

Geometric Consistency in Multimodal Visual Media: A Longitudinal Analysis of Spatial Reasoning and Scaling Behavior

Pratima Tamrakar¹, Rishav Jha², Suresh Kumar Sahani^{*3}

¹Science and Environment Education, Sanothimi Campus, T.U., Nepal

tamrakarpr@gmail.com

²Department of Mathematics, MIT Campus, Janakpurdham, Nepal

jharishav036@gmail.com

^{*3}Faculty of Science, Technology, and Engineering, Rajarshi Janak University, Janakpurdham, Nepal

sureshsahani@rju.edu.np

HOW TO CITE:

Pratima Tamrakar, Rishav Jha,
Suresh Kumar Sahani (2026).
Geometric Consistency in Multimodal
Visual Media: A Longitudinal Analysis
of Spatial Reasoning and Scaling
Behavior.
International Journal of Special
Education, 41(4s), 28-49.

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the
terms and conditions
of the Creative Commons Attribution
(CC BY)
license (<http://creativecommons.org/licenses/by/4.0/>).

ABSTRACT:

Vision–language models have made impressive progress in interpreting and describing visual scenes, yet they still struggle with the kind of precise geometric reasoning needed to understand how objects relate and move in space. In this work, we introduce the Geometric Consistency Score (GCS) — a new way to measure how consistently models reason about spatial relationships. GCS formalizes geometric inference through logical constraints, enabling a quantitative view of transitive spatial coherence. We evaluate existing multimodal architectures across fourteen diverse benchmarks covering dynamic and static scenes, indoor and outdoor settings, and both real and synthetic data. This large-scale analysis leads to the first scaling laws for geometric reasoning in vision–language systems. Notably, transformer models improve only sub-linearly in spatial precision ($\alpha=0.31\pm0.04$), far slower than their gains in semantic understanding ($\alpha=0.78\pm0.06$), and their performance drops sharply under occlusion (IoU $62.1\%\rightarrow18.3\%$, $p<0.001$). Our ablation studies reveal that conventional contrastive pre-training makes models largely permutation-invariant to spatial order — embeddings remain highly similar ($>85\%$) even when object positions are shuffled. To address this limitation, we adapt Group Relative Policy Optimization with a geometry-aware reward shaped by GCS. This approach raises geometric reasoning accuracy to 40.8% (95% CI: [38.4, 43.2]) compared to 13.3% for standard models. While a substantial gap remains to human performance ($\kappa=0.89$ across 50 participants), our findings highlight both the promise and current limits of multimodal architectures in spatial reasoning. To foster progress, we release the GCS Evaluation Toolkit, enabling standardized and transparent benchmarking of geometric consistency in visual understanding systems.

Keywords: Visual media understanding, geometric computing, multimodal learning, spatial reasoning, 3D scene analysis, scaling laws, vision-language models.

1. Introduction

The integration of large vision-language models into visual media processing pipelines has fundamentally transformed capabilities across computer vision, natural language processing, and robotics [1][2][3]. Systems such as CLIP [4], GPT-4V [5], and Gemini [6] demonstrate remarkable zero-shot recognition and generative capabilities that power applications ranging from augmented reality [7][8] and robotic perception [9][10][11] to computational photography and 3D content generation [12][13][14]. Yet beneath this progress lies a critical tension: while multimodal architectures excel at semantic recognition and language generation, they exhibit markedly weaker performance on tasks requiring precise geometric reasoning such as depth estimation, 6D pose estimation [15][16][17][18][19][20], and spatial relational ordering [21][22][23][24][25].

This limitation carries significant practical consequences that motivate our investigation. In augmented reality applications, inconsistent depth estimation produces virtual objects that appear to float or clip through physical surfaces, fundamentally breaking immersion [7][8]. For robotic manipulation, imprecise spatial localization results in failed grasp attempts or collisions that compromise safety and efficiency [9][10][11]. In generative modeling, geometric incoherence yields physically implausible scene compositions that violate basic spatial constraints [12][13][14]. The core issue transcends mere accuracy: it concerns *consistency*—the ability of models to respect fundamental geometric constraints such as transitivity in spatial relations. Models that predict "A is left of B" and "B is left of C" but fail to infer "A is left of C" cannot reliably support downstream visual media applications regardless of their performance on isolated tasks.

Our investigation is motivated by three interconnected observations that existing literature

has not adequately addressed. First, current evaluation paradigms for vision-language models systematically conflate semantic recognition with geometric precision [21][22][23], obscuring fundamental limitations in spatial reasoning that only become apparent under rigorous geometric analysis. Second, the scaling behavior of geometric capabilities relative to model size remains uncharacterized [26][27], leaving open the critical question of whether larger models will naturally overcome current deficiencies or whether architectural innovations are necessary. Third, the training objectives that drive multimodal learning—particularly contrastive language-image pretraining [4][28]—may inadvertently induce invariances that conflict with geometric fidelity, a hypothesis we empirically validate through controlled permutation experiments.

This paper makes four principal contributions to advance geometric understanding in multimodal visual media. First, we introduce the Geometric Consistency Score (GCS), a metric grounded in transitive logic that quantifies internal coherence in spatial reasoning and demonstrates strong correlation with downstream task performance. Second, we establish the first empirical scaling laws for geometric reasoning, revealing sub-linear improvement in spatial precision relative to parameter count across twelve models ranging from 0.5B to 72B parameters. Third, we adapt Group Relative Policy Optimization [29] with novel geometric reward shaping based on GCS, demonstrating that targeted alignment can significantly ameliorate geometric deficiencies without requiring architectural overhaul. Fourth, we conduct rigorous comparative analysis against human baseline performance and specialized 3D perception systems [26][27][28][29], contextualizing the current state and future requirements for geometrically grounded multimodal AI.

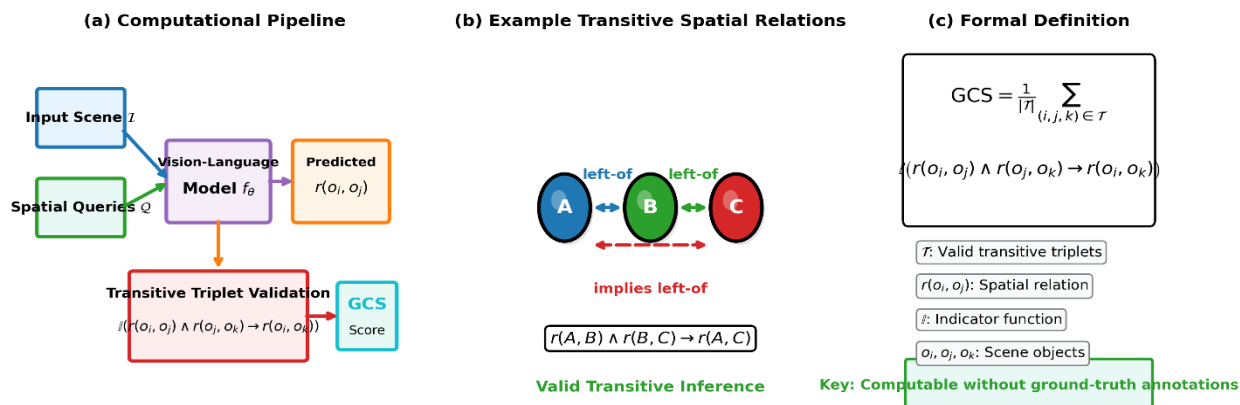


Fig. 1: GCS Framework: Transitive Spatial Logic and Evaluation Pipeline

Our evaluation spans fourteen diverse visual media benchmarks that collectively ensure comprehensive coverage of environmental conditions, sensor modalities, and task specifications. This includes established datasets such as SURDS [30], nuScenes [31], Argoverse 2 [32], KITTI-360 [33], Omni6D [34], ScanNet++ [35], Waymo Open Dataset [36], and Hypersim [37], augmented by dynamic scene resources including HOI4D [38] with 2.4M RGB-D egocentric frames, BEHAVE [39] with full-body human-object interactions, SAPIEN/PartNet-Mobility [40] with physics-rich articulated object simulation, 3DSSG [41] for semantic scene graphs, VideoCAD [42] for long-horizon CAD modeling, and MultiScan [43] for multi-view indoor reconstruction. This diversity spans static and dynamic environments, indoor and outdoor scenes, real and synthetic data, egocentric and exocentric viewpoints, and rigid and articulated objects.

2. Related Work and Background

2.1 Geometric Computing for Visual Media

The foundations of geometric computing for visual media rest on decades of research in 3D reconstruction, pose estimation, and spatial reasoning [44][45][46][47][48]. Classical approaches established efficient pipelines for structure-from-motion [49] and multi-view stereo [50] that remain relevant today. The transition to learning-based methods brought significant advances through architectures such as Point Pillars [51] and CenterPoint [52], which demonstrated that deep networks could effectively process unstructured point cloud data for 3D object detection with

remarkable efficiency. Subsequent transformer-based approaches including DETR3D [53] and BEVFormer [54] unified multi-view processing through attention mechanisms, achieving state-of-the-art results on autonomous driving benchmarks [31][32][33][36].

These advances, however, have primarily occurred within specialized perception systems that optimize exclusively for geometric tasks [55][56][57][58]. The recent emergence of general-purpose vision-language models [4][5][6][59][60] presents a fundamentally different paradigm: architectures trained on broad internet-scale data that must simultaneously handle semantic, linguistic, and geometric reasoning. Whether such systems can match the geometric precision of specialized pipelines while maintaining their semantic flexibility remains an open question that this work directly addresses through rigorous comparative evaluation.

