Assignment: Research Paper
MULTIMODAL SPATIAL INTELLIGENCE

Annie Burke
Department of Architecture, Florida International University
ARC7980: Doctoral Project; Class 80623; Section RVC
Professor Neil Leach
December 6, 2025

ABSTRACT: Generative models increasingly construct spatial environments from incomplete, modality-specific inputs such as appearance, depth, motion, and semantics. Because each modality supplies only part of the spatial evidence, generative systems infer structure through modality-dependent constraint sets rather than fused multimodal cues. Biological spatial intelligence, by contrast, integrates heterogeneous cues into a single, continuously updated spatial hypothesis grounded in ecological invariants. These architectural differences produce persistent representational divergence: humans maintain coherence through invariant-based expectations, while models rely on statistical priors that may not preserve continuity, identity, or geometric stability.
As synthetic environments proliferate, their spatial tendencies enter human perceptual experience, creating hybrid spatial expectations and requiring coordination between fundamentally different representational logics. This coordination functions as an adaptive negotiation shaped by model-side constraints, human interpretive routines, and tool-mediated interaction. Interactive spatial inference—where human input functions as an additional modality—further shapes how coherence is maintained or adjusted in hybrid environments. This framework clarifies how modality shapes spatial structure in generative systems and how these structures influence human interpretation across physical, digital, and synthetic contexts.

MULTIMODAL SPATIAL INTELLIGENCE

1 INTRODUCTION

Computational models create environments whose spatial structure is constrained by interactions among their input modalities (Yang et al., 2024). Because particular modality configurations supply their own spatial cues, models infer different forms of spatial structure depending on which signals they receive (Mildenhall et al., 2020; Sajjadi et al., 2022).
Humans can interpret an AI-generated space to the extent that the model's internal spatial logic makes sense. A model's spatial logic does not come only from its input cues (RGB, depth, motion, semantics, etc.) but also from the architectural decisions that determine how those cues are coordinated or separated (fusion modules, attention patterns, loss functions, training regimes, etc.). Because these interactions and architectural choices are encoded in ways that remain largely uninterpretable, our understanding of the spatial forms produced by generative models—and how humans will interpret them—remains limited. Recent creative workflows also reveal that real-time human intervention can act as an additional input modality, altering constraint weighting during inference and contributing directly to spatial outcomes.
This gap motivates the following research questions:
RQ1: How do different systems—biological and computational—form spatial intelligence from incomplete cues?
RQ2: How do different combinations of input modalities shape spatial structures that generative models can infer?
RQ3: How will differences between biological spatial intelligence and model-generated spatial structure influence one another?
This paper examines how generative models construct spatial structure by analyzing the spatial cues supplied by each modality, the interactions amongst those modalities, and the architectural mechanisms that coordinate those cues. Together, these factors shape the structures a model infers.

input modality → constraints → constraint interaction → constraint alignment or separation → architecture-shaped structure → representation → human interpretation

2 FOUNDATIONS OF SPATIAL INTELLIGENCE

2.1 Definition of Spatial Intelligence

Before examining how generative models build spatial structure, this section defines spatial intelligence as the underlying capacity that enables systems to infer structure from limited information.

Biological Spatial Intelligence:

Biological perceptual systems, such as those of humans and other animals, fuse heterogeneous sensory cues into a coherent-enough spatial hypothesis (Spelke et al., 2007; O'Keefe & Nadel, 1978). They produce unified spatial representations from partial and noisy input, relying on approximations rather than perfect alignment.

Computational Spatial Intelligence:

Computational systems construct spatial structure from explicit constraints such as sensors, inputs, or latent patterns (Mildenhall et al., 2020; Sajjadi et al., 2022; Yang et al., 2024). They do not inherently fuse constraints; instead, structure emerges from the particular signals an architecture can absorb or align through design, reflecting architecture-dependent coordination rather than cue fusion. Interactive spatial inference introduces an additional condition: real-time human input, which can serve as a modality that modulates constraint weighting during generation.

2.2 Spatial Structure, Coherence, and the Mechanisms that Support Spatial Inference

Spatial structure refers to the latent organization of surfaces, geometry, boundaries, and spatial relations that underlie any environment. Because no single observation provides complete access to this organization, biological and computational systems encounter only partial or projected information and must infer a structural hypothesis from incomplete cues (Gibson, 1979; O'Keefe & Nadel, 1978; Mildenhall et al., 2020; Poole et al., 2022; Eslami et al., 2018; Sajjadi et al., 2022). Foundational accounts identify perceptual primitives—surface layout, depth relations, occlusion structure, and geometric configuration—that provide evidence to constrain inference (Gibson, 1979; Ullman, 1984). These primitives function as a constraint system, ruling out spatial configurations that violate physical layout, boundary relations, or geometric viability (Spelke & Kinzler, 2007; Gibson, 1979).

