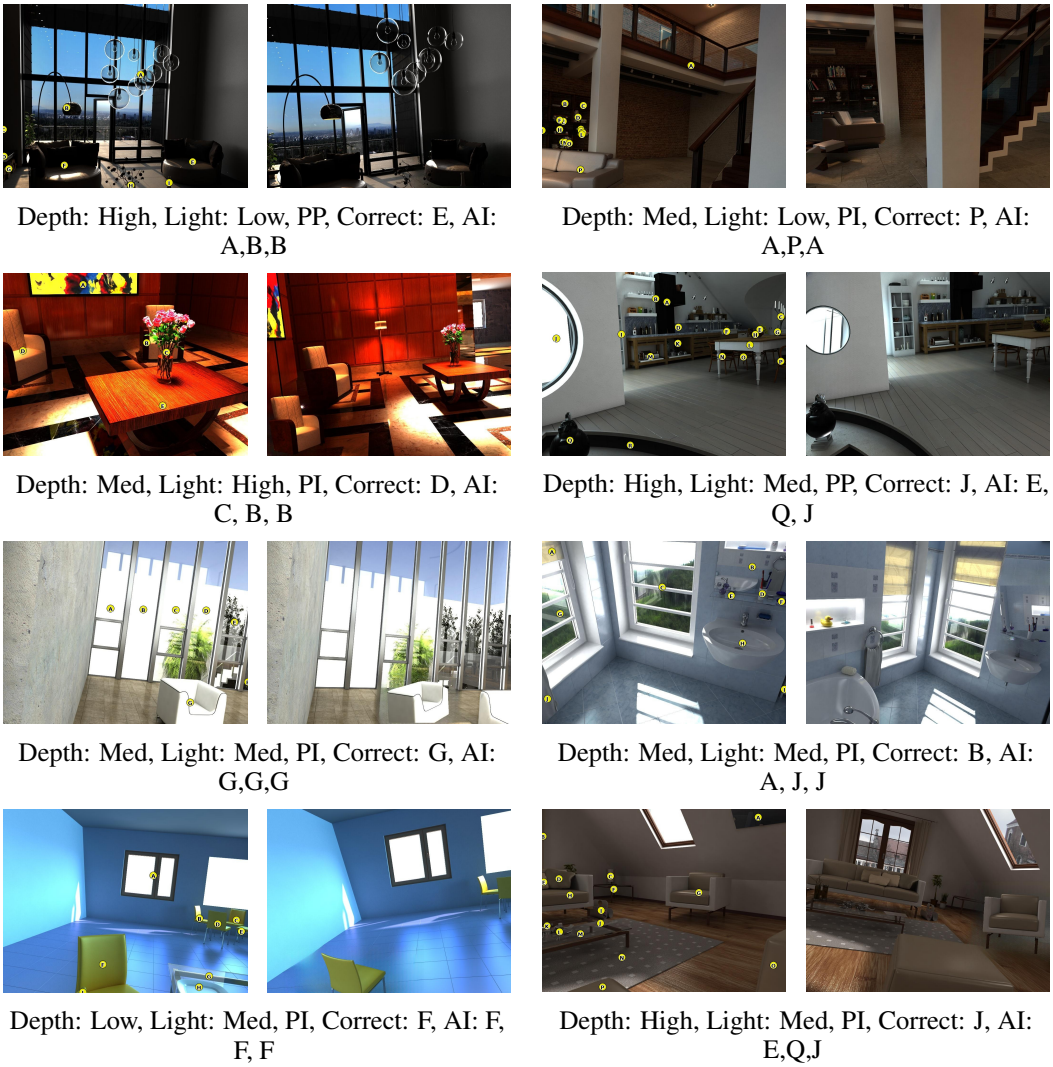

Figure 2: **Example figures with labels and predictions.** AI labels are ordered as GPT-5 (Low), Gemini 2.5 Pro (Medium) and Qwen3 VL 8B Thinking. Zoom in to see labels in detail.

