Fluid Layouts and 2D Probes: Mitigating the Micro-Adjustment Paradox in Generative Spatial Interfaces

YANG Yi

School of Design, Hong Kong Polytechnic University, Hong Kong, CHINA

francoyeoh7@gmail.com

KEYWORDS:

Generative spatial user interfaces; cross-dimensional interaction; human-machine control; spatial abstract syntax trees; constraint-driven spatial solvers; the fine-tuning paradox; spatial computing.

ABSTRACT:

With the explosive evolution of large language models (LLMs), computational interaction is shifting from graphical user interfaces (GUIs) to intent-driven generative spatial user interfaces (Gen-SUIs). However, applying LLMs to mixed reality triggers a severe "cross-dimensional compilation crisis": 1D linear tokens struggle to map stably into 3D layouts that respect human cognitive and physical constraints. Consequently, users face extreme operational friction when fine-tuning complex generative business topologies—a phenomenon we define as the "Micro-Adjustment Paradox." To address this, we propose Gestalt, a foundational multi-modal compilation architecture designed for Gen-SUI. Gestalt introduces three core mechanisms. First, the Spatial-AST middleware transforms non-deterministic LLM semantics into rigorous 3D topological constraints, breaking the crude mapping of direct 3D coordinates. Second, a spatial solver utilizing multi-objective optimization and ergonomic penalty terms enables adaptive, fluid-like physical layouts of complex networks. Finally, the "Dimensional Probes" interaction metaphor allows users to extract specific 3D node clusters and instantly reduce them into a precise 2D workspace for parameter intervention. A system-level within-subject study tracking expert users' kinematic trajectories revealed that Gestalt reduces spatial friction by 81.2% and task completion time by 61.7% during large-scale situational awareness and high-precision fine-tuning tasks, exponentially increasing the sense of control. Ultimately, this paper establishes a new theoretical framework and rendering benchmark for non-deterministic spatial operating systems in the AI era.

1. INTRODUCTION

The history of human-computer interaction is, at its core, a history of how humans have reduced the mental effort required to communicate with machines. Today, computer science stands at the intersection of two revolutionary waves: first, intent-driven programming and natural language interfaces enabled by large language models; and second, spatial computing enabled by mixed reality devices. The convergence of these two trends has given rise to the highly ambitious Generative Spatial User Interface (Gen-SUI). In the Gen-SUI vision, users do not need to write code or manually drag and drop components to build software; instead, by simply expressing high-level semantic intent through natural language, artificial intelligence can generate and render concrete, complex system topologies and control logic in the user's physical space in real time. However, when this generative capability is applied to high-level productivity tasks—such as building B2B backends with multi-tiered permissions or troubleshooting dynamic business networks based on geographic locations—existing underlying operating system architectures (such as visionOS and Horizon OS) reveal critical limitations. The UI rendering protocols of these systems rely heavily on rule-based, deterministic 2D anchor

mechanisms. They are unable to handle the business logic trees—comprising dozens of interdependent and dynamically evolving elements—generated in real time by large language models. Forcing LLMs to directly output three-dimensional absolute coordinates not only easily triggers spatial hallucinations in the model but also leads to severe interface clipping, visual occlusion, and logical confusion resulting from elements falling outside the user's ergonomic reach. A deeper, system-level catastrophe lies in the complete collapse of human control. In a purely three-dimensional space, humans must rely on 6-DoF (six degrees of freedom) gestures or eye tracking for direct manipulation. Once AI rapidly generates a vast, macro-level business network, errors in local, granular parameters are inevitable due to the nature of its probabilistic models. When users attempt to precisely fine-tune these non-deterministic outcomes—for example, accurately pinching and adjusting the probability slider of a specific edge node among hundreds of floating data pipelines—the excessive degrees of freedom in three-dimensional manipulation actually become a massive source of friction. Users are forced to expend vast amounts of cognitive bandwidth on "understanding the chaotic AI topology" and "clumsily aligning hand movements in three-dimensional space." This paper formally defines this interactive deadlock—arising from the evolution from one-dimensional language to three-dimensional space—as the "fine-tuning paradox" in generative systems.