Human spatial intelligence depends on this constrained structure because coherent-enough spatial relations reduce interpretive effort and allow relational computation to be offloaded to the environment rather than reconstructed internally (Clark, 1997; Gibson, 1979). Sufficient spatial organization supports prediction and planning by providing dependable cues about how objects and surfaces will behave over time and during movement (O'Keefe & Nadel, 1978), and enables narrative continuity by allowing people to track actions and viewpoints without recalculating layout changes (Ullman, 1984; Spelke & Kinzler, 2007). Together, these properties enable humans to maintain a coherent-enough world from incomplete information.

Biological spatial intelligence relies on perceptual and representational processes that maintain spatial relations across movement, while computational spatial intelligence depends on constraint coordination shaped by architecture and input conditions. This contrast frames later analysis of how coherence emerges in hybrid environments.

3 INPUT MODALITIES AND THE SPATIAL INFORMATION THEY PROVIDE

3.1 What an Input Modality Is

Input modalities are the channels through which a system receives spatial information, each providing a particular signal type that exposes a distinct subset of available spatial cues (Gibson, 1979). A modality's signal type determines the kind of spatial evidence it supplies and thus shapes which aspects of spatial structure a system can detect, infer, or reconstruct (Ullman, 1984; Spelke & Kinzler, 2007). Because these modalities carry fundamentally different forms of spatial content, no single modality provides full access to the underlying structure of a scene; each reveals specific structural properties while leaving others unspecified, establishing the initial conditions under which spatial inference proceeds (Gibson, 1979; Mildenhall et al., 2020; Sajjadi et al., 2022).

Input modalities vary by the cues they disclose—appearance, geometry, motion, continuity, or semantics—and each cue type imposes different constraint conditions on spatial inference. Because each modality contributes a distinct pattern of evidence, the modality configuration jointly determines which spatial hypotheses are viable at the outset. Downstream inference can refine these conditions but cannot override the structural limitations set by the inputs (Eslami et al., 2018). Together, the available modalities define the interpretive space from which a coherent-enough spatial structure can be inferred.

3.3 Modalities in Biological Systems

Biological perceptual systems extract spatial information through the simultaneous intake of multiple signals, with each channel providing complementary cues that jointly reveal structural properties no single source can specify alone (Gibson, 1979; Spelke & Kinzler, 2007). These heterogeneous cues are automatically fused into a single, coordinated spatial representation rather than maintained as separate streams, enabling stable enough interpretation across changing viewpoints and conditions (O'Keefe & Nadel, 1978).

3.4 Modalities in Computational Systems

Computational systems receive spatial information through modality-indexed input channels, each supplying its own signal type (Eslami et al., 2018). Unlike biological perception, these channels are not inherently aligned and enter the system as separate streams with no built-in correspondence (Yang et al., 2025). A single channel rarely supports full spatial inference, so the system begins with fragmented, modality-specific views of the scene (Mildenhall et al., 2020).

Integration occurs only when architectures provide explicit fusion or alignment mechanisms—such as shared latent spaces, cross-modal attention, correspondence objectives, or supervision signals (Sajjadi et al., 2022). Without such mechanisms, each modality constrains interpretation independently, and spatial inference unfolds in a multi-stream interpretive space where constraint sets remain only partially coordinated (Poole et al., 2022). The spatial structure a computational system infers, therefore, reflects how its modality configuration interacts with the system's coordination mechanisms. Coherence arises only when the model's coordination mechanisms impose shared structure or when statistical regularities produce emergent alignment (OpenAI, 2024).

3.5 Structural Claim: Why Modality Matters

Spatial structure provides the organizational substrate that interpretation, prediction, and action require (Gibson, 1979; O'Keefe & Nadel, 1978). Generative systems rely on this substrate to construct usable, interpretable environments; the spatial organization a model produces sets the practical boundaries of the worlds it generates, shaping how those environments can be understood and navigated by human users (Mildenhall et al., 2020; Sajjadi et al., 2022).

As learning, design, and analysis increasingly occur within synthetic environments, the spatial patterns embedded in these environments influence how people interpret, navigate, and make decisions. Because humans rely on spatial cues to reduce interpretive effort and maintain continuity (Gibson, 1979; Clark, 1997), variations in generated structure incrementally expand the interpretive repertoire humans apply to layout, identity, and spatial relations (OpenAI, 2024; Qin et al., 2025). Regularities in synthetic spatial structure gradually influence the assumptions users apply across digital and physical contexts (Yang et al., 2025).

Modality configuration determines the kinds of spatial structures generative systems can produce. Each modality supplies a distinct pattern of cues and omissions (Mildenhall et al., 2020; Poole et al., 2022), and different combinations open or close different portions of the interpretive space. These modality-governed constraint sets determine whether a system converges on a spatial hypothesis or a characteristic structural pattern (Sajjadi et al., 2022). As a result, modality configuration becomes a primary driver of a model's spatial strengths and limitations, establishing the interpretive conditions analyzed in the next section.

