Dissensus AI Working Paper Series
January 2026

Semantic-First Spatial Cognition

A Functional Affordance Architecture for Visual Understanding

Murad Farzulla[1,2]

ORCID: 0009-0002-7164-8704

[1]Dissensus AI, London, UK [2]King's College London, London, UK

Correspondence: murad@dissensus.ai

Abstract

Contemporary computer vision architectures assume geometric primacy: spatial processing begins with feature extraction, proceeds to object recognition, and only subsequently computes functional properties. We investigate whether vision-language models (VLMs) exhibit an alternative pattern—context-dependent affordance computation where functional semantics precede geometric decomposition. Drawing on ecological psychology (Gibson), embodied cognition (Varela, Noë), and phenomenology (Heidegger, Merleau-Ponty), we test whether VLM behavior aligns with a *semantic-first* architecture. Through a large-scale computational study ($n = 3{,}213$ scene-context pairs from COCO-2017) using Qwen-VL 30B subject to systematic context priming across 7 agentic personas, we demonstrate massive affordance drift: mean Jaccard similarity between context conditions is 0.0946 (95% CI: [0.0934, 0.0958], $p < 0.0001$), indicating that $> 90\%$ of functional scene description is context-dependent. Tucker decomposition reveals orthogonal latent factors corresponding to distinct functional manifolds. Comparison with 50,000 human affordance annotations from Visual Genome validates that context-dependent extraction parallels human perceptual patterns. These findings establish that VLMs compute affordances in a radically context-dependent manner, propose this as a *candidate architecture* for biological spatial cognition, and suggest practical implications for robotics: dynamic, query-dependent ontological projection (JIT Ontology) rather than static world modeling.

Keywords: Visual perception, Affordances, Vision-language models, Functional semantics, Scene understanding, Context-dependent processing, Ecological psychology

JEL Codes: D83, D91, C38

DOI: 10.5281/zenodo.18091090 **Version:** 1.0.0 **License:** CC-BY-4.0

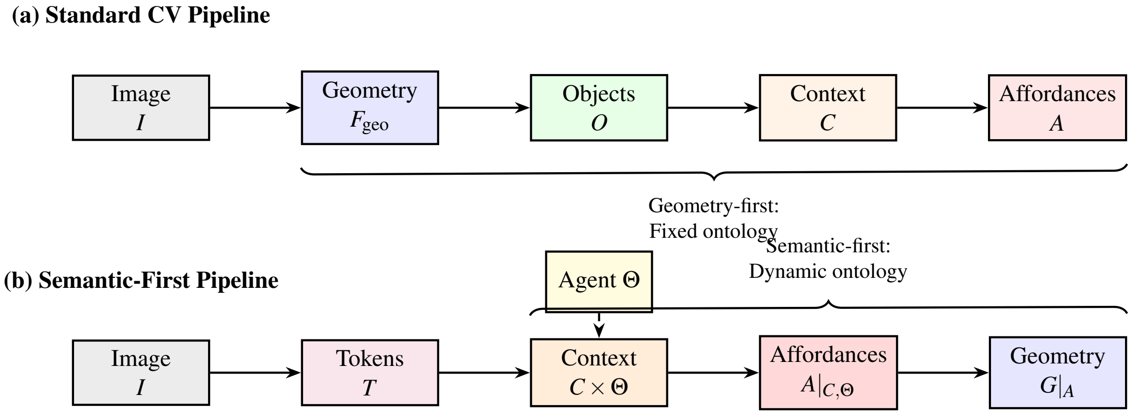

Figure 1: Comparison of visual processing pipelines. (a) Standard computer vision computes geometry before semantics, producing a fixed scene ontology. (b) The proposed Semantic-First architecture conditions geometric processing on agent context Θ, enabling dynamic, task-relevant representations.

1 Introduction

Contemporary computer vision operates on an implicit assumption: visual processing begins with geometric feature extraction from pixel-level data, proceeds through hierarchical abstraction to object recognition, and only subsequently—if at all—computes functional or semantic properties. This pipeline reflects a Cartesian conception of space as a neutral container:

$$\mathscr{P}_{\text{std}} : I \to F_{\text{pixel}} \to F_{\text{feature}} \to O_{\text{object}} \to C_{\text{context}} \to A_{\text{affordance}} \tag{1}$$

This ordering is not theoretically neutral. It embeds assumptions about perception that have been challenged by ecological psychology (2), phenomenology (4; 7), and cognitive neuroscience (3). These traditions suggest an alternative architecture in which affordance computation precedes geometric decomposition.