To break this deadlock, this paper argues that spatial computing systems should not be crude imitations of the physical world, but rather flexible extensions of human cognitive dimensions. We propose the Gestalt system architecture. Its core architectural insight is that the depth vision of three-dimensional physical topology is best suited for humans to achieve large-scale, parallel situational awareness, while flat two-dimensional symbolic systems (2D Canvas) are best suited for deterministic, low-degree-of-freedom, and absolutely precise calibration. Based on this, this paper makes the following three core contributions: Innovations in the theoretical and protocol layers: We propose the Spatial-AST compilation protocol, which breaks the direct mapping from natural language to 3D physical space. It reconstructs LLM outputs into a purely spatial logical topological syntax tree, eliminating semantic ambiguities arising from cross-dimensional generation. Innovations in the Physical Layout Engine Layer: We designed and implemented a constraint-driven spatial layout solver. By formulating a total energy equation that incorporates "semantic topological attraction," "ergonomic reachability penalties," and "environmental collision elimination," this solver enables dynamic, adaptive fluid layout of generated interfaces. Innovation in Cross-Dimensional Interaction Paradigms: We invented the "Dimensional Probes" interaction metaphor. This provides a seamless mechanism for transferring control across dimensions, enabling users to freely navigate between 3D macroscopic and 2D microscopic views, thereby completely resolving the challenge of precision loss in spatial operations.

2. RELATED WORK AND THEORETICAL FRAMEWORK

This study is situated at the intersection of generative artificial intelligence, spatial computing, and physical engines for human-computer interaction. To establish the theoretical foundation of the Gestalt architecture, we first examine the system-level shortcomings of existing Gen-SUI paradigms and formally propose a mathematical model of the "fine-tuning paradox" and an interaction philosophy of "cross-dimensional focus separation." 2.1 The Fallacy of Direct Spatial Mapping With the maturation of large language models in code generation and logical reasoning, Generative User Interfaces (GenUI) have become a hot topic. However, the vast majority of current GenUI research is limited to the dynamic assembly of two-dimensional DOM (Document Object Model) trees and the adaptive layout of web components. When academia and industzry attempt to extend this generative capability into three-dimensional space, they often fall into the fallacy of "direct spatial mapping." The core rendering protocols of industry-level operating systems (such as visionOS and Horizon OS) remain window managers based on planar anchors. Developers are forced to use deterministic coordinate systems (i.e., X, Y, Z, and quaternions) to position components. Recent hybrid initiatives have attempted to have LLMs directly output JSON data with 3D coordinates to arrange spatial environments. However, LLMs are fundamentally probabilistic prediction models based on one-dimensional token sequences, lacking embodied cognition of the real physical world. Forcing LLMs to act as "spatial

layout engines" inevitably leads to severe "spatial illusions"—such as UI nodes clipping through each other, violating ergonomic reach zones, or generating critical business data in the user's blind spots. This practice of tightly coupling logical semantics with physical coordinates is the root cause of the current fragility of Gen-SUI. 2.2 The Fine-Tuning Paradox and the Kinematic Friction Model In generative workflows, due to the probabilistic nature of AI outputs, the system state is inherently "non-deterministic." Humans inevitably need to perform secondary corrections and precise fine-tuning on the initial topology generated by the AI. However, three-dimensional space introduces catastrophic resistance to this fine-tuning. In traditional GUIs, the degrees of freedom of input devices (2-DoF for a mouse) align with the dimensions of the interface (2D). In spatial computing, however, users must employ 6-DoF gestures or tracked controllers. We propose a formal model that defines the "interactive editing friction" experienced by users when correcting AI errors in generative spaces as:

$$F_{edit} = \int_{t_0}^{t_1} \left(\lambda \cdot A_{intent}(t) \times D_{spatial}\right) dt$$