4 CONSTRAINTS AND INTERPRETIVE REDUCTION

4.1 Constraints, Modality-Dependent Evidence, and Interpretive Reduction

Constraints define the informational conditions under which spatial inference can occur. A constraint is any cue-dependent condition that reduces the number of structural interpretations a system must consider (Ullman, 1984; Gibson, 1979). By eliminating configurations that conflict with the cues provided by a modality, constraints bound the range of viable hypotheses and establish the system's initial interpretive space (Spelke & Kinzler, 2007).

This cue-driven elimination process—interpretive reduction—removes spatial hypotheses that conflict with geometric, temporal, semantic, or relational evidence (Gibson, 1979; Ullman, 1984). It reduces the ample, undifferentiated hypothesis space into one structured by cue-dependent constraints, establishing the baseline conditions under which spatial structure becomes inferable (Sajjadi et al., 2022). Constraints are not obstructions but informational filters that clarify which structures are possible given available cues.

Input modalities determine which constraints a system begins with, since each modality reveals only a subset of spatial information. Different cue types expose different structural properties, establishing the initial conditions under which inference unfolds. Because modalities differ in the constraints they impose, each modality shapes its own region of the interpretive space.

When multiple modalities are available, their constraint sets interact rather than accumulate. They may reinforce one another by eliminating the same alternatives, conflict by eliminating incompatible possibilities, operate in independent regions of the hypothesis space, or cancel when one modality neutralizes constraints imposed by another (Sajjadi et al., 2022). These multimodal dynamics determine how much of the interpretive space becomes structured, how much remains unresolved, and which structural hypotheses remain viable for subsequent inference.

4.2 What Structural Coherence Requires

Coherence emerges when constraints eliminate enough incompatible spatial interpretations for a workable organizational pattern to form (Ullman, 1984). Because input cues are partial and often noisy, coherence depends not on perfect information but on a sufficiently reduced hypothesis space in which one interpretation explains the evidence better than its alternatives (Gibson, 1979).

This reduction is achieved when constraint sets provide enough overlapping support for a system's coordination mechanisms to organize them into a coherent-enough structure. Sufficient cue compatibility allows the system to suppress alternatives even when some signals are ambiguous or partially misaligned (Spelke & Kinzler, 2007). Constraint interactions shape the available structural possibilities: they may reinforce one another, conflict, or cancel depending on how eliminations combine across modalities (Sajjadi et al., 2022). Coherence emerges when overlapping eliminations collectively reduce the hypothesis space enough for one structurally plausible organization—or a closely related set of organizations—to dominate the system's interpretation (O'Keefe & Nadel, 1978).

4.3 Patterns of Constraint Coordination

Biological perceptual systems coordinate spatial constraints through continuous multimodal fusion. Cue combinations are integrated in parallel, allowing constraints to jointly support one coherent-enough interpretation (Gibson, 1979; Spelke & Kinzler, 2007). Conflicting cues are managed through complementarity and selective weighting, while residual ambiguities remain contained within a unified representational frame (O'Keefe & Nadel, 1978).

Computational systems coordinate constraints through architecture-dependent mechanisms rather than inherent fusion. Modality-specific inputs enter as separate streams with no built-in alignment (Eslami et al., 2018; Yang et al., 2025). Coherence emerges only when the system's coordination mechanisms impose alignment via shared latent spaces, cross-modal attention, or learned correspondences (Sajjadi et al., 2022; Mildenhall et al., 2020). Some models exhibit emergent partial alignment through training dynamics or statistical regularities (OpenAI, 2024; Hu et al., 2025), producing coherence within the model's representational logic.

Across systems, constraint coordination can follow multiple patterns. When heterogeneous cues align, their eliminations converge on mutually compatible interpretations, narrowing the hypothesis space in a coordinated way (Spelke & Kinzler, 2007; Ullman, 1984). When cues remain unfused, each restricts interpretation within its own representational domain, and coherence depends on the system's ability to relate these domains effectively (Mildenhall et al., 2020). Alignment and separation reflect the system's structure and objectives, determining how much of the spatial hypothesis becomes shared across constraints and how much remains distributed.

Because systems coordinate constraints according to their own organizing logic, they arrive at spatial interpretations with distinct coherence profiles (Poole et al., 2022; Qin et al., 2025). Biological systems tend toward unified hypotheses; computational systems tend toward distributed representations shaped by modality configuration and system design (Sajjadi et al., 2022; Mildenhall et al., 2020).

5 REPRESENTATION

5.1 Biological Spatial Representation

Biological perceptual systems form spatial representations by integrating multimodal constraints into a single coherent-enough hypothesis (Gibson, 1979; Spelke & Kinzler, 2007). Coherence emerges not from perfect alignment but from overlapping constraint support: conflicting cues are handled through complementarity or selective weighting, and ambiguities remain localized (Spelke & Kinzler, 2007). This enables perception, prediction, and action to be guided by a dominant spatial account that is stable enough for behavior even under partial or noisy input (Gibson, 1979). Biological fusion relies on heuristics that prioritize efficiency over exhaustive accuracy (O'Keefe & Nadel, 1978).