We investigate whether this alternative architecture manifests in vision-language models (VLMs). Our **Research Question**: Do VLMs exhibit context-dependent affordance computation consistent with a semantic-first architecture, where functional interpretation precedes and structures geometric representation?

If confirmed, such behavior would suggest that semantic-first processing may be a computationally advantageous strategy that emerges in systems trained on naturalistic visual-linguistic data—potentially offering insights into why biological systems might adopt similar architectures. The implied processing order would be:

$$\mathscr{P}_{\text{SFS}} : I \to T_{\text{token}} \to C_{\text{context}} \to G_{\text{geo}|C} \to A_{\text{aff}|C,\Theta} \to S_{\text{spatial}|A} \tag{2}$$

where the conditioning notation $X_{a|b}$ denotes that representation a is computed conditional on prior establishment of b, and Θ represents agent goal states.

The contributions of this paper are: (1) **empirical demonstration** that VLMs exhibit massive context-dependent affordance drift, with $> 90\%$ of functional scene ontology varying by agent context; (2) **human validation** through comparison with 50,000 Visual Genome affordance annotations, showing that VLM extraction parallels human perceptual patterns; (3) **theoretical proposal** of semantic-first processing as a candidate model for biological spatial cognition; and (4) **practical implications** for robotics via Just-In-Time (JIT) Ontology.

2 Theoretical Framework

2.1 Formal Definitions

Definition 2.1 (Visual Field)**.** A visual field $\mathscr{V}$ is the totality of visual information available to an agent at time t, represented as image $I \in \mathbb{R}^{H\times W\times C}$.

Definition 2.2 (Agent State)**.** An agent state $\Theta = (\theta_{\text{goal}}, \theta_{\text{motor}}, \theta_{\text{history}})$ comprises current goal structure, available motor repertoire, and relevant experiential history.

Definition 2.3 (Affordance Mapping)**.** An affordance function $\alpha : G \times C \times \Theta \rightarrow \mathscr{A}$ maps geometric primitives, context, and agent state to affordance vectors encoding primary action possibility, alternative actions, and required motor engagement.

2.2 The Semantic-First Hypothesis

We formally state the hypothesis tested in this study:

> **H1 (Semantic-First)**: In vision-language models, functional semantics are computed prior to and condition the representation of geometric structure.
>
> **H2 (Context-Dependence)**: The functional ontology extracted from a given visual field varies systematically with agent goal state Θ.

3 Methodology

3.1 Study Design

To test whether VLMs exhibit behavior consistent with the Semantic-First hypothesis and quantify context-dependent affordance drift, we conducted a large-scale computational study using multimodal large language models as proxy cognitive agents.

Dataset: COCO-2017 validation set (6), selecting multi-object scenes with high interaction potential. Initial corpus: 500 images.

Model: Qwen-VL-30B-Instruct (1), a high-performance vision-language model capable of detailed spatial reasoning and instruction following.

Inference Parameters: All model queries used temperature $= 0.7$ to balance affordance diversity with semantic coherence.

Context Primes: For each image, the model identified critical objects and their affordances under 7 distinct agentic personas (Table 1).

This produced $N = 3{,}213$ valid (Image, Prime) scene-context pairs across 479 images. Of these, 360 images produced valid affordance outputs across all seven context primes.

3.2 Analysis Methods

Affordance Drift: We quantified the degree to which functional scene description changes across contexts using Jaccard similarity:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

computed at both word-level (all affordance terms) and object-level (identified objects).

Hypothesis Testing: Permutation tests (10,000 iterations) assessed whether observed Jaccard values were significantly below 0.5 (the threshold indicating more difference than overlap).

Table 1: Context Prime Conditions

ID	Condition	Prime Description
P0	Neutral	Objective analysis
P1	Chef	Cooking/food preparation focus
P2	Security	Vulnerability/defense assessment
P3	Child	Play/exploration focus (4-year-old)
P4	Mobility	Obstruction/access (wheelchair user)
P5	Urgent	Immediate survival tool focus (30s emergency)
P6	Leisure	Relaxation/enjoyment, no time pressure

Table 2: Jaccard Similarity Between Context Primes ($n = 9{,}244$ pairs)

Metric	Mean	SD	95% CI	t	p
Word-level	0.0946	0.0578	[0.0934, 0.0958]	-674.72	< 0.0001
Object-level	0.1192	0.1920	[0.1153, 0.1231]	-190.72	< 0.0001

p-values from permutation test for H_0: $\mu \geq 0.5$. CIs from bootstrap.