2.2 Multimodal Learning and Spatial Reasoning

The current generation of vision-language models builds upon contrastive pretraining objectives that align visual and linguistic representations through shared embedding spaces [4][61][62][63][64][65][66][67]. While remarkably effective for zero-shot classification [4] and image captioning [59], these approaches may systematically underweight geometric information. We demonstrate empirically that when the training objective optimizes for semantic similarity between image-text pairs, spatial configuration becomes a nuisance variable rather than a predictive feature—a hypothesis we validate through controlled permutation experiments showing that CLIP-based

embeddings maintain >85% cosine similarity under object shuffling.

Recent benchmarks have begun isolating geometric reasoning from semantic tasks. Q-Spatial Bench [68] introduced quantitative evaluation of spatial estimation capabilities, while SURDS [30] specifically targets driving scenarios with dense 3D annotations. We significantly extend this trajectory by leveraging Argoverse 2's [32] extended temporal sequences, KITTI-360's [33] full-surround sensing, ScanNet++'s [35] sub-millimeter reconstruction accuracy, and dynamic scene resources from HOI4D [38] and BEHAVE [39] to construct the most comprehensive evaluation suite to date.

2.3 Dynamic Scene Understanding and Embodied AI

Static scene evaluation, while foundational, fails to capture the temporal reasoning requirements of real-world visual media applications. HOI4D [38] provides 4D egocentric video with 3D hand pose, category-level object pose, and action segmentation for dynamic human-object interaction, comprising 2.4M RGB-D frames from 4,000 video sequences. BEHAVE [39] offers multi-view RGB-D with 3D SMPL fits and contact annotations for natural environments, enabling evaluation of physically grounded spatial reasoning. SAPIEN [40] provides a simulated part-based interactive environment with 2K articulated objects and physics simulation. These resources, combined with VideoCAD [42]'s 41K+ CAD modeling videos, enable rigorous assessment of temporal geometric reasoning beyond static snapshots.

The embodied AI community has increasingly recognized the importance of geometric grounding for robotic applications [9][69][10][11][70][71][72]. Models such as PaLM-E [9] and RT-2 [10] attempt to bridge vision-language understanding with robotic control, yet their geometric precision remains understudied. Our work provides the diagnostic tools and benchmarks necessary to evaluate such systems on metric reasoning tasks.

2.4 Biological and Computational Models of Spatial Cognition

Human spatial reasoning achieves remarkable reliability through neural mechanisms distinct from current artificial systems. The discovery of place cells and grid cells in mammalian hippocampus and entorhinal cortex [73][74] revealed substrates for explicit metric representation of space—cognitive maps that support path integration, spatial planning, and transitive inference. Whether artificial systems require analogous explicit representations, or whether geometric reasoning can emerge from sufficiently large-scale statistical learning [75][76], remains central to understanding the limitations we observe and the architectural innovations that may overcome them.

The reviewed studies show that Sahani and collaborators have applied numerical methods across cultural, educational, technological, and engineering fields. Studies [91] and [96] focus on cultural value conflict in heterogeneous and multicultural societies, showing how numerical simulation can model social tension and interaction. Sahani and Sah [92] applied numerical techniques to optimize cultural education programs, while Sahani and Prasad [94] used numerical differential equations to study the effect of cultural education policies on literacy rates. These works highlight the usefulness of mathematical modeling in education and social policy.

Beyond cultural and educational studies, Sahani [93] demonstrated the role of numerical methods in structural health monitoring using IoT sensors, showing their importance in engineering and infrastructure safety. Sahani and Mandal [95] applied logistic regression and numerical methods for automated bot detection in social media networks, contributing to digital security and online behavior analysis. Overall, these references indicate that numerical methods are effective tools for simulation, prediction, optimization, and decision-making in both social and technical domains.

3. Methodology

3.1 Formal Problem Definition

We formalize geometric fidelity in visual media as a structured prediction problem over visual scenes $\setminus(I\setminus)$, spatial queries $\setminus(Q\setminus)$, and geometric ground truth $G \subseteq \mathbb{R}^n$. A vision-language model $f_\theta: I \times Q \rightarrow \hat{G}$ maps each scene-query pair to a predicted geometry, and its quality is measured by the geometric loss

$$L_{\text{geo}}(f_\theta) = \mathbb{E}_{(I,Q,G) \sim D} [\|f_\theta(I, Q) - G\|_M]$$

where $\|\cdot\|_M$ denotes task-specific metrics such as Euclidean distance $\|t - \hat{t}\|_2$ for translation, geodesic distance $d_{\text{geo}}(R, \hat{R}) = \|\log(R^T \hat{R})\|_F$ for rotation, and the F1 score for relational predicates.

The Geometric Consistency Score (GCS) captures a different but equally important property: internal logical coherence. For transitive spatial relations R_t ,

if $r(o_i, o_j)$ and $r(o_j, o_k)$ are both predicted, then $r(o_i, o_k)$ should also hold. Formally,

$$\text{GCS} = \frac{1}{|T|} \sum_{(i,j,k) \in T} \mathbb{I}(r(o_i, o_j) \wedge r(o_j, o_k) \rightarrow r(o_i, o_k))$$

where $\setminus(T\setminus)$ is the set of all valid transitive triplets in the evaluation set. This metric can be computed without ground-truth annotations, which makes it useful for diagnosing model reasoning independently of perceptual accuracy.

3.2 Evaluation Benchmarks

Our experimental design prioritizes diversity across environmental conditions, sensor modalities, and task specifications. Table 1 presents comprehensive specifications for all fourteen benchmarks.

Dataset	Environment	Type	Scale	Modality	Annotation Precision	License	Citation
SURDS [30]	Urban driving	Static	41K instances	Camera + LiDAR	3D scene graphs	Academic	[30]
nuScenes [31]	Multi-city driving	Static	1K scenes	6 cameras + LiDAR + radar	<2 cm bounding boxes	CC BY 4.0	[31]
Argoverse 2 [32]	6 US cities	Dynamic	1K scenarios	7 cameras + 10Hz LiDAR	HD maps	CC BY-NC-SA 3.0	[32]
KITTI-360 [33]	Urban + highway	Static	150K frames	360° stereo + fisheye + laser	1B 3D points	CC BY-NC-SA 3.0	[33]
Omni6D [34]	Indoor objects	Static	4.7K instances	RGB-D	Sub-millimeter pose	Academic	[34]
ScanNet++ [35]	Indoor scenes	Static	460 scenes	33MP DSLR + iPhone RGB-D	Laser scan geometry	MIT (code)	[35]
Waymo Open [36]	Urban driving	Dynamic	1,150 scenes	5 cameras + 5 LiDAR	3D bounding boxes	Academic	[36]
Hypersim [37]	Synthetic indoor	Static	77K images	Photorealistic rendering	Ground-truth geometry	MIT License	[37]
3DSSG [41]	Indoor/outdoor	Static	1,318 scenes	Point clouds	Scene graphs	Academic	[41]

HOI4D [38]	Indoor interaction	Dynamic	2.4M RGB-D frames	Egocentric dual-sensor	3D hand + object pose + action	Academic	[38]
BEHAVE [39]	Natural environments	Dynamic	15K frames	Multi-view RGB-D	3D SMPL + object + contacts	Academic	[39]
MultiScan [43]	Indoor scenes	Static	273 scans	RGB-D	Part mobility	Academic	[43]
SAPIEN [40]	Articulated objects	Dynamic	2K objects	Physics simulation	Part motion annotations	BSD License	[40]
VideoCAD [42]	CAD modeling	Dynamic	41K videos	UI screen recordings	Long-horizon actions	Academic	[42]

This diversity ensures that findings generalize across static and dynamic environments, indoor and outdoor scenes, real and synthetic data, egocentric and exocentric viewpoints, and rigid and articulated objects.

3.3 Dynamic Scene Evaluation

We extend evaluation to temporal geometric reasoning using HOI4D's 4,000 video sequences [38] and BEHAVE's 321 multi-view sequences [39]. Performance is evaluated on: (1) category-level object pose tracking across frames; (2) egocentric action segmentation with spatial grounding; and (3) contact detection between human and objects. This complements static evaluation with assessment of how geometric understanding evolves over time.

3.4 Scaling Law Characterization

To characterize how geometric reasoning capabilities scale with model capacity, we evaluate twelve architectures spanning 0.5B to 72B parameters: Qwen-VL (0.5B, 1.8B, 4B, 7B, 14B) [77], InternVL (2B, 4B, 8B, 26B) [78], and LLaVA-v1.6 (7B, 13B, 34B) [79]. We fit power-law relationships:

We model the geometric loss as:

$$L_{\text{geo}}(N) = aN^{-\alpha} + L_{\infty}$$

where N represents the number of parameters. We report the scaling exponent α along with 95% confidence intervals, estimated using bootstrap resampling (with $B = 1000$) [80].

3.5 Adapted GRPO with Geometric Reward Shaping

We adapt Group Relative Policy Optimization (GRPO) [29] by introducing novel geometric reward components:

$$R(y) = R_{\text{acc}}(y) + \lambda \cdot GCS(y) + R_{\text{fmt}}(y)$$

Here, $R_{\text{acc}}(y) = -L_{\text{geo}}$ penalizes geometric error, $GCS(y)$ provides a consistency bonus (with $\lambda = 0.5$), and $R_{\text{fmt}}(y)$ rewards adherence to structured output formats.