Spatial hypotheses are updated incrementally, adjusting specific relations while retaining the broader organizing structure, supporting fast interpretive cycles across changing viewpoints (Ullman, 1984). Biological systems also coordinate egocentric and allocentric reference frames: egocentric components track near-body positions, movement

directions, and immediate affordances, while allocentric components encode stable, observer-independent layouts and landmark relations (O'Keefe & Nadel, 1978). These frames interact continuously, with egocentric updates interpreted relative to allocentric structure and allocentric maps refined through moment-to-moment experience. Through incremental updating and reference-frame interaction, biological systems maintain a coherent sense of the environment across both immediate interactions and extended activity (Gibson, 1979; Spelke & Kinzler, 2007). This fused representation is not flawless or complete, but it is an efficient and resilient strategy for producing actionable spatial coherence under real-time perceptual constraints.

5.2 Computational Spatial Representation

Computational systems represent space by encoding modality-specific constraint sets into separate latent components rather than fusing them into a unified spatial hypothesis (Spelke & Kinzler, 2007; O'Keefe & Nadel, 1978). Because each modality carries a different cue profile, each latent component provides only a partial structural account, producing a distributed representational workspace (Mildenhall et al., 2020; Sajjadi et al., 2022; Poole et al., 2022; OpenAI, 2024). Coherence is therefore not an automatic consequence of cue fusion but a contingent outcome of whether the architecture can enforce correspondence across these components (Sajjadi et al., 2022). When alignment mechanisms are absent, modality-specific representations can remain internally consistent yet weakly coupled, yielding only loose cross-modal spatial agreement (Sajjadi et al., 2022). When alignment mechanisms are present—shared latent spaces, cross-modal attention, learned correspondences, or supervision signals—they impose relationships among components and produce coordinated spatial organization (Mildenhall et al., 2020). In this sense, the inferred spatial structure reflects the subset of constraints the system can absorb and the coordination processes available to relate them, rather than the inherent organization of the external world (Mildenhall et al., 2020; Sajjadi et al., 2022; OpenAI, 2024).

5.3 Systems Diverge In How They Build and Maintain Spatial Structure

Biological and computational systems construct and maintain spatial structure in different ways, so they arrive at different accounts of coherence, identity, and continuity—setting up the interpretive consequences in Section 6. Biological systems maintain spatial coherence by fusing heterogeneous cues into a unified spatial hypothesis constrained by ecological invariants (Gibson, 1979; Spelke & Kinzler, 2007). Generative systems, by contrast, assemble structure from distributed, modality-indexed latent components shaped by learned priors; coherence depends on whether the architecture can enforce correspondence across those components (Mildenhall et al., 2020; Poole et al., 2022). The same scene can therefore yield a single dominant interpretation for human perception while supporting multiple structurally plausible hypotheses for a model. In shared environments—AR/VR, simulation, design workflows, and hybrid human–AI contexts—this mismatch becomes operational: humans expect coherent-enough continuity, while models can treat unresolved ambiguity as structural uncertainty, producing different behavior around geometry, boundaries, and identity.

5.4 Interactive Spatial Inference

Interactive spatial inference describes how spatial structure in hybrid environments emerges from ongoing coordination between a model's multimodal constraint sets and real-time human intervention. Human adjustments—whether through prompting, node manipulation, camera steering, or constraint tuning—function not as new sensory modalities but as operations that modulate constraint weighting, selectively reinforcing some cues while suppressing or overriding others. This modulation reshapes the inference trajectory, producing spatial outcomes that the system did not generate autonomously.

Interactive spatial inference, therefore, operates as a hybrid constraint loop. The model proposes a provisional structure conditioned by its modality configuration; human intervention perturbs that structure by altering the distribution of viable hypotheses, and the system re-infers local or global organization in response. Across these iterative cycles, coherence becomes co-produced rather than internally generated: neither participant fully determines the spatial hypothesis, yet both shape its evolution.

This interactive loop also makes the constraint dynamics described in Section 4 observable in practice. Human adjustments may reinforce model tendencies, cancel unstable cues, introduce conflicts that force reorganization, or isolate independent regions of the hypothesis space. As these interactions accumulate, they produce the hybrid continuity seen in contemporary creative workflows—a spatial logic reflecting both ecological expectations and model-shaped regularities. Interactive spatial inference thus constitutes a distinct representational condition,

increasingly central as synthetic tools enter design, analysis, and worldbuilding workflows, and will play a growing role in how hybrid human–AI systems generate, evaluate, and communicate spatial structure.

6 CONSEQUENCES

6.1 Biological Expectations and Synthetic Departures

Human expectations about continuity, geometry, and causality follow from biological mechanisms that infer coherent structure from incomplete input. Perceptual systems register surfaces, layout, occlusion relations, and environmental invariants (Gibson, 1979), and early perceptual capacities establish durable expectations for cohesion, persistence, and identity across time (Spelke & Kinzler, 2007; Spelke et al., 2007). Together, these mechanisms set a baseline for what counts as a plausible spatial world.