Here, $A_{(intent)}$ represents the system's intention ambiguity at time t, and $D_{(spatial)}$ represents the degrees of freedom of the current interaction environment (in pure VR, $D_{(spatial)} = 6$). λ is the individual's cognitive load coefficient. According to this equation, when a user attempts to adjust a high-dimensional parameter slider—generated by AI and with ambiguous states—in a 6-DoF space, the kinematic friction will increase exponentially. This leads to a severe "Gorilla Arm Effect" and overload of pre-attentive vision. The user will lose control of the system; we can approximately define control as the inverse of the friction force:

$$Agency = \frac{\Phi(Context)}{F_{edit} + \epsilon}$$

Here, $\Phi(Context)$ represents the user's awareness of the overall system situation. This is precisely the core of the "fine-tuning paradox" proposed in this paper: three-dimensional space provides unparalleled situational awareness (high Φ), yet simultaneously imposes friction that undermines execution precision (high F_{edit}), ultimately rendering the system unusable.

2.3 Dimensional Separation of Concerns

To maximize system control $Agency$, we must break the rigid constraints of $D_{(spatial)}$. Inspired by the "Separation of Concerns" principle in software engineering, the Gestalt architecture proposes the theoretical hypothesis of "Cross-Dimensional Reduction" at the physical interaction level.

This hypothesis posits that the optimal cognitive use of three-dimensional space (3D Space) is to render directed acyclic graphs (DAGs) with complex physical topologies and business dependencies, thereby establishing a macro-level "system situational awareness"; whereas the optimal cognitive use of a two-dimensional plane (2D Canvas) is to process structured text, forms, and deterministic parameter editing.

Therefore, rather than pointlessly optimizing ray tracing or gesture anti-shake algorithms in a pure 3D environment, it is preferable to design a compilation protocol: one that allows the system to unfold logical structures within a 3D topology, while simultaneously compressing target objects to a state of absolute precision ($D_{(spatial)}$ =2) via "Dimensional Probes" the moment the user requires localized, precise intervention. This cross-dimensional paradigm shift fundamentally circumvents the physical limits of micro-operations in three-dimensional space, establishing theoretical legitimacy for high-level system supervision in the era of strong artificial intelligence.

3. SYSTEM DESIGN AND IMPLEMENTATION

Gestalt is not merely a simple virtual reality front-end UI library, but rather a low-level "Interaction Compiler" tailored for non-deterministic spatial intent. Built on the Unity 3D engine and the C# asynchronous task system, the system completely abandons traditional front-end DOM tree rendering logic, forming a complete closed-loop system of "multimodal perception → semantic compilation → physics simulation → cross-dimensional feedback."

3.1 Overview of the Gestalt Pipeline

The UI rendering pipeline in traditional operating systems is deterministic: developers write layout code, and the engine renders according to specified coordinates. In the

Gestalt architecture, however, the generation of the interface is entirely unknown at runtime. To manage this uncertainty, our underlying pipeline is divided into three decoupled stages:

Intent Alignment Layer: Captures users' colloquial, ambiguous commands and subconscious physical movements, encapsulating them into high-density multimodal intent objects.

Spatial-AST Compilation Layer: Forces the LLM to decouple business intents from physical coordinates, outputting a pure logical topological tree.

Constraint-Driven Rendering Layer: Receives the topological tree and calculates absolute coordinates using a multi-object physics solver.

3.2 Multimodal Intent Fusion Tensor

In natural language-driven spatial computing, the primary challenge facing large language models is "semantic grounding." When a user says in a 3D space, "Lower the priority of this data stream and move it over there," a text-only LLM will completely fail because it cannot parse the pronouns ("this part," "over there").

To solve this spatial pronoun resolution problem, we built a Multimodal Intent Fusion Tensor at the input end of Gestalt. The system no longer listens to the audio stream in isolation but maintains a circular buffer of high-frequency time series data, including eye-tracking gaze, hand pose matrices, and spatial speech-to-text data.