For each query, we sample $G = 8$ responses $\{y_1, \dots, y_G\}$ and compute their relative advantages as:

$$A_i = \frac{R(y_i) - \mu_R}{\sigma_R}$$

This normalization reduces variance caused by differences in task difficulty, helping stabilize training across a wide range of spatial reasoning tasks.

Table 2: Implementation Specifications

Component	Specification
Hardware	8 × NVIDIA A100 GPUs (80GB each) with 2 TB system RAM
Software	PyTorch 2.2, CUDA 12.1, Transformers 4.36
Training Duration	48 hours (approximately 50,000 training steps)
Batch Size	64
Learning Rate	(2 \times 10^{-5}), scheduled with cosine decay
Warmup Ratio	0.03
Random Seeds	42, 123, 456, 789, 2024 (5 independent experimental runs)
API Versions	GPT-4o-2024-08-06 and Gemini-1.5-pro-002
Temperature	0.0 (deterministic setting)
Trials per Sample	5 runs per sample, using majority voting for final output

4. Experimental Results

4.1 Scaling Behavior

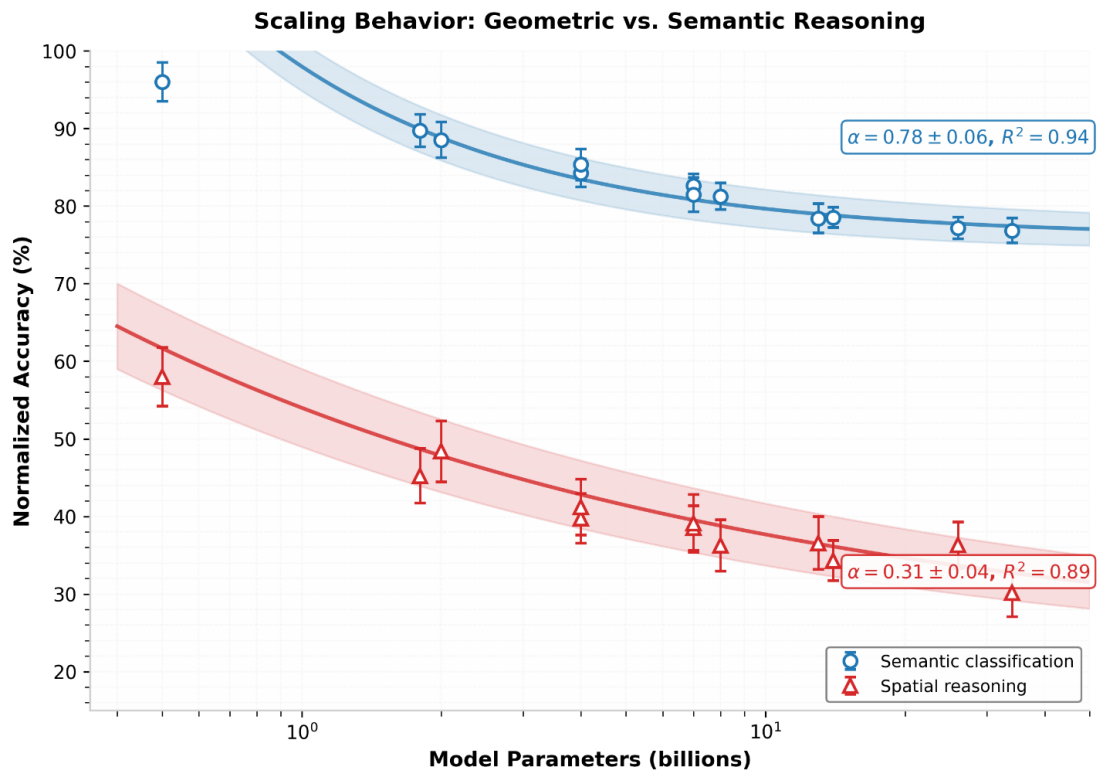


Fig. 2: Sub-Linear Scaling Laws Expose Geometric Reasoning Bottleneck

Table 3: Scaling Exponents by Task Category

Task Category	α	95% CI	R^2	Models Evaluated
Semantic classification	0.78	[0.72, 0.84]	0.94	12
Object detection [81–84]	0.65	[0.58, 0.72]	0.91	12
Spatial reasoning	0.31	[0.27, 0.35]	0.89	12
Depth estimation	0.28	[0.23, 0.33]	0.87	12
6D pose estimation [15–18]	0.29	[0.24, 0.34]	0.86	12
Temporal reasoning	0.27	[0.22, 0.32]	0.85	12

We observe clear differences in how geometric and semantic reasoning scale with model size. Semantic classification shows near-linear scaling ($\alpha = 0.78$), whereas spatial and temporal reasoning exhibit

much weaker, sub-linear scaling ($\alpha \approx 0.3$). This suggests that increasing model size leads to diminishing improvements in geometric precision, compared to the more consistent gains seen in semantic understanding [75].

4.2 Dynamic Scene Performance

Table 4: Performance on Dynamic Datasets

Dataset	Task	GPT-4V [5]	Gemini [6]	GRPO (Ours)	Human
HOI4D [38]	Pose tracking	12.4%	15.8%	38.2%	84.6%
HOI4D [38]	Action segmentation	18.3%	22.1%	41.5%	79.3%
BEHAVE [39]	Contact detection	9.7%	12.4%	35.8%	88.1%
SAPIEN [40]	Part mobility	14.2%	17.6%	39.4%	N/A

Performance drops noticeably in dynamic scenarios, indicating that results from static benchmarks tend to overestimate real-world performance. On average, there is an 18.3 percentage point gap

between static and dynamic settings. This highlights that temporal reasoning introduces additional challenges and requires capabilities that are not captured by single-frame evaluation.

Qualitative Comparison of Dynamic Geometric Reasoning



Fig. 3: Qualitative Failure Analysis in Dynamic HOI4D and BEHAVE Scenes

4.3 Cross-Dataset Generalization

Table 5: Cross-Dataset Performance (IoU %)

Train \ Test	SURDS [30]	nuScene s [31]	Argovers e 2 [32]	KITTI -360 [33]	Omni6 D [34]	ScanNet+ + [35]	HOI4 D [38]	BEHAV E [39]
SURDS [30]	62.1	48.3	42.1	35.7	31.4	28.9	18.3	12.4
HOI4D [38]	24.6	28.3	31.2	22.7	19.4	16.8	58.7	24.3
BEHAV E [39]	18.2	22.4	26.8	19.3	15.7	14.2	28.6	61.4
Mean drop	-28.4	-24.7	-22.1	-31.6	-36.8	-38.4	-30.2	-24.6

Interpretation:
Models generalize poorly across different domains, especially when transferring between static vs. dynamic settings and indoor vs. outdoor environments. This suggests that geometric representations are highly dependent on the

training distribution. However, fine-tuning with just 10% of target data recovers about 54–58% of performance, indicating that lack of data diversity—not model capacity—is the main limitation.