B EXPERIMENTAL DETAILS

B.1 MODELS EVALUATED

We evaluate a comprehensive suite of Multimodal Large Language Models (MLLMs) spanning various architectures, sizes, and training strategies. Our selection includes leading proprietary frontier models, specifically the **GPT-5** family, **GPT-4o**, and **Gemini 2.5** (Pro and Flash) under a wide range of reasoning budgets. We also assess state-of-the-art open-weights models, including **Llama 3.2 Multimodal** (et al., 2024) **Qwen 2.5 and 3 VL Instruct and Thinking** (Bai et al., 2025b;a), **Gemma 3** (Team et al., 2025), **Idefics2 and 3** (Laurençon et al., 2024), **LLaVA OneVision 1.5** (Li et al., 2024b), and **InternVL 3.5** (Wang et al., 2025). Finally, to investigate the impact of domain-specific fine-tuning and inference-time reasoning, we include spatially specialized variants such as

SpaceQwen (Chen et al., 2024). For GPT-5 and Gemini 2.5 Pro, we also report low (LR), medium (MR) and high (HR) reasoning efforts.

B.2 Benchmark Construction

Following established practice to validate the benchmark (Fu et al., 2024), we manually inspect all pairs. Because there are only 452 valid scenes in Hypersim, we manually remove similar pairs and obtain a final set of 615 inconsistent frame pairs. Note that this manual step is not needed when generating training data as similar pairs can be sees as data augmentation. To probe MLLMs' spatial reasoning, we convert each pair into a forced-choice question: given views (V_1, V_2'), identify the object that violates 3D consistency.

Object labeling. We label all objects in view V_1 larger than 1000 pixels (assuming 1024×768 image size) using a raster-scan ordering from top-left to bottom-right, assigning letter labels `A...Z` following prior work (Fu et al., 2024). This labeling scheme allows models to refer to objects unambiguously in natural language. If more than 26 objects satisfy the size threshold, we keep the largest 26 while ensuring that O is included. If too few objects qualify, we slightly relax the threshold but never include instances smaller than 300 pixels that lack meaningful visual detail. These labels define the answer space for both human and model evaluation.

Prompt and evaluation. At the test time, the model receives the labeled V_1 and the edited V_2' along with the following prompt:

```
Here are two photos of a static scene from different views.
However, in between taking the photos, I edited the second image
such that, for one object, its positioning is inconsistent with
the camera motion between the two frames.  Which letter marks the
modified object?  Please use the labels from the first image.
Format your reply exactly as:
Final Answer:  <LETTER>
...where <LETTER> is ONE capital letter A to Z only.  Do not write
anything else.
```

We parse the final output and mark the item correct if the letter matches the single edited object. Our primary metric is the average accuracy across pairs.

Human evaluation protocol. We evaluate both humans and multimodal language models under the same forced-choice protocol. Given an image pair, participants and models must indicate which image contains the 3D-inconsistent object. No additional textual hints about the manipulated object or category are provided. We measure accuracy over all 615 pairs. Section 3 reports detailed results broken down by depth, lighting, and physical plausibility.

C Additional Experiments

C.1 How does the scene affect spatial inconsistency spotting?

Depth and lighting factors. Because Hypersim provides ground-truth depth and per-pixel illumination, we can directly examine how geometric and photometric factors influence performance. For each inconsistent pair, we compute the depth of the modified object and the average scene brightness, then partition the dataset into 3 equally sized bins based on the empirical distributions: *close*, *medium*, and *far* for object depth, and *dark*, *medium*, and *bright* for lighting. For object depth we use the monocular depth within frame V_1, and for lighting we use the luminance averaged across V_1 and V_2. This categorization enables systematic comparisons of human and model accuracy under different viewing conditions. Surprisingly, we find that most multimodal models perform substantially better on objects that are *far* relative to those that are *close*, with differences exceeding 15 percentage points, whereas human accuracy is highest for *medium*-distance objects and remains stable for *close* and *far* ones. This pattern suggests that current MLLMs exhibit geometric biases related to