We propose a "Time-window Decay Algorithm." When the natural language processing module (NLP Tokenizer) identifies demonstrative pronouns at time step t_{token}, the system extracts a time window of width $=\Delta t$ 200 ms. The fusion engine extracts the average gaze ray vector V_{gaze} within this window and the collision intersection coordinates $P_{(intersect)}$ with the target bounding box to construct the fusion tensor $\mathbf{T_{intent}}$.

$$\mathbf{T_{intent}} = \left[Token_{text}, \int_{t_{token}-\Delta t}^{t_{token}+\Delta t} \left(\omega \cdot V_{gaze}(t) \times \mathbf{M_{hand}}(t) \right) dt \right]$$

This tensor mathematically stitches together vague voice commands with extremely precise spatial gaze points at the lower level, and then sends them to the backend parser as a "Rich-Intent Object."

3.3 The Spatial-AST Parser

The biggest flaw in the system design of traditional GenSUI lies in attempting to have the LLM directly serve as a "spatial layout engine" (for example, requiring the large model to output JSON data containing x: 1.5, y: 0.8, z: -2.0). Large models lack the ability to perform real-world collision detection, and the coordinates they output can easily lead to logical breakdowns in the physical dimension.

The key strength of the Gestalt architecture lies in the fact that, through strict schema validation in the underlying network communication protocol, we have completely blocked the LLM's ability to output absolute physical coordinates.

We strictly limit the backend large model (such as GPT-4o) to functioning as a "pure logic compiler." After receiving a multimodal intent tensor, the large model must output a custom intermediate data structure we define—the Spatial Abstract Syntax Tree (Spatial-AST).

Spatial-AST is essentially a weighted directed acyclic graph (DAG) used to express complex business logic. Within the Spatial-AST, each UI component or data monitoring panel contains no coordinate information whatsoever; instead, they are abstracted into vector structures comprising the following properties:

$$\mathbf{V_{node}} = \langle Id, Semantics, Hierarchy, \mathbf{W_{edges}}, Priority$$

Semantics (Business Semantics): Defines whether the

node is a front-end input field, a global data dashboard, or an edge alert node.

Hierarchy (Nested Hierarchy): Determines the parent-child tree structure between nodes. $\mathbf{W_{edges}}$ (Topological Weight Matrix): Records the strength of business relationships between this node and other nodes. For example, in an LBS marketing network, the edge weight between the fund pool node and the coupon issuance node is 1.0 (strong relationship), while the edge weight with the user avatar node is 0.1 (weak relationship).

Priority (Interaction Priority): Indicates whether the node requires frequent manual intervention by the user in the current business context (e.g., adjusting the anti-cheating probability).

Asynchronous Streaming Topology: To resolve the clock domain conflict between LLM inference latency and the 90Hz refresh rate of the spatial physics engine, this parser does not employ traditional blocking wait mechanisms. We designed Spatial-AST to support a streaming chunk parsing architecture. As tokens from the large model are streamed down one by one, the middleware dynamically instantiates isolated business nodes and feeds them into the solver as "free-floating particles." When the $\mathbf{W_{edges}}$ weight tokens representing connectivity arrive, semantic springs are instantly generated between nodes, triggering self-layout. This design completely masks the AI's network latency, transforming the traditional "wait-and-load" process into a highly visually engaging "Spatial Crystallization" visual feedback.

3.4 Constraint-Driven Spatial Solver

After obtaining the Spatial-AST generated by large language models, the most significant challenge the system faces is how to transform abstract semantic relationships into three-dimensional absolute coordinates that align with real-world physical environments and human physiological limits. Existing systems (such as the default UI layout engine in Unity) typically rely on predefined grid systems; once the number of generated business nodes becomes massive and their relationships become highly complex, the system inevitably descends into spatial chaos.

The Gestalt architecture abandons static coordinate systems and pioneers a Spatial Solver based on Multi-Objective Optimization and the Spring-Damper Physics Metaphor.

The core mechanism of the solver is to transform the UI rendering task into a process of finding the minimum of the "Total Spatial Energy (E_{total})". Every frame (Δt = 11ms, to maintain a 90Hz refresh rate), the solver engine executes a gradient descent algorithm in a compute shader, driving all UI modules to "flow" toward optimal spatial coordinates. The total energy function is composed of four weighted core physical constraints:

$$E_{total} = w_1 E_{semantics} + w_2 E_{ergonomics} + w_3 E_{collision} + w_4 E_{gaze\_strain}$$

3.4.1 Semantic Gravity and Coulomb Repulsion

To ensure that the business logic topology is perfectly mapped to the visual distribution, we treat each UI node in the Spatial-AST as a charged particle with mass.