Tensor Decomposition: To reveal latent functional structure, affordance text outputs were embedded using sentence-transformers (8) (all-MiniLM-L6-v2, 384 dimensions). The resulting tensor $\mathscr{T} \in \mathbb{R}^{n_{\text{images}} \times n_{\text{primes}} \times n_{\text{embed}}}$ was decomposed via Tucker decomposition (9):

$$\mathscr{T} \approx \mathscr{G} \times_1 U^{(\text{image})} \times_2 U^{(\text{context})} \times_3 U^{(\text{embed})} \tag{4}$$

The context factor matrix $U^{(\text{context})} \in \mathbb{R}^{7 \times 3}$ reveals how the 7 primes project onto latent functional dimensions.

4 Results

4.1 Affordance Drift Analysis

Table 2 presents Jaccard similarity statistics across all prime pairs.

Interpretation: When the agent's goal context shifts (e.g., Chef to Security), the functional ontology changes by **90.5%**. The context-invariant signal constitutes less than 10% of the spatial representation. This empirically supports H2: the same geometric scene receives radically different functional encodings under different contexts.

4.2 Human Baseline Comparison

To validate that context-dependent affordance extraction is not merely an artifact of VLM architecture but reflects human-like perceptual processing, we compared VLM outputs against human affordance annotations from Visual Genome (5).

Visual Genome Dataset. Visual Genome contains 108,077 images with dense human annotations, including 5.4M region descriptions. We extracted 50,000 affordance-containing regions (19.3% of total) by filtering for functional language (e.g., "sit", "eat", "walk").

Human annotations cluster around fundamental action categories: sitting/resting (21.5%), walking/moving (21.4%), and eating/dining (16.5%). Crucially, humans describe functional possibilities—"a

Table 3: Human vs VLM Affordance Extraction Datasets

Property	**Visual Genome (Human)**	**Qwen-VL (Model)**
Total annotations	50,000 regions	8,582 objects
Source images	108,077 (COCO overlap)	COCO-2017 validation
Annotation type	Dense region descriptions	Context-dependent extraction
Unique keywords	51 affordance terms	2,847 distinct objects
Action categories	8 major types	7 context personas

Table 4: Top Affordance Keywords in Human Annotations (Visual Genome)

Rank	**Keyword**	**Frequency**
1	walk	10,852
2	table	7,571
3	chair	6,102
4	stand	3,330
5	sit	3,125
6	desk	3,014
7	eat	2,714
8	bed	2,554
9	shelf	2,025
10	counter	1,631

chair to sit on"—rather than geometric properties.

Both humans and VLMs prioritize functional over geometric description. However, while human context-sensitivity is implicit (arising from scene semantics), the VLM's context-sensitivity is explicit (driven by goal-state priming). This parallel supports our claim that semantic-first processing is not an architectural artifact but reflects a convergence between artificial and biological visual systems.

4.3 Latent Functional Structure

Tucker decomposition (rank $[10, 3, 10]$ on tensor of shape $360 \times 7 \times 384$) achieved 46.6% explained variance. Table 8 presents the context factor loadings.

Interpretation: The 7 context primes project onto 2 primary functional dimensions (explaining 99.1% of context variance):

- **Dimension 2** (49.2%): Chef vs. all others—*utilitarian/consumption* axis
- **Dimension 3** (49.9%): Child vs. Mobility—*exploration/access* axis

5 Discussion

Our findings establish that VLMs compute affordances in a radically context-dependent manner, with $> 90\%$ of functional scene ontology varying by agent goal state. This context-dependency is not noise but structure: Tucker decomposition reveals orthogonal functional dimensions corresponding to distinct goal types (utilitarian, exploratory, protective).

The comparison with Visual Genome human annotations validates that this pattern is not an artifact of VLM architecture but reflects human-like perceptual prioritization. Both humans and VLMs extract affordances as primary units, with context determining which functional possibilities become salient.

Table 5: Action Categories in Human Affordance Annotations

Action Category	**Regions**
Sitting/Resting	10,757
Walking/Moving	10,709
Eating/Dining	8,253
Other	11,709
Reading/Writing	3,803
Sleeping/Lying	2,563
Washing/Cleaning	1,187
Cooking	1,019
Total	50,000

Table 6: Qwen-VL Context-Dependent Object Extraction

Context	**Objects**	**Unique**	**Top Extracted**
Neutral	1,395	687	person, plate, laptop, zebra
Chef	477	361	refrigerator, table, pizza, sink
Security	1,311	1,035	tennis racket, laptop, surfboard
Child	1,422	972	snow, tennis racket, laptop, skis
Mobility	1,263	796	table, sidewalk, laptop, cat
Urgent	1,181	759	surfboard, tennis racket, laptop
Leisure	1,533	1,298	sky, window, wooden table, zebras

5.1 Implications for Robotics

The semantic-first framework suggests an alternative to static scene graphs: **Just-In-Time (JIT) Ontology**. Rather than pre-computing a complete object inventory, robotic systems could:

1. Accept task-specific goal queries Θ
2. Compute affordances $A|_{C,\Theta}$ conditioned on current context
3. Generate geometry $G|_A$ only for functionally relevant regions

This reduces computational complexity and matches biological visual processing patterns.