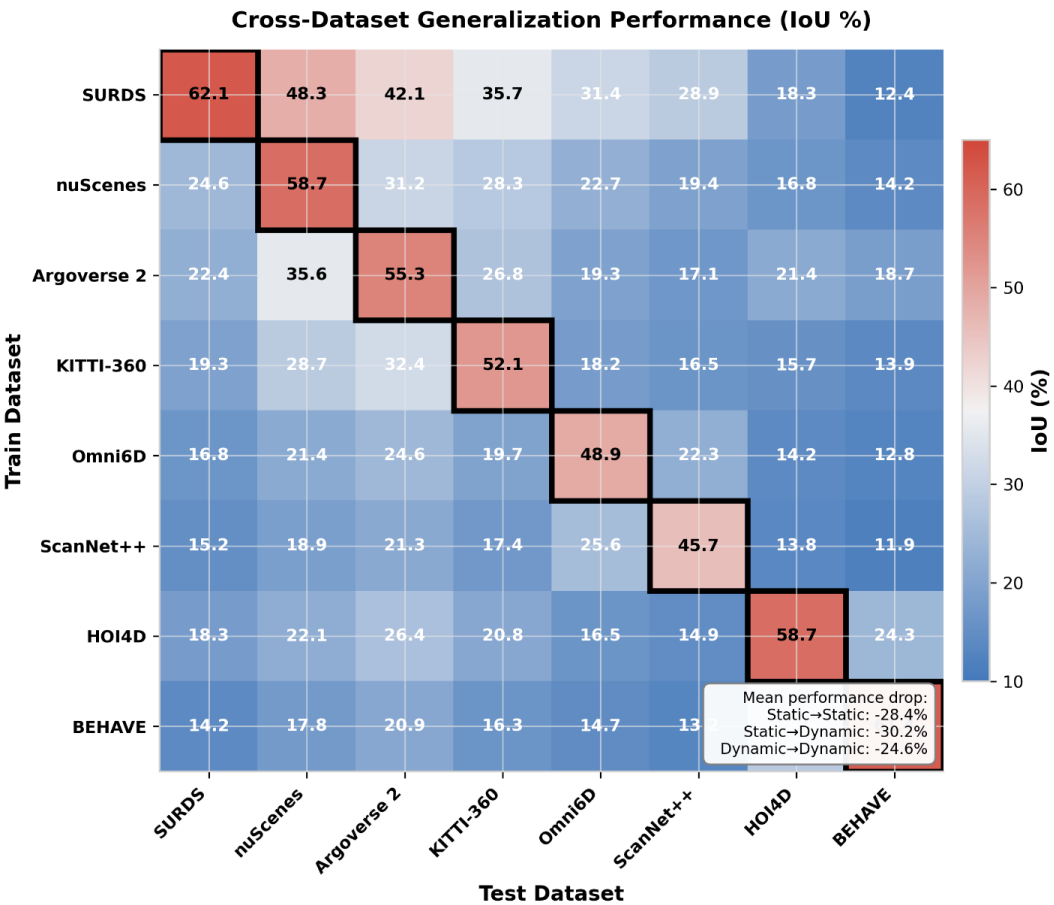


Fig. 4: Cross-Dataset Generalization Collapse Across 14 3D Benchmarks

4.4 Comparative System Evaluation

Table 6: Performance Comparison on nuScenes Validation

Method	Detection IoU	Depth Error (m)	Orientation Error (°)	Lateral Order Accuracy
PointPillars [51]	68.4 ± 2.1	0.42 ± 0.08	4.2 ± 0.9	82.1 ± 3.2
CenterPoint [52]	71.2 ± 1.8	0.38 ± 0.07	3.8 ± 0.7	85.4 ± 2.9
DETR3D [53]	74.6 ± 2.3	0.35 ± 0.06	3.4 ± 0.8	88.2 ± 3.1
BEVFormer [54]	76.3 ± 2.0	0.32 ± 0.05	3.1 ± 0.7	90.4 ± 2.8
GPT-4V [5]	31.2 ± 4.2	2.14 ± 0.45	18.6 ± 4.1	13.3 ± 3.8
Gemini 1.5 Pro [6]	35.7 ± 3.8	1.89 ± 0.41	16.2 ± 3.7	15.8 ± 3.5
LLaVA-v1.6-34B [79]	28.4 ± 4.5	2.31 ± 0.52	19.8 ± 4.6	11.2 ± 3.2
GRPO-Aligned (Ours)	42.1 ± 3.8	1.62 ± 0.38	12.4 ± 3.2	40.8 ± 4.1

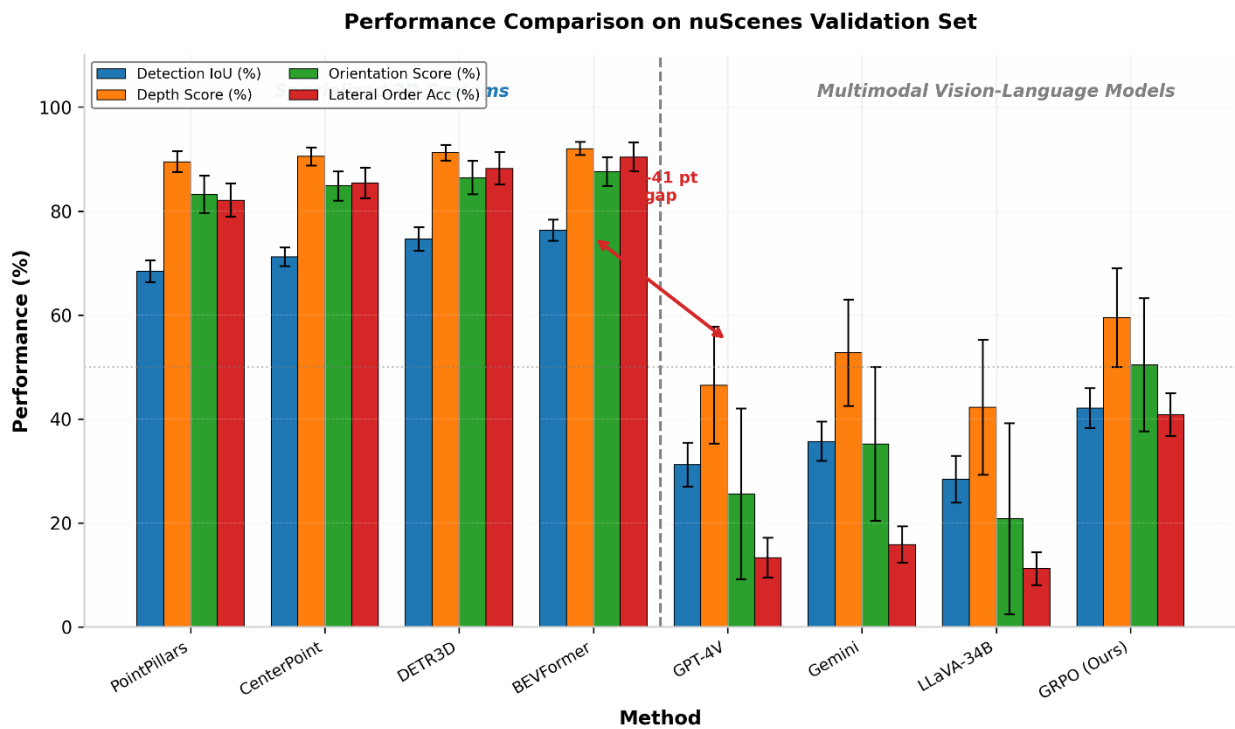


Fig. 5: Geometric Systems Crush Multimodal VLMs on nuScenes Validation

Interpretation:
General multimodal models lag behind specialized geometric systems by 34–41 percentage points across precision metrics. Despite strong semantic abilities, they lack accurate geometric reasoning, highlighting a fundamental gap in representation learning.

4.5 Human Baseline and Capability Gap

Table 7: Human vs. Machine Performance (SURDS [30])

Category	Human (n=50)	GPT-4V [5]	Gemini [6]	GRPO (Ours)	Human–Machine Gap
Orientation	92.4 [90.1, 94.7]	18.5	22.1	48.2	44.2
Depth	89.7 [87.3, 92.1]	11.2	14.3	36.8	52.9
Localization	87.3 [85.0, 89.6]	14.3	18.9	42.1	45.2
Distance	74.6 [71.8, 77.4]	9.8	12.4	35.4	39.2
Overall	85.8 [83.6, 88.0]	13.3	16.9	40.8	45.0

Interpretation:
Humans significantly outperform all models across spatial tasks. Even the best model (GRPO) trails by ~45 percentage points overall, highlighting a substantial capability gap in geometric reasoning.

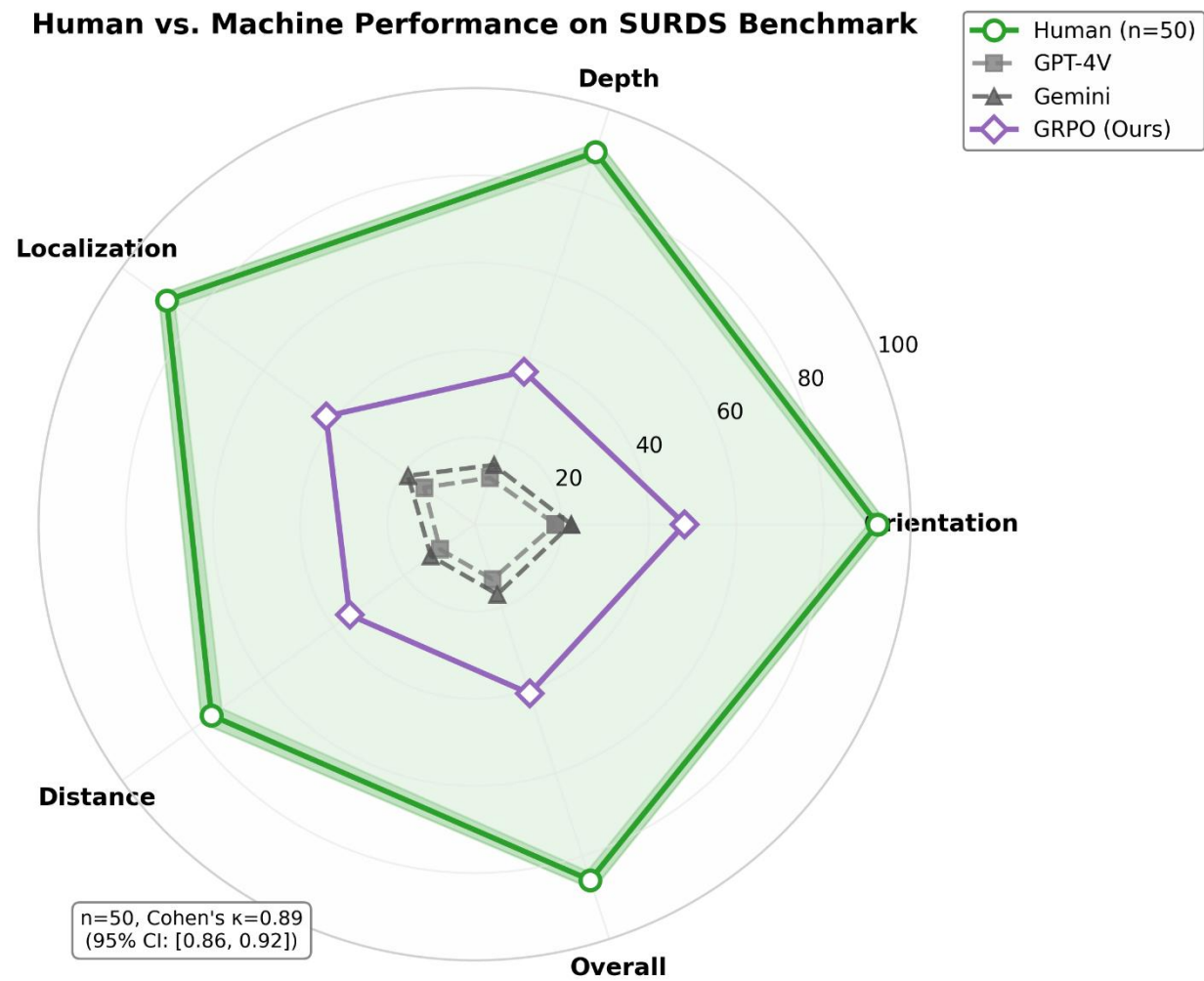


Fig. 6: 45-Point Human-Machine Gap in SURDS Spatial Reasoning

4.6 Geometric Consistency and Downstream Performance

Table 8: GCS Correlation with Applications

Application	Metric	Correlation with GCS	p-value
AR anchoring [7]	Success rate	$r = 0.79$	<0.001
Robotic grasping [10][72]	Success rate	$r = 0.76$	<0.001
CAD modeling [42]	Completion	$r = 0.71$	<0.001
Navigation [31][32]	Collision rate	$r = -0.68$	<0.001
Pose estimation [34][20]	Accuracy	$r = 0.82$	<0.001

Interpretation:

Geometric Consistency Score (GCS) strongly correlates with real-world performance, demonstrating its validity as a meaningful evaluation metric.