Synthetic spatial environments introduce additional patterns shaped by statistical learning rather than ecological regularity (OpenAI, 2024; Hu et al., 2025; Yang et al., 2025). Because the cues in these scenes can agree with one another without being governed by real-world constraints, the output can read as spatially coherent even when it violates ecological invariants. The departures that result are recognizable and repeatable, giving viewers stable targets for learning. Over repeated exposure, primarily through practical use, viewers learn to interpret these synthetic regularities alongside ecological ones, expanding what is treated as plausible without erasing ecological expectations.

6.2 Structural Bias and Feedback Loops

6.2.1 Models Inject Modality Biases into Synthetic Environments

Because generative systems infer spatial structure through the constraint profiles implied by their available modalities, their outputs carry modality-conditioned spatial biases (Mildenhall et al., 2020; Sajjadi et al., 2022). Different modality profiles weight different evidence—surface appearance, temporal flow, geometric consistency, or semantic relations—so generated scenes tend to express the signals the system can satisfy most reliably. These preferences become visible as recurrent structural signatures: appearance-weighted pipelines often show view-dependent geometry, video-weighted pipelines privilege temporal continuity, and text–image systems reflect cultural or linguistic priors about scene organization (Sajjadi et al., 2022). When multiple modalities are present, these constraint signals interact—reinforcing, suppressing, or skewing one another—rather than automatically fusing into a single unified spatial hypothesis (Mildenhall et al., 2020).

This interaction yields stable-looking tendencies in layout, continuity, and object organization that remain modality-bound. As these signatures propagate through repeated generation and reuse, synthetic environments increasingly embody the representational tendencies of the modality configurations that produced them (OpenAI, 2024; Hu et al., 2025). Section 6.2.2 then traces how this propagation becomes a training feedback loop.

6.2.2 Synthetic Feedback Loops and the Emergence of a Parallel Spatial Grammar

As synthetic outputs are scraped, curated, and reused in training corpora, the modality-conditioned signatures that shaped them re-enter the learning pipeline as data, forming a feedback loop (Hu et al., 2025; Qin et al., 2025; Poole et al., 2022). Through repeated cycles of generation and retraining, these synthetic spatial statistics develop regularities that evolve in parallel with the ecological invariants present in physical environments (Gibson, 1979; O'Keefe & Nadel, 1978). Because physical-world constraints are not inherent to generative architectures, the spatial organization of generated scenes reflects model-internal priors: layout shaped by training distributions, continuity aligned with learned transitions, and identity inferred through statistical association rather than physical persistence (OpenAI, 2024; Hu et al., 2025).

Over successive model generations, these learned regularities shape what becomes typical within the generative ecosystem (Qin et al., 2025), gradually producing a parallel spatial grammar shaped by modality- and architecture-dependent learning. In multimodal models, this drift reflects not only individual modality biases but also the way cues reinforce, suppress, or asymmetrically reshape one another through learned interactions. This grammar forms the model-side representational foundation examined in Section 6.3.

6.3 Divergent Structures, Adaptive Interaction

6.3.1 Structural Divergence Persists Even as Both Systems Adapt

Given the parallel spatial grammar described in 6.2, coordination improves through coadaptation rather than representational convergence. Humans and models can reach "coherent-enough" agreement for a task while relying on different constraint regimes to generate and validate spatial structure. Biological perception stabilizes a unified spatial hypothesis anchored by ecological invariants (Gibson, 1979; Spelke & Kinzler, 2007; O'Keefe & Nadel, 1978),

while generative systems assemble coherence from modality-conditioned components and learned regularities (Mildenhall et al., 2020; Sajjadi et al., 2022; Poole et al., 2022; OpenAI, 2024). *What changes over time is not a shared internal spatial logic, but the interaction practices that make these mismatched logics interoperable.*

As synthetic environments proliferate, users learn recurring model signatures and develop compensatory routines for design, navigation, and analysis (Clark, 1997; Martin-Brualla et al., 2021). These routines externalize constraints where coordination is fragile—stabilizing identity, continuity, and layout through added references, checks, and corrective steps—so outputs remain usable even when underlying representations diverge. Models shift in parallel as they absorb spatial patterns shaped by those workflows and by curated synthetic corpora (Bruce et al., 2024; Hu et al., 2025; Qin et al., 2025). The result is persistent divergence with partial, task-dependent bridges built through use rather than a unified spatial account.

6.3.2 Shared Scenes Elicit Divergent but Navigable Interpretations

When a scene is shared—through prompts, reference media, or interactive outputs—the same cues can support task-level coordination while still stabilizing different underlying spatial hypotheses. Biological perception compresses heterogeneous evidence into a single coherent-enough spatial account anchored by ecological invariants (Gibson, 1979; Spelke & Kinzler, 2007; O'Keefe & Nadel, 1978). Generative systems stabilize "coherence" through learned, modality-conditioned regularities distributed across latent components, so multiple internally permissible 3D arrangements can remain viable from the same cues (Ullman, 1984; Poole et al., 2022; Sajjadi et al., 2022; Mildenhall et al., 2020; OpenAI, 2024; Yang et al., 2025). Coordination remains possible when action is organized around shared descriptors—objects, paths, relations—and agreed constraints, even if the underlying 3D reconstruction is not identical.