Model	Object Depth			Scene Lighting			Plausibility	
	Close	Medium	Far	Dark	Medium	Bright	PI	PP
Random Chance	7.2	8.0	8.4	7.7	8.1	7.8	7.9	7.7
Human	83.3	87.3	83.8	82.3	85.4	86.8	86.3	80.1
Gemma 3 12B	13.2	7.3	4.9	7.8	11.2	6.3	9.4	5.4
Idefics3 8B	8.8	10.7	7.3	8.8	9.8	8.3	7.7	12.8
Idefics2 8B	9.8	12.2	13.7	12.2	13.2	10.2	12.4	10.1
SpaceOm	19.0	15.6	8.8	14.6	16.6	12.2	15.0	12.8
GPT-5 Nano	14.6	17.6	13.7	16.1	15.1	14.6	16.1	12.8
Qwen2.5 VL 7B	11.7	12.7	23.4	19.0	19.0	9.8	17.1	12.2
InternVL 3.5 8B	14.2	13.7	20.5	16.6	16.1	15.6	16.5	14.9
LLaVA OneVision 1.5 8B	11.2	14.6	23.9	18.1	19.0	12.7	17.3	14.2
Gemini 2.5 Flash	18.1	18.1	16.6	18.1	21.0	13.7	18.4	14.9
SpaceQwen2.5 VL 3B	23.4	16.6	14.6	20.5	19.5	14.6	19.3	14.9
GPT-4o	16.1	17.1	23.9	22.9	14.2	20.0	19.9	16.2
Llama 3.2 Multimodal 11B	30.7	23.4	16.1	21.0	25.4	23.9	24.2	21.0
Qwen3 VL 4B	31.2	24.9	18.1	23.9	25.4	24.9	25.9	21.0
Qwen3 VL 8B Thinking	18.1	24.4	33.2	31.2	27.8	16.6	26.8	20.3
Qwen3 VL 8B Instruct	26.8	25.4	30.7	27.8	32.2	22.9	29.8	21.0
Gemini 2.5 Pro (HR)	28.3	30.2	28.3	29.8	33.7	23.4	29.1	28.4
Gemini 2.5 Pro (LR)	**31.7**	29.3	27.3	30.2	31.7	26.3	30.6	25.7
Gemini 2.5 Pro (MR)	30.7	**32.2**	25.4	30.7	32.7	24.9	29.8	28.4
GPT-5 (HR)	28.3	25.9	36.6	28.8	32.2	29.8	30.6	29.1
GPT-5 (MR)	23.4	30.2	40.5	**32.2**	33.2	28.8	33.2	25.7
GPT-5 (LR)	27.3	29.8	**45.4**	30.2	**36.6**	**35.6**	**35.3**	**30.4**

Table 2: **Performance breakdown (% accuracy) by depth, lighting, and physical plausibility.** We report accuracy for identifying the single spatially inconsistent object, stratified by inconsistent object depth (Close/Medium/Far), average pair brightness (Dark/Medium/Bright), and whether the augmented object is physically plausible (PI vs. PP). While humans remain comparatively robust across conditions, all models show substantial sensitivity to depth and lighting, and plausibility also impacts accuracy.

depth perception. Lighting reveals similar discrepancies: humans are consistently more accurate in *bright* scenes, while models tend to perform best under *medium* illumination, potentially reflecting biases inherited from visual instruction tuning data.

Physical plausibility of poses. Not all 3D-inconsistent poses are equally salient. Some pose manipulations are obviously impossible in everyday scenes (e.g., a chair with two legs hovering above the ground), while others are subtle geometric violations that are harder to notice at a glance. To capture this variation, we categorize each pair by the *physical plausibility* of the inconsistent pose. We estimate the camera transformation between frames V_2 and V_3 of the underlying Hypersim sequence and use the induced roll to approximate whether the inserted object appears stably supported or not. Pairs with less than 5° roll are labeled as *physically plausible* (PP), meaning the object appears roughly stable under gravity, while those with larger roll are labeled as *physically implausible* (PI). This allows us to analyze how the physical compatibility of the inserted pose affects detection difficulty. We find that both humans and models more easily identify inconsistencies in PI pairs, as these poses tend to be unusual in real-world settings. Interestingly, while models with limited test-time compute perform better overall on our task, models with greater test-time compute exhibit only a small performance gap between PP and PI pairs, suggesting that additional compute helps mitigate sensitivity to physical plausibility.

C.2 Does Learning to Spot Inconsistencies Improve Broader 3D Perception?

Given the poor zero-shot accuracy of state-of-the-art multimodal models, a natural question arises: is our task fundamentally beyond the capabilities of current architectures, or do existing models simply lack exposure to the kind of geometric violations we introduce? Moreover, if a model can be trained to detect these inconsistencies, does it merely learn a narrow heuristic, or does it acquire a more generalizable internal representation of 3D structure?

Model	Base Model	Finetuned	Min	Max
Our Benchmark	24.7	66.8±1.5	65.1	68.7
Multiview Reasoning	37.6	39.9±1.4	38.4	42.1
Visual Correspondence	80.2	81.7±0.7	80.8	82.6
Relative Depth	84.7	86.9±1.1	85.5	87.1

Table 3: **Fine-tuning on our benchmark improves performance on external perception and reasoning tasks.** We report accuracy (%) of the base model and the same model fine-tuned on our spatial inconsistency dataset across several BLINK subtasks: Multiview Reasoning, Visual Correspondence, and Relative Depth. For each task, we show the mean and standard deviation over 5 runs, as well as the minimum and maximum accuracy. The gray row reproduces our benchmark performance for reference. Fine-tuning on spatial inconsistency provides consistent, though modest, gains on multiview and depth reasoning, indicating that learning to detect geometric violations transfers beyond our specific evaluation. We report the 95% conf. int. and min and max accuracies over 5 training runs.