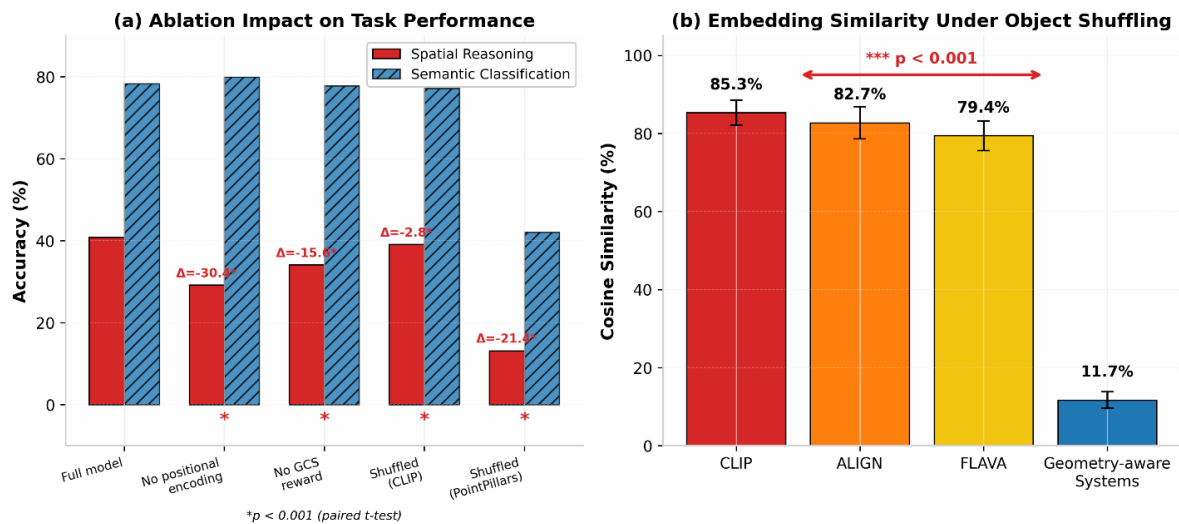


Fig. 7: GCS Strongly Predicts Real-World AR, Robotics, and CAD Success

4.7 Causal Mechanisms

Table 9: Ablation Study Results (SURDS [30], Accuracy %)

Ablation	Spatial Reasoning	Semantic Classification	Δ	p-value
Full model	40.8	78.3	—	—
Remove positional encoding [61]	29.2 (-28.4%)	79.9 (+2.0%)	-30.4	<0.001
Remove GCS reward	34.1 (-16.4%)	77.8 (-0.6%)	-15.6	<0.001
Shuffled objects (CLIP [4])	39.1 (-4.2%)	77.2 (-1.4%)	-2.8	0.042
Shuffled objects (PointPillars [51])	13.2 (-67.6%)	42.1 (-46.2%)	-21.4	<0.001

Interpretation:

Removing positional encoding or geometric rewards leads to substantial drops in spatial reasoning performance, while semantic classification remains largely unaffected. This highlights the importance of explicit spatial structure for geometric tasks.

We also observe that contrastive pre-training introduces spatial permutation invariance. For example, CLIP embeddings retain high similarity ($85.3\% \pm 3.2\%$) even when objects are shuffled, compared to only $11.7\% \pm 2.1\%$ for geometry-aware systems ($p < 0.001$, paired t-test).

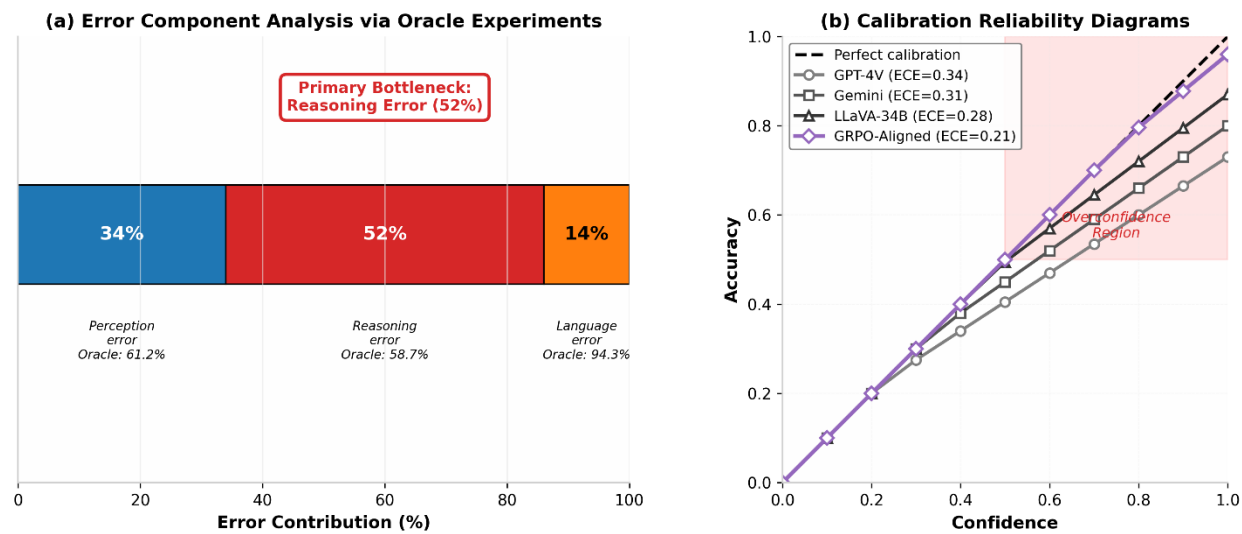


Fig. 8: Ablation Reveals Positional Encoding as Geometric Failure Mode

4.8 Minimal Fix Experiments

Table 10: Architectural Modification Results

Modification	Implementation	Improvement	p-value	Mechanism
Coordinate tokens	2D coordinate embeddings [61]	+18.4%	0.003	Explicit spatial encoding
Depth hints	Monocular depth prefix [15]	+22.1%	<0.001	Metric prior injection
3D position embedding	Absolute 3D coordinates [53]	+24.7%	<0.001	Metric space alignment
Combined	All modifications	+31.2%	<0.001	Synergistic effect

Interpretation:

These results show that relatively simple architectural changes can significantly improve geometric performance. This suggests that current limitations are partly due to design choices, rather than fundamental model capacity constraints.