Modality-conditioned constraint hierarchies make this divergence systematic rather than incidental. Because different modality profiles privilege different evidence, identical-looking inputs can leave 3D structure underdetermined for a model, supporting multiple internally permissible reconstructions; in multimodal settings, cues can interact—reinforcing, suppressing, or asymmetrically reshaping—rather than fusing automatically into a single unified spatial hypothesis (Mildenhall et al., 2020; Sajjadi et al., 2022). A scene that appears spatially determinate to a human observer can therefore map to several structurally plausible world hypotheses for a generative model (Poole et al., 2022; Sajjadi et al., 2022), while a model-generated scene can remain statistically coherent within its learned distributions yet violate human expectations of persistence, occlusion, or support (OpenAI, 2024; Hu et al., 2025; Qin et al., 2025; Martin-Brualla et al., 2021). The mismatch is a consequence of how coherence is constructed, not a simple performance error.

Coordination becomes workable when users treat divergence as a predictable condition and externalize constraints where model bias is most likely to surface. In practice, this means tightening prompts, adding reference media or additional modalities, and inserting verification steps that stabilize identity, layout, and continuity at the points most prone to drift, so the scene remains usable for the task even when its internal spatial logic remains nonconvergent (Clark, 1997; Martin-Brualla et al., 2021). The result is a shared operational space—sufficient for navigation, design, or analysis—assembled through explicit commitments and checks rather than representational agreement. Section 6.3.3 specifies how those commitments and checks are produced and maintained.

6.3.3 Coordination Emerges in Practice, Even When Systems Differ

Coordination in shared spatial tasks is achieved at the level of commitments and checks, not at the level of shared representation. Humans anticipate stable trajectories, cohesive surfaces, and invariant layout as default constraints on what counts as a coherent-enough world (Gibson, 1979; Spelke & Kinzler, 2007; O'Keefe & Nadel, 1978). Generative models, by contrast, can treat the same partial cues as supporting multiple structurally plausible continuations, with "coherence" grounded in learned statistical regularities rather than ecological invariants (OpenAI, 2024; Yang et al., 2025; Qin et al., 2025). Ambiguity, therefore, lands differently: biological perception tends to collapse uncertainty into a single maintained world-model, whereas model inference can preserve competing hypotheses until an external constraint collapses the alternatives (Ullman, 1984; Poole et al., 2022; Sajjadi et al., 2022). Coordination depends on making those constraints explicit at the points where underdetermination would otherwise reorganize layout, continuity, or identity.

In practice, this explicitness is supplied through a coupled loop of user commitments and interface-mediated checks. Users become fluent in where model-specific tendencies predictably surface and compensate by stating invariants (what must persist) and then tightening prompts, adding references, and inserting verification steps that stabilize

identity, layout, and continuity for the task at hand (Clark, 1997; Martin-Brualla et al., 2021). Tools that expose additional constraint signals—multi-view context, depth cues, action-conditioning, or other mechanisms that bind structure across time and viewpoint—reduce the model's degrees of freedom, translating ecological expectations into model-legible evidence by binding structure across viewpoint and time (Mildenhall et al., 2020; Bruce et al., 2024; Hu et al., 2025). What results is not representational convergence, but an operational middle layer in which humans and models can act as if they share "the same scene" because the interaction provides sufficient constraints to keep their divergent spatial logics interoperable within the bounded demands of the task.

This middle layer is pragmatic and contingent: it stabilizes a scene for the task while leaving representational divergence intact. When constraints are loose, underdetermination reappears and structure can reorganize; when constraints are tight, the output becomes more interoperable with ecological expectations without becoming ecologically grounded. Coordination is therefore an ongoing negotiation in which usability is maintained by repeatedly asserting and testing commitments, not by convergence on a single internal spatial logic.

6.4 Hybrid Spatial Assumptions in Design and Worldbuilding

6.4.1 Hybridization in Professional Practice

Fields that design, simulate, *or automate* space—architecture, VFX/game production, AR/VR, and robotics—already require translation across multiple spatial representations. Architecture moves between plan, section, rendering, and built form; VFX composites physical footage with simulated dynamics; AR/VR systems introduce display-specific depth cues and interaction logics that reorganize how depth, scale, and occlusion are read (Gibson, 1977; O'Keefe & Nadel, 1978; Clark, 1997). What makes these practices workable is not representational uniformity but ongoing translation across projections, engines, conventions, and constraints. Generative systems introduce another grammar into that existing ecology—one governed by modality-shaped assumptions about geometry, continuity, and occlusion (Mildenhall et al., 2020; Sajjadi et al., 2022; Poole et al., 2022; OpenAI, 2024).