To answer these questions, we fine-tune Qwen3-4B using a scaled-up version of our dataset. We automatically generate a training set of 19K inconsistent (V_1, V_2') pairs from Hypersim by extending our cut-and-paste pipeline with additional constraints on object depth and relative view pose, ensuring that augmented examples are neither trivial nor unrealistic. We apply LoRA (Hu et al., 2021) with rank 16 to reduce overfitting. For Table 3, we train for 80 iterations with minibatch size 128 on the full training set (roughly half an epoch). The training started overfitting after this since the data is synthesized from only 452 scenes. We believe that one can use our method on a larger set of 3D scenes like (Ling et al., 2024) to scale up the generated data with more diversity. For Table 4, we hold out 20% of scenes to build a disjoint test set and fine-tune for one epoch on the remaining scenes.

In-domain generalization. As shown in Table 4, fine-tuning yields a substantial improvement in performance: Qwen3 VL-4B jumps from 24.7% to 69.2% accuracy on entirely unseen scenes. This corroborates that the task is learnable and that our synthetic pairs provide a consistent and meaningful signal for the model to acquire a generalizable notion of 3D violations, rather than memorizing scene layouts or object textures.

Transfer to broader 3D benchmarks. The more important question is whether this training transfers beyond our task. We evaluate the fine-tuned model on the BLINK benchmark (Fu et al., 2024), focusing on categories most closely tied to spatial geometry: "Multiview Reasoning", "Visual Correspondence", and "Relative Depth". As summarized in Table 3, training on our inconsistency-spotting task yields consistent improvements across all three metrics. We observe gains of approximately 2.3 percentage points in both "Multiview Reasoning" and "Relative Depth", and 1.3 percentage points in "Visual Correspondence". These improvements are notable because the model receives no direct supervision on depth estimation, correspondence matching, or camera motion. Instead, identifying *which* object breaks cross-view consistency implicitly encourages the model to learn depth (for scale compatibility), correspondence (for appearance matching), and also camera motion understanding.

Together, these findings suggest that our inconsistency-spotting task provides an effective proxy objective for learning 3D structure. By leveraging our scalable, annotation-free data generation pipeline, we can inject meaningful 3D priors into multimodal models without relying on costly 3D supervision, handcrafted geometric annotations, or specialized rendering engines.

C.3 Do VQA Judges Notice Spatial Inconsistencies?

Despite struggling at directly identifying inconsistent objects, modern MLLMs are widely used as VQA-style judges for evaluating image and video generation. We therefore ask: do these judges notice our 3D inconsistencies when used in their standard Yes/No evaluation protocol? Given a set of frames and a textual description, the model is asked a constrained question such as: *"Does this video show* {`caption`}*? Please answer Yes or No."* The alignment score is then estimated from the probability assigned to the `Yes` token. In practice, this setup is used not only to measure

Model	Base Model	Finetuned
Our Benchmark, disjoint set	24.7	69.2

Table 4: **Fine-tuning on our spatial inconsistency task yields large gains on a disjoint test split.** Accuracy (%) of the base vision–language model versus the same model fine-tuned on our benchmark, evaluated on a held-out set of image pairs with no scene overlap. Performance jumps from 24.7% to 69.2%, showing that the task is learnable and that targeted training substantially improves spatial reasoning on new scenes.

Model	Pairwise	Pearson	Kendall
GPT–4o	50.1	-0.14	0.32

Table 5: **VQAScore evaluation of GPT–4o on our spatial inconsistency benchmark.** We report the pairwise accuracy (%) of VQAScore (Lin et al., 2024) when asked to score both the physically consistent image and the manipulated one, along with Pearson and Kendall correlations between VQAScore and human accuracy across image pairs. The near-chance pairwise score and weak correlations indicate that the SOTA VQA-based judge poorly aligns with human assessments of spatial consistency.

semantic alignment, but also as a proxy for visual plausibility: if two candidates match the prompt, the expectation is that the judge will assign a higher `Yes` score to the one that is more realistic.