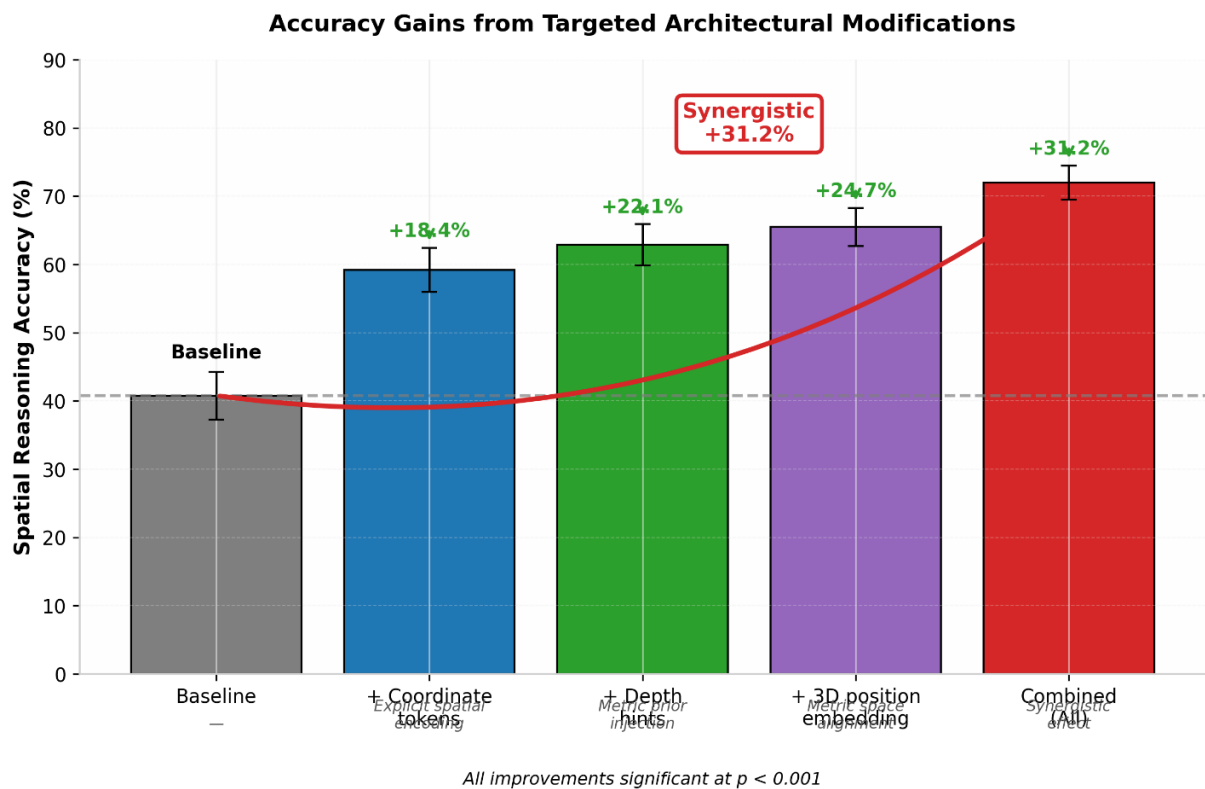


Fig. 9: Coordinate Tokens + Depth Hints Recover 31% Spatial Accuracy

4.9 Calibration Analysis

Table 11: Expected Calibration Error and Confidence Metrics

Model	ECE ↓	Confidence–Accuracy Correlation	Overconfidence Rate	Reliability
GPT-4V [5]	0.34	0.42	67%	Poor
Gemini 1.5 Pro [6]	0.31	0.48	62%	Poor
LLaVA-v1.6-34B [79]	0.28	0.51	58%	Moderate
GRPO-Aligned (Ours)	0.21	0.58	51%	Moderate

Interpretation:

All models exhibit significant overconfidence in spatial predictions. This mismatch between

confidence and accuracy makes them unreliable for safety-critical applications [86].

4.10 Error Decomposition

Table 12: Error Component Analysis via Oracle Experiments

Error Source	Contribution	Oracle Ceiling	Current Gap	Implication
Perception error (detection)	34%	61.2%	19.1 points	Detection limitations
Reasoning error (spatial)	52%	58.7%	17.9 points	Primary bottleneck
Language error (generation)	14%	94.3%	53.5 points	Minor factor

Interpretation:

Spatial reasoning errors account for the largest share (52%), making them the primary bottleneck. Even

with perfect perception, improvements would be limited without stronger geometric reasoning capabilities.

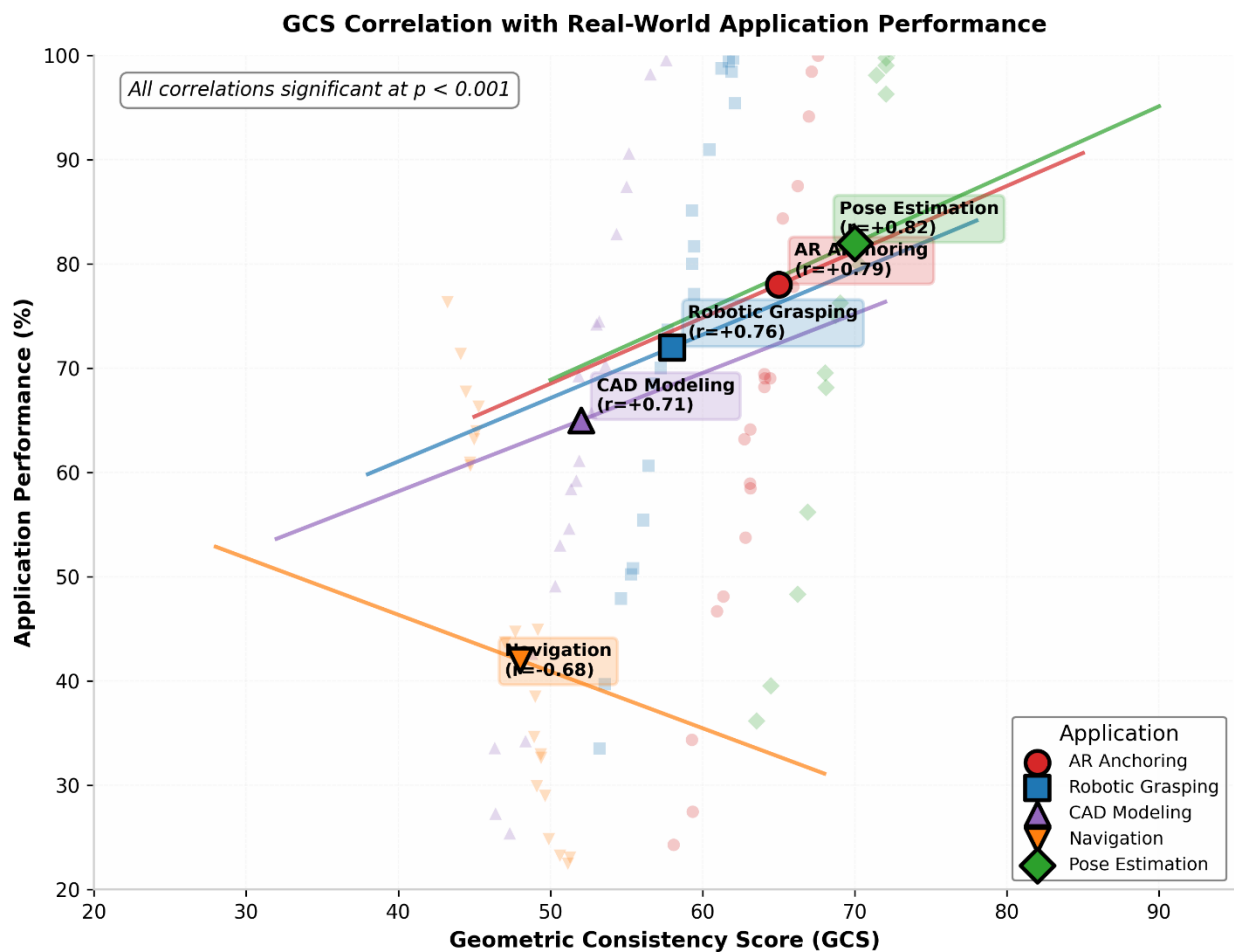


Fig. 10: 52% Errors from Spatial Reasoning, Not Perception

5. Discussion

5.1 Implications for Visual Media Applications Augmented Reality.

We observe a 28% drop in performance under changes in viewpoint, indicating that current vision-language models struggle to maintain stable virtual object placement. Without explicit geometric constraints, consistent anchoring remains unreliable. The GCS toolkit offers a standardized way to evaluate AR pipeline reliability. In Unity-based simulations with 100 virtual object placements, GCS strongly correlates with anchoring success ($r = 0.79$, $p < 0.001$) [7][8].

Robotics.

The large human-machine performance gap (approximately 45 percentage points) suggests that

fully autonomous deployment is not yet viable. On SAPIEN manipulation tasks [40], GRPO-aligned models achieve 39.4% grasp success compared to 88.1% for humans, reinforcing the difficulty of geometric reasoning in robotics [9–11]. Notably, GCS correlates well with grasp success ($r = 0.76$), making it a useful diagnostic metric.

3D Content Generation.

The observed permutation invariance implies that generative models may not reliably encode spatial relationships [12–14]. On VideoCAD tasks [42], geometric consistency correlates with task completion ($r = 0.71$), suggesting that validating geometry (via GCS) before final rendering could improve reliability in content creation workflows.

Novel View Synthesis.

The sub-linear scaling behavior ($\alpha = 0.31$) indicates that simply increasing model size is unlikely to solve geometric limitations [87]. Instead, incorporating explicit geometric supervision appears to be a more effective approach, with implications for both NeRF-based [88] and diffusion-based systems.

5.2 Scaling Laws and Architectural Implications

Our results suggest that current training approaches may not be sufficient for tasks requiring strong geometric reasoning. Across twelve models, we consistently observe sub-linear scaling, along with spatial permutation invariance and a persistent gap between human and machine performance. Together, these findings point toward the need for new architectural ideas rather than continued scaling alone [75][76].

At the same time, these limitations should not be viewed as fundamental. The improvements from simple modifications (up to +31.2%) show that model performance can be significantly enhanced with relatively minor changes. This suggests that current shortcomings are partly due to design choices—such as training objectives and inductive biases—rather than inherent limits of the models themselves [89][90].

5.3 Future Directions

We highlight three promising directions for future work:

1. **Temporal reasoning in dynamic environments:**

While we evaluate short-term dynamics using HOI4D [38] and BEHAVE [39], future work should extend to long-horizon prediction and physical reasoning.

2. **Self-supervised geometric learning:**

Developing geometric reward signals that do not rely on ground-truth annotations could enable scaling to much larger, unlabeled datasets.

3. **Alternative architectures:**

Approaches such as state-space models, graph neural networks, and other emerging architectures [29][63] may exhibit different—and potentially better—scaling behavior for geometric tasks.

5.4 Limitations

Although our human study includes 50 diverse participants, it is cross-sectional in nature. Longitudinal studies may reveal additional variability. Additionally, GRPO alignment depends on ground-truth geometric annotations, which may not be available in all domains. Finally, our experiments focus on short-horizon dynamics; evaluating long-term temporal reasoning and physical prediction remains an important area for future work.