In practice, workflows reorganize around where this synthetic grammar is reliable and where it must be corrected, constrained, or treated as provisional. A generated room with an inconsistent back corner can still function as a usable cue—much as wide-angle optics exaggerate depth or a stylized shader encodes non-literal surfaces—so long as the practitioner knows what to trust and what to overwrite (Martin-Brualla et al., 2021). Designers, therefore, patch outputs, re-anchor them with references, or use them as prompts for downstream modeling and simulation, turning recurrent model regularities into manageable boundary conditions. Earlier transitions followed the same pattern—CGI in film, parametric modeling in architecture, physics solvers in animation—each adding new regularities that demanded new interpretive habits without collapsing practice into error (Smelik et al., 2014; Clark, 1997).

Hybridization, on this account, is the accumulation of additional constraint grammars in professional work: an expanded repertoire of coherent-enough spatial cues that must be negotiated and selectively enforced, not a replacement of prior standards.

6.4.2 Hybridization in Human Experience

Human spatial experience routinely adapts to patterned mediation, and synthetic spatial cues now enter that adaptive loop alongside ecological ones. Repeated interaction with predictive systems—autocorrect, shorthand conventions, typing rhythms—trains expectations about completion and correction as normal parts of interpretation, while GPS-assisted navigation shifts how people externalize spatial memory and normalize device-mediated guidance. Comparable dynamics appear in perception: smartphone cameras accustom users to stabilized motion, depth segmentation, and continuity processing that differ from ecological optics, and AR filters introduce synthetic shadows, inferred occlusion, and algorithmic continuity that become familiar through repetition. Touchscreens further establish gesture grammars that reshape expectations about how objects should respond to contact, movement, and attention across contexts. Over time, these patterned cues become intuitive, and the interpretive habits they support travel with users, influencing how scenes are read and how spatial relations are anticipated even in physical environments.

These adaptations do not replace biological perception; they layer synthetic cues onto ecological ones, producing hybrid expectations about what counts as coherent-enough space and how continuity should behave. In interactive spatial inference, that hybrid repertoire becomes actionable: prompts, reference media, selections, camera moves, edits, and verification checks externalize commitments about identity, layout, and causality. Users therefore stabilize outputs by placing constraints where inference is most underdetermined—identity, layout, and continuity—so a scene stays usable for the task at hand. In tools such as Marble, TouchDesigner, or Unreal, coherence becomes a co-produced property: models propose statistically plausible structure, while humans iteratively impose, test, and revise

the constraints that make that structure navigable, editable, or buildable. Translation is the skill that makes this possible: converting hybrid expectations into constraints that a tool can register and enforce.

6.4.3 Translation as an Ongoing Spatial Skill

Translation is a routine spatial skill: moving commitments across physical environments, conventional representations, and model-generated scenes. Each environment encodes its own constraints, so coherence in one representation does not guarantee coherence in another; a drawing must be reinterpreted to become a building, and a synthetic scene may require revision to remain coherent when treated as physical space (Gibson, 1979; O'Keefe & Nadel, 1978). Model-generated environments reflect statistical priors rather than ecological invariants, introducing additional structures that can reshape—or expand—how designers and viewers reason about spatial possibilities (Poole et al., 2022; Sajjadi et al., 2022; Qin et al., 2025).

Effective translation depends on recognizing the representational logic that produced the scene. Generative systems introduce a stance shaped by multimodal correlations, dataset composition, and architectural choices (Mildenhall et al., 2020; Bruce et al., 2024; Hu et al., 2025), and people learn to read that stance much as they learned to read perspective drawing, cinematic continuity, or game-engine space. This layering expands the range of spatial interpretations available without requiring representational unity; it requires local rules for what to trust, what to treat as provisional, and how to test coherence when moving between representations. Synthetic scenes may diverge from physical constraints yet serve as prompts for exploration, while physical environments ground and reshape assumptions carried over from generative outputs.

The result is a hybrid spatial environment in which multiple grammars coexist and interact. Historically, people have adapted through representational shifts (Clark, 1997); generative spatial systems extend that pattern by making statistical regularities a visible part of the representational landscape. Designers and viewers learn when to treat synthetic structure as exploratory prompt material and when to re-anchor it in ecological constraints, because downstream tasks (building, navigation, simulation, fabrication) re-impose different coherence demands. In interactive tools, translation is enacted through constraint channels—references, edits, camera moves, and checks—that keep scenes navigable, editable, and buildable even when the underlying spatial logic remains hybrid.

Sources

Bar-Tal, O., et al. (2024). Lumiere: A space-time diffusion model for video generation. ArXiv.org.https://arxiv.org/abs/2401.12945

Barron, J. T., et al. (2021). Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. ArXiv.org. https://arxiv.org/abs/2103.13415

Bruce, J., et al. (2024). Genie: Generative interactive environments. ArXiv.org. https://arxiv.org/abs/2402.15391

Clark, A. (1997). Being there: Putting brain, body, and world together again. MIT Press.