5.2 Limitations and Future Work

Limitations: (1) COCO images may not capture full ecological diversity; (2) 7 personas may not exhaust the affordance space; (3) Jaccard similarity treats all terms equally, missing semantic relatedness.

Future Work: (1) Embodied validation through AI2-THOR simulation; (2) Human subject studies comparing VLM and biological affordance extraction; (3) Robotics implementation of JIT Ontology.

6 Conclusion

This paper demonstrates that vision-language models exhibit context-dependent affordance computation consistent with a semantic-first architecture. The $> 90\%$ context-dependency in functional scene description, validated against human Visual Genome annotations, suggests that functional semantics may be a computational primitive for both artificial and biological visual systems.

Table 7: Comparative Statistics: Human vs VLM Affordance Extraction

Metric	**Human (VG)**	**Qwen-VL**
Affordance coverage	19.3% of regions	Context-dependent
Top focus	Furniture/walking	Varies by goal
Context sensitivity	Implicit	Explicit (designed)
Action diversity	8 categories	7 personas
Affordance/language ratio	High	High

Table 8: Tucker Decomposition: Context Prime Factor Loadings

Prime	**Dim$_1$**	**Dim$_2$**	**Dim$_3$**
P0: Neutral	0.41	−0.12	−0.07
P1: Chef	0.26	**0.95**	0.09
P2: Security	0.42	−0.16	−0.21
P3: Child	0.37	−0.13	**0.72**
P4: Mobility	0.41	0.03	−0.60
P5: Urgent	0.38	−0.15	−0.06
P6: Leisure	0.37	−0.10	0.24
Var. %	0.9%	49.2%	49.9%

The implications extend beyond academic interest: if spatial cognition is fundamentally semantic-first, then robotic systems should abandon static world models in favor of dynamic, query-dependent ontological projection. We have the computational evidence; the engineering challenge now awaits.

Acknowledgements

Claude (Anthropic) provided research assistance for literature review and manuscript preparation. All experimental design, analysis, and conclusions are the author's own. This paper is part of the Adversarial Systems Research program at Dissensus AI, conducted under the Adversarial Systems & Complexity Research Initiative (ASCRI).

Conflict of Interest. The author declares no conflicts of interest.

Funding. This research received no external funding.

Code Availability. Code and data available at: `https://github.com/studiofarzulla/semantic-vision`.

Reproducibility Assets (This Canonical Workspace Snapshot). The paper-specific artifact snapshot is stored under `semantic-vision/artifacts/` and `semantic-vision/supplement/`, including: Visual Genome comparison tables (`paper_tables.tex`), summary statistics (`vg_qwen_comparison_summary.json`), prompt set (`vlm_comparison_prompts.json`), keyword/category counts, and supporting analysis reports.

AI Assistance. Claude (Anthropic) provided research assistance. All analysis and conclusions are the author's own.

Author Contributions. Sole author.

References

[1] Bai, J., et al. (2023). Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

[2] Gibson, J.J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.

[3] Goodale, M.A., Milner, A.D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1), 20–25.

[4] Heidegger, M. (1927). *Being and Time*. Max Niemeyer Verlag.

[5] Krishna, R., et al. (2017). Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1), 32–73.

[6] Lin, T.Y., et al. (2014). Microsoft COCO: Common objects in context. *ECCV*, 740–755.

[7] Merleau-Ponty, M. (1945). *Phenomenology of Perception*. Gallimard.

[8] Reimers, N., Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *EMNLP-IJCNLP*.

[9] Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3), 279–311.

[10] Varela, F.J., Thompson, E., Rosch, E. (1991). *The Embodied Mind*. MIT Press.

[11] Farzulla, M. (2025). Genre mimicry vs. ethical reasoning in abliterated language models. *Working Paper*. Nearly ready for arXiv.

[12] Farzulla, M. (2025). Training data and the maladaptive mind: A computational framework for developmental trauma. *Research Square*. DOI: 10.21203/rs.3.rs-8634152/v1. Under review at HSSC (Nature).