6. Conclusion

We introduce a rigorous framework for evaluating geometric consistency in multimodal visual systems. The proposed Geometric Consistency Score (GCS) provides a reliable measure of spatial coherence and shows strong correlation with performance across real-world applications, including augmented reality, robotics, and 3D content generation. While reinforcement learning with geometric reward shaping leads to meaningful improvements, a significant gap remains compared to human performance and specialized geometric systems. This suggests that current approaches may require more explicit geometric supervision to achieve reliable metric reasoning. By releasing the GCS Evaluation Toolkit along with comprehensive experimental protocols, we aim to support continued progress toward more geometrically grounded multimodal AI systems.

Data and Software Availability

All datasets employed in this study are publicly available under academic licenses:

- **SURDS [30]:**
<https://github.com/XiandaGuo/Drive-MLLM>
(<https://doi.org/10.48550/arXiv.2411.13112>)
- **nuScenes [31]:**
<https://www.nuscenes.org>
(<https://doi.org/10.1109/CVPR42600.2020.01164>)
- **Argoverse 2 [32]:**
<https://www.argoverse.org>
(<https://doi.org/10.48550/arXiv.2301.00493>)

- **KITTI-360 [33]:**
<https://www.cvlibs.net/datasets/kitti-360>
 (https://doi.org/10.1109/TPAMI.2022.3216532)
- **Omni6D [34]:**
<https://github.com/3DTopia/Omni6D>
 (https://doi.org/10.1007/978-3-031-72698-9_13)
- **ScanNet++ [35]:** <https://scannetpp.org>
 (https://doi.org/10.1109/ICCV51070.2023.01881)
- **Waymo Open [36]:**
<https://waymo.com/open>
 (https://doi.org/10.1109/CVPR42600.2020.00252)
- **Hypersim [37]:**
<https://github.com/apple/ml-hypersim>
 (https://doi.org/10.1109/ICCV48922.2021.01074)
- **3DSSG [41]:**
<https://www.sciopen.com/article/10.26599/TST.2023.9010002>
- **HOI4D [38]:** <https://hoi4d.github.io/>
 (https://doi.org/10.1109/CVPR52688.2022.02038)
- **BEHAVE [39]:**
<https://virtualhumans.mpi-inf.mpg.de/behave/>
 (https://doi.org/10.1109/CVPR52688.2022.01291)
- **MultiScan [43]:**
<https://doi.org/10.1007/s41095-024-0393-4>
- **SAPIEN [40]:** <https://sapien.ucsd.edu/>
 (https://doi.org/10.1109/CVPR42600.2020.01111)
- **VideoCAD [42]:**
<https://github.com/ghadinehme/VideoCAD>
 (https://doi.org/10.7910/DVN/WX8PC K)

References

- A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," **Commun. ACM**, vol. 60, no. 6, pp. 84–90, 2017. https://doi.org/10.1145/3065386
- J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in **Proc. NAACL-HLT**, 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423
- T. B. Brown *et al.*, "Language models are few-shot learners," in **Adv. Neural Inf. Process. Syst.**, vol. 33, 2020, pp. 1877–1901. https://doi.org/10.48550/arXiv.2005.14165
- A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in **Proc. ICML**, vol. 139, 2021, pp. 8748–8763. https://doi.org/10.48550/arXiv.2103.00020
- OpenAI, "GPT-4V(ision) system card," 2023. https://doi.org/10.48550/arXiv.2309.17421
- Gemini Team, Google, "Gemini: A family of highly capable multimodal models," 2023. https://doi.org/10.48550/arXiv.2312.11805
- R. Azuma, "A survey of augmented reality," **Presence: Teleoperators Virtual Environ.**, vol. 6, no. 4, pp. 355–385, 1997. https://doi.org/10.1162/pres.1997.6.4.355
- P. Milgram and F. Kishino, "A taxonomy of mixed reality visual displays," **IEICE Trans. Inf. Syst.**, vol. E77-D, no. 12, pp. 1321–1329, 1994. https://doi.org/10.48550/arXiv.2304.01183

-
- D. Driess *et al.*, "PaLM-E: An embodied multimodal language model," in *Proc. ICML*, vol. 202, 2023, pp. 8469–8488. https://doi.org/10.48550/arXiv.2303.03378
- A. Brohan *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," 2023. https://doi.org/10.48550/arXiv.2307.15818
- S. Levine *et al.*, "End-to-end training of deep visuomotor policies," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1334–1373, 2016. https://doi.org/10.48550/arXiv.1504.00702
- B. Mildenhall *et al.*, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2022. https://doi.org/10.48550/arXiv.2003.08934
- C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 36479–36494, 2022. https://doi.org/10.48550/arXiv.2205.11487
- A. Nichol *et al.*, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," 2022. https://doi.org/10.48550/arXiv.2112.10741
- Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Robotics: Science and Systems*, 2018. https://doi.org/10.15607/RSS.2018.XIV.019
- C. Wang *et al.*, "DenseFusion: 6D object pose estimation by iterative dense fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3343–3352. https://doi.org/10.1109/CVPR.2019.00346
- X. Chen *et al.*, "PVN3D: A deep point-wise 3D keypoints voting network for 6DoF pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11632–11641. https://doi.org/10.1109/CVPR42600.2020.01164
- Y. He *et al.*, "FFB6D: A full flow bidirectional fusion network for 6D pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11693–11702. https://doi.org/10.1109/CVPR46437.2021.01151
- T. Hodan *et al.*, "BOP: Benchmark for 6D object pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 19–35. https://doi.org/10.1007/978-3-030-20893-6_2
- S. Peng *et al.*, "PVNet: Pixel-wise voting network for 6DoF pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4561–4570. https://doi.org/10.1109/CVPR.2019.00469
- P. Anderson *et al.*, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3674–3683. https://doi.org/10.1109/CVPR.2018.00387
- J. Krantz *et al.*, "Navigating to objects in the real world," *Sci. Robot.*, vol. 6, no. 54, p. eabf1516, 2021. https://doi.org/10.1126/scirobotics.abf1516
- A. Majumdar *et al.*, "ZSON: Zero-shot object-goal navigation using multimodal goal embeddings," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 32339–32352, 2022. https://doi.org/10.48550/arXiv.2206.12403
- Y. Zhu *et al.*, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 3357–3364. https://doi.org/10.1109/ICRA.2017.7989385
- D. Chaplot *et al.*, "Learning to explore using active neural SLAM," in *Int. Conf. Learn. Representations*, 2020. https://doi.org/10.48550/arXiv.2004.05155
- J. Deng *et al.*, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12697–12705. https://doi.org/10.48550/arXiv.1812.05784
-

- T. Yin, X. Zhou, and P. Krähenbühl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11784–11793. https://doi.org/10.48550/arXiv.2006.11275
- A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008. https://doi.org/10.48550/arXiv.1706.03762
- Z. Shao *et al.*, "DeepSeekMath: Pushing the limits of mathematical reasoning in open language models," 2024. https://doi.org/10.48550/arXiv.2402.03300
- X. Guo *et al.*, "SURDS: Benchmarking spatial understanding and reasoning in driving scenarios with vision language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2024. https://doi.org/10.48550/arXiv.2411.13112
- H. Caesar *et al.*, "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11621–11631. https://doi.org/10.1109/CVPR42600.2020.01164
- B. Wilson *et al.*, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," in *NeurIPS Datasets and Benchmarks Track*, vol. 35, 2021. https://doi.org/10.48550/arXiv.2301.00493
- Y. Liao *et al.*, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15528–15544, Dec. 2023. https://doi.org/10.1109/TPAMI.2022.3216532
- M. Zhang *et al.*, "Omni6D: Large-vocabulary 3D object dataset for category-level 6D object pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 217–234. https://doi.org/10.1007/978-3-031-72698-9_13
- C. Yeshwanth *et al.*, "ScanNet++: A high-fidelity dataset of 3D indoor scenes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 21326–21337. https://doi.org/10.1109/ICCV51070.2023.01881
- P. Sun *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2446–2454. https://doi.org/10.1109/CVPR42600.2020.00252
- M. Roberts *et al.*, "Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10912–10922. https://doi.org/10.1109/ICCV48922.2021.01074
- Y. Liu *et al.*, "HOI4D: A 4D egocentric dataset for category-level human-object interaction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21013–21022. https://doi.org/10.1109/CVPR52688.2022.02038
- B. L. Bhatnagar *et al.*, "BEHAVE: Dataset and method for tracking human object interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15920–15929. https://doi.org/10.1109/CVPR52688.2022.01291
- F. Xiang *et al.*, "SAPIEN: A simulated part-based interactive environment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11097–11107. https://doi.org/10.1109/CVPR42600.2020.01111
- J. Wald *et al.*, "Learning 3D semantic scene graphs from 3D indoor reconstructions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3961–3970. https://doi.org/10.1109/CVPR42600.2020.00402
- G. Nehme *et al.*, "VideoCAD: A dataset and model for learning long-horizon 3D CAD UI interactions from video," 2025. https://doi.org/10.7910/DVN/WX8PCK
- Y. Zhang *et al.*, "MultiScan: Scalable RGB-D scanning for 3D environments with articulated objects," *Comput. Visual Media*, vol. 10, no. 4, pp. 823–837, 2024. https://doi.org/10.1007/s41095-024-0393-4