Durante, Z., et al. (2025). Agent AI: Surveying the horizons of multimodal interaction. ArXiv.org. https://arxiv.org/abs/2504.01512

Erdem, U., et al. (2019). Applications of spatial and temporal reasoning in cognitive robotics. Cognitive Robotics Lab.

Eslami, S. M. A., et al. (2018). Neural scene representation and rendering. Science, 360(6394), 1204–1210. https://doi.org/10.1126/science.aar6170

Gao, J., et al. (2021). Dynamic view synthesis from dynamic monocular video. ArXiv.org. https://arxiv.org/abs/2105.06468

Gibson, J. J. (1979). The ecological approach to visual perception. Houghton Mifflin.

Gibson, J. J. (1979). The theory of affordances. In The ecological approach to visual perception (pp. 127–143).

Ha, D., & Schmidhuber, J. (2018). World models. ArXiv.org. https://arxiv.org/abs/1803.10122

Hafner, D., et al. (2023). Mastering diverse domains through world models. ArXiv.org. https://arxiv.org/abs/2311.01460

Hu, S., et al. (2024). Simulating the real world: A unified survey of multimodal generative models. ArXiv.org. https://arxiv.org/abs/2405.14034

Hu, Y., et al. (2025). EWMBENCH: Evaluating scene, motion, and semantic quality in embodied world models. ArXiv.org. https://arxiv.org/abs/2505.09694

Huang, Z., et al. (2024). EnerVerse: Envisioning embodied future space for robotic manipulation. ArXiv.org. https://arxiv.org/abs/2407.08768
Jiang, L., et al. (2024). EnerVerse-AC: Envisioning embodied environments with action-conditioned world models. ArXiv.org. https://arxiv.org/abs/2407.08769
Liu, Y., et al. (2024). Aligning cyberspace with the physical world: A comprehensive survey on embodied AI. IEEE TPAMI.
Kerbl, B., et al. (2023). 3D Gaussian splatting for real-time radiance field rendering. ArXiv.org. https://arxiv.org/abs/2303.13495
Kumar, A., et al. (2019). Consistent generative query networks. ArXiv.org. https://arxiv.org/abs/1807.02033
Martin-Brualla, R., et al. (2021). NeRF in the Wild: Neural radiance fields for unconstrained photo collections. ArXiv.org. https://arxiv.org/abs/2008.02268
Mildenhall, B., et al. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. ArXiv.org. https://arxiv.org/abs/2003.08934
O'Keefe, J., et al. (1978). The hippocampus as a cognitive map. Oxford University Press.
OpenAI. (2024). Video generation models as world simulators. OpenAI. https://cdn.openai.com/papers/world-simulators.pdf
Park, K., et al. (2021). HyperNeRF: A higher-dimensional representation for topologically varying neural radiance fields. ArXiv.org. https://arxiv.org/abs/2103.16376
Poole, B., et al. (2022). DreamFusion: Text-to-3D using 2D diffusion. ArXiv.org. https://arxiv.org/abs/2209.14988
Qin, C., et al. (2024). WorldSimBench: Towards video generation models as world simulators. ArXiv.org. https://arxiv.org/abs/2403.12031
Rombach, R., et al. (2022). High-resolution image synthesis with latent diffusion models. ArXiv.org. https://arxiv.org/abs/2112.10752
Sajjadi, M. S. M., et al. (2022). OSRT: Object scene representation transformer. NeurIPS. https://osrt-paper.github.io
Sajjadi, M. S. M., et al. (2022). Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. CVPR. https://srt-paper.github.io
Spelke, E. S., et al. (2007). Core knowledge. Developmental Science. https://doi.org/10.1111/j.1467-7687.2007.00569.x
Sun, T., et al. (2024). Generative multimodal models are in-context learners. ArXiv.org. https://arxiv.org/abs/2402.10894
Tewari, A., et al. (2020). State of the art on neural rendering. ArXiv.org. https://arxiv.org/abs/2004.04776
Ullman, S. (1984). Visual routines. Cognition.
Wang, Z., et al. (2025). SITE: Towards spatial intelligence through evaluation. ArXiv.org. https://arxiv.org/abs/2503.09733
Wu, G., et al. (2024). 4D Gaussian splatting for real-time dynamic scene rendering. CVPR. https://guanjunwu.github.io/4dgs
Yang, M., et al. (2024). Cambrian-S: Towards spatial supersensing in video. ArXiv.org. https://arxiv.org/abs/2403.12843
Yang, M., et al. (2025). Thinking in space: How multimodal large language models see, remember, and recall spaces. ArXiv.org. https://arxiv.org/abs/2505.02520
Yu, A., et al. (2021). PlenOctrees for real-time rendering of neural radiance fields. ArXiv.org. https://arxiv.org/abs/2103.14024
Zhao, X., et al. (2024). Pseudo-generalized dynamic view synthesis from a video. ICLR. https://arxiv.org/abs/2310.08587
Zhen, X., et al. (2024). 3D-VLA: A 3D vision-language-action generative world model. ArXiv.org. https://arxiv.org/abs/2403.15952