Recent benchmarks (e.g., GenAI-Bench (Li et al., 2024a)) support this intuition by showing that VQA-style judges can successfully rank images that all satisfy the prompt but differ in fidelity and artifacts, suggesting some sensitivity to realism beyond pure text alignment. We hypothesize that this sensitivity to realism does *not* extend to 3D understanding, and thus may be unable to distinguish videos where object poses are modified such that the frames are impossible in a static scene.

We use our constructed pairs to test this assumption directly. For each multiview set (V_1, V_2, V_2'):

1. We first generate a caption C_1 from frame V_1 alone (without labels) using GPT-5.
2. We treat (V_1, V_2) and (V_1, V_2') as the frames sampled from two candidate videos depicting the same static scene.
3. We compute the accuracy as the fraction of items where the unmodified pair (V_1, V_2) receives a higher `Yes` score than the manipulated pair (V_1, V_2').

Assuming GenAI-Bench's finding holds, a judge that truly understands scene geometry should consistently prefer (V_1, V_2), the unmodified pair over (V_1, V_2') which is spatially inconsistent. However, as shown in Table 5, GPT-4o (the current SOTA for VQAScore eval) performs near random chance. This suggests that VQA judges does not spot spatial inconsistencies when ranking alignment of generated videos and provided captions. Additionally, when utilizing VQA judges to *control* diffusion-based generators (Luo et al., 2025; Anonymous, 2025), they may amplify these spatial inconsistencies due to their inability to identify them.

D Related Works

D.1 Evaluating Perception in Multimodal Language Models

Since the early days of VQA, many have questioned the perceptual limits of multimodal models. For example, the CLEVR dataset (Johnson et al., 2017) was introduced as a diagnostic to test whether models were learning superficial statistics or genuinely understanding low-level spatial and compositional relationships. This concern has re-emerged with general-purpose MLLMs, which excel at semantic tasks but often fail at core perception. A crucial line of work, exemplified by BLINK (Fu et al., 2024), found that many MLLMs "can see but not perceive." They showed that while models can solve high-level VQA, they fail at basic vision tasks like correspondence or object localization that cannot be solved with captions alone. These findings spurred a new generation of more rigorous benchmarks.

To probe a model's intuitive "world model," benchmarks like Physion++ (Tung et al., 2023) and PhysBench (Chow et al., 2025) evaluate understanding of physical properties and dynamics. Similarly, to test reasoning across multiple images, benchmarks such as MVBench (Li et al., 2024c), MMRB (Cheng et al., 2025), and MM-Spatial (Daxberger et al., 2025) evaluate a model's ability to aggregate spatial and semantic information from different viewpoints. While these works ask models to describe physical scene properties (e.g., stability, mass) or temporal details (e.g., "what is the person doing after cleaning?"), we isolate and evaluate a model's understanding of *projective geometry*. Additionally, we test not *what* is in a scene, but whether the scene's geometry is *fundamentally possible*.

D.2 Counterfactual Scene Understanding

A robust visual system should not only recognize what is present but also reject what is impossible, and counterfactual evaluation has become crucial to identify flaws in MLLMs visual representations. Early foundational works in this domain, such as Winoground (Thrush et al., 2022) utilize semantic counterfactuals (e.g., swapping object-attribute bindings or word order) to demonstrate that models often ignore syntax and relational logic in favor of bag-of-words matching. This line of inquiry has since expanded to "hard negatives," with benchmarks like CREPE (Ma et al., 2023) and EqBen (Wang et al., 2023) testing a model's ability to distinguish between minimally different captions or images.

More recently, this paradigm has shifted toward visual and physical counterfactuals, where images are manipulated to contradict physical laws rather than just textual descriptions. HallusionBench (Guan et al., 2024) introduced a suite of 2D visual illusions to probe whether MLLMs effectively hallucinate details when faced with confusing visual inputs. In the video domain, VideoHallu (Li et al., 2025) and Vinoground (Zhang et al.) generate synthetic videos that violate physical principles (e.g., gravity, object permanence) or temporal logic, identifying a critical gap in dynamic causal reasoning. Similarly, Visual Jenga (Bhattad et al., 2025) introduces a new task to evaluate a generative model's notion of scene stability and object dependence by asking the model to progressively remove objects from the scene without deforming it or creating a physical inconsistency.

However, these existing counterfactual benchmarks largely focus on semantic plausibility (is the correct object present?) or dynamic plausibility (did it move correctly over time?). Our work focuses on creating multiview counterfactuals that do not rely on generative models to synthesize the inconsistency.