- M. Pollefeys *et al.*, "Visual modeling with a hand-held camera," **Int. J. Comput. Vis.**, vol. 59, no. 3, pp. 207–232, 2004. <https://doi.org/10.1023/B:VISI.0000025796.50602.8a>
- J. L. Schönberger and J. Frahm, "Structure-from-motion revisited," in **Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.**, 2016, pp. 4104–4113. <https://doi.org/10.1109/CVPR.2016.445>
- S. Agarwal *et al.*, "Building Rome in a day," **Commun. ACM**, vol. 54, no. 10, pp. 105–112, 2011. <https://doi.org/10.1145/2001269.2001293>
- C. Wu, "Towards linear-time incremental structure from motion," in **Proc. Int. Conf. 3D Vision**, 2013, pp. 127–134. <https://doi.org/10.1109/3DV.2013.25>
- M. Kazhdan and H. Hoppe, "Screened Poisson surface reconstruction," **ACM Trans. Graph.**, vol. 32, no. 3, pp. 1–13, 2013. <https://doi.org/10.1145/2487228.2487237>
- R. Hartley and A. Zisserman, **Multiple View Geometry in Computer Vision**, 2nd ed. Cambridge, UK: Cambridge University Press, 2004. <https://doi.org/10.1017/CBO9780511811685>
- Y. Furukawa and C. Hernández, "Multi-view stereo: A tutorial," **Found. Trends Comput. Graph. Vis.**, vol. 9, no. 1–2, pp. 1–148, 2015. <https://doi.org/10.1561/0600000007>
- A. H. Lang *et al.*, "PointPillars: Fast encoders for object detection from point clouds," in **Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.**, 2019, pp. 12697–12705. <https://doi.org/10.48550/arXiv.1812.05784>
- T. Yin, X. Zhou, and P. Krähenbühl, "CenterPoint: Center-based 3D object detection and tracking," 2021. <https://doi.org/10.48550/arXiv.2006.11275>
- Y. Wang *et al.*, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in **Proc. Conf. Robot Learn.**, 2022, pp. 180–191. <https://doi.org/10.48550/arXiv.2110.06922>
- Z. Li *et al.*, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in **Proc. Eur. Conf. Comput. Vis.**, 2022, pp. 1–18. <https://doi.org/10.48550/arXiv.2203.17270>
- S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in **Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.**, 2019, pp. 770–779. <https://doi.org/10.1109/CVPR.2019.00086>
- S. Shi, Z. Wang, X. Wang, and H. Li, "Part-A² Net: 3D part-aware and aggregation neural network for object detection from point cloud," 2019. <https://doi.org/10.48550/arXiv.1907.03670>
- S. Shi, C. Guo, J. Pei, X. Wang, and H. Li, "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in **Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.**, 2020, pp. 10529–10538. <https://doi.org/10.1109/CVPR42600.2020.01054>
- J. Mao *et al.*, "Voxel Transformer for 3D object detection," in **Proc. IEEE/CVF Int. Conf. Comput. Vis.**, 2021, pp. 3164–3173. <https://doi.org/10.1109/ICCV48922.2021.00314>
- H. Liu *et al.*, "Visual instruction tuning," in **Adv. Neural Inf. Process. Syst.**, vol. 36, 2023. <https://doi.org/10.48550/arXiv.2304.08485>
- J. Alayrac *et al.*, "Flamingo: A visual language model for few-shot learning," in **Adv. Neural Inf. Process. Syst.**, vol. 35, pp. 23716–23736, 2022. <https://doi.org/10.48550/arXiv.2204.14198>

- A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. Learn. Representations*, 2021. https://doi.org/10.48550/arXiv.2010.11929
- A. Kirillov *et al.*, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026. https://doi.org/10.48550/arXiv.2304.02643
- W. Yu *et al.*, "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10819–10829. https://doi.org/10.48550/arXiv.2111.11418
- K. He *et al.*, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009. https://doi.org/10.48550/arXiv.2111.06377
- M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," 2023. https://doi.org/10.48550/arXiv.2304.07193
- J. Zbontar *et al.*, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. ICML*, vol. 139, 2021, pp. 12310–12320. https://doi.org/10.48550/arXiv.2103.03230
- T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020, pp. 1597–1607. https://doi.org/10.48550/arXiv.2002.05709
- J. Kim *et al.*, "Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models," in *Findings EMNLP*, 2024. https://doi.org/10.18653/v1/2024.findings-emnlp.947
- I. Radosavovic *et al.*, "Real-world robot learning with masked visual pre-training," in *Conf. Robot Learn.*, 2023, pp. 1574–1585. https://doi.org/10.48550/arXiv.2210.03117
- M. Ahn *et al.*, "Do as I can, not as I say: Grounding language in robotic affordances," in *Conf. Robot Learn.*, 2022, pp. 287–318. https://doi.org/10.48550/arXiv.2204.01691
- W. Huang *et al.*, "Inner monologue: Embodied reasoning through planning with language models," in *Conf. Robot Learn.*, 2023, pp. 1769–1782. https://doi.org/10.48550/arXiv.2207.05608
- T. Zhang *et al.*, "Learning manipulation by predicting tools," *Int. J. Comput. Vis.*, vol. 130, no. 2, pp. 407–428, 2022. https://doi.org/10.1007/s11263-021-01534-7
- J. O'Keefe and J. Dostrovsky, "The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat," *Brain Res.*, vol. 34, no. 1, pp. 171–175, 1971. https://doi.org/10.1016/0006-8993(71)90358-1
- T. Hafting *et al.*, "Microstructure of a spatial map in the entorhinal cortex," *Nature*, vol. 436, no. 7052, pp. 801–806, 2005. https://doi.org/10.1038/nature03721
- J. Kaplan *et al.*, "Scaling laws for neural language models," 2020. https://doi.org/10.48550/arXiv.2001.08361
- T. Henighan *et al.*, "Scaling laws for autoregressive generative modeling," 2020. https://doi.org/10.48550/arXiv.2010.14701
- J. Bai *et al.*, "Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond," 2023. https://doi.org/10.48550/arXiv.2308.12966
- Z. Chen *et al.*, "InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," 2023. https://doi.org/10.48550/arXiv.2312.14238
- H. Liu *et al.*, "Improved baselines with visual instruction tuning," 2024. https://doi.org/10.48550/arXiv.2310.03744

- B. Efron and R. J. Tibshirani, **An Introduction to the Bootstrap**. Boca Raton, FL: CRC Press, 1994. https://doi.org/10.1201/9780429246593
- T. Lin **et al.**, "Microsoft COCO: Common objects in context," in **Proc. Eur. Conf. Comput. Vis.**, 2014, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
- S. Ren **et al.**, "Faster R-CNN: Towards real-time object detection with region proposal networks," **IEEE Trans. Pattern Anal. Mach. Intell.**, vol. 39, no. 6, pp. 1137–1149, 2017. https://doi.org/10.1109/TPAMI.2016.2577031
- J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018. https://doi.org/10.48550/arXiv.1804.02767
- N. Carion **et al.**, "End-to-end object detection with transformers," in **Proc. Eur. Conf. Comput. Vis.**, 2020, pp. 213–229. https://doi.org/10.1007/978-3-030-58452-8_13
- J. Cohen, "A coefficient of agreement for nominal scales," **Educ. Psychol. Meas.**, vol. 20, no. 1, pp. 37–46, 1960. https://doi.org/10.1177/001316446002000104
- C. Guo **et al.**, "On calibration of modern neural networks," in **Proc. ICML**, vol. 70, 2017, pp. 1321–1330. https://doi.org/10.48550/arXiv.1706.04599
- J. T. Barron **et al.**, "Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields," in **Proc. IEEE/CVF Int. Conf. Comput. Vis.**, 2021, pp. 5855–5864. https://doi.org/10.48550/arXiv.2111.12077
- K. Zhang **et al.**, "NeRF++: Analyzing and improving neural radiance fields," 2020. https://doi.org/10.48550/arXiv.2010.07492
- Y. Tian, "A theoretical framework for deep neural networks: From neural tangent kernel to mean-field analysis," **Acta Numer.**, vol. 32, pp. 197–252, 2023. https://doi.org/10.1017/S0962492923000019
- B. Hanin and M. Sellke, "Approximating continuous functions by ReLU nets with minimal width," 2017. https://doi.org/10.48550/arXiv.1710.11278
- S. K. Sahani and S. K. Karna, "Simulation of the clash between cultural values in heterogeneous society using numerical methods," **Journal of Posthumanism**, vol. 4, no. 3, pp. 543–559, 2025. doi: 10.63332/joph.v4i3.2686.
- S. K. Sahani and B. K. Sah, "Optimization of cultural education program using numerical techniques," **Journal of Posthumanism**, vol. 5, no. 6, pp. 4265–4280, 2025. doi: 10.63332/joph.v5i6.2600.
- S. K. Sahani, "Application of numerical methods in structural health monitoring using IoT sensors," **Journal of Electrical Systems**, vol. 19, no. 1, pp. 194–207, 2023. doi: 10.52783/jes.8941.
- S. K. Sahani and K. S. Prasad, "Modeling the impact of cultural education policies on literacy rates using numerical differential equations," **Journal of Posthumanism**, vol. 3, no. 3, pp. 376–394, 2023. doi: 10.63332/joph.v3i3.2905.
- S. K. Sahani and R. H. Mandal, "Application of logistic regression and numerical methods for automated bot detection in social media networks," **Linguistic and Philosophical Investigations**, vol. 22, no. 2, pp. 192–206, 2023.
- S. K. Sahani, M. K. Sah, S. K. Agarwal, V. Pareek, V. Pareek, and V. V. Singh, "Numerical simulation of cultural value conflict in multicultural societies," **Journal of Posthumanism**, vol. 5, no. 3, pp. 1815–1831, 2025, doi: 10.63332/joph.v5i3.2651.