For nodes i and j that exhibit high connectivity (i.e., strong business dependencies) in the AST topology weight matrix, $\mathbf{W_{edges}}$ the engine applies a semantic gravitational force $\mathbf{F_s}$ that conforms to Hooke's Law:

$$\mathbf{F_s} = -k_s \cdot \mathbf{W_{ij}} \cdot \left(|x_i - x_j| - L_{ideal}\right) \frac{x_i - x_j}{|x_i - x_j|}$$

Here, L_{ideal} represents the optimal preset distance for human vision without overlap. Additionally, to prevent clipping and crowding between nodes with no direct connection, the engine applies Coulomb repulsion to global node pairs. However, traditional $O(N^2)$ repulsion calculations result in severe performance bottlenecks. Therefore, we have introduced the Barnes-Hut octree approximation algorithm (Octree Approximation) from astrophysics into the rendering pipeline. The system dynamically constructs a spatial hash tree; for clusters of distant UI nodes where the distance exceeds the threshold θ, the engine reduces them to a single center of mass for repulsion calculations. This algorithmic breakthrough successfully reduces the time complexity of topological layout to $O(N \log N)$, ensuring real-time performance for ultra-large-scale generative workflows on mobile computing platforms.

3.4.2 Ergonomic Golden-Zone Constraint

A major flaw in traditional VR systems is the "Gorilla Arm Effect." When AI generates interactive panels in unpredictable locations, users often have to stretch their arms to the extreme to reach the sliders.

The Gestalt solver reads the user's headset matrix and controller skeleton data in real time, setting the head position as the coordinate origin $P_{head}(0,0,0)$. Based on Kinematic Ergonomics, we define a circular hemispherical space $[R_{min}, R_{max}]$(typically set to 0.3m to 0.6m) around the user's chest as the "Golden Reach Zone."

For core monitoring or fine-tuning points in Spatial-AST that are marked as "High" priority, once they are pushed out of the golden reach zone by other forces $(d > R_{max})$ the engine will apply a strong nonlinear elastic penalty force $\mathbf{F_{ergo}}$.

$$\mathbf{F_{ergo}} = \lambda \cdot (d - R_{max})^2 \cdot \hat{\mathbf{u}}$$

This mechanism ensures that, no matter how vast or complex the background topology generated by the AI may be, the core components requiring precise user intervention will smoothly and automatically "slide" into the user's most comfortable range of motion, as if captured by a gravitational singularity.

3.4.3 Collision & Gaze Strain Prevention

The unpredictability of generative interfaces can easily cause the UI to be embedded in real walls (physical conflict) or to fall outside the user's effective field of view (physiological conflict).

- Traditional mesh intersection volume calculations cause significant performance bottlenecks on mobile XR devices. Therefore, the Gestalt engine extracts raw LiDAR data, voxelizes it, and constructs Signed Distance Fields (SDFs). For each UI node's center of mass $\overrightarrow{x_i}$, the engine queries its distance scalar $D_{sdf}(\overrightarrow{x_i})$ in the SDF in real time. When a node approaches the boundary of a physical obstacle $(\nabla E_{collision} \to \infty)$, the environmental repulsive field experiences an exponential surge in its gradient $(D_{sdf} < \epsilon)$, enabling extremely smooth zero-clipping physical collision avoidance with an extremely low query complexity of $O(1)$.
- Gaze Fatigue Prevention: Based on the comfortable rotation limits of the human cervical spine, the system defines an optimal frustum with a horizontal range of $\pm 30°$ and a vertical range of $\pm 20°$. When critical business nodes are generated outside this frustum,E_{gaze_strain} generates a centripetal force, significantly reducing the neck strain associated with visual searching in 3D space.

Through this series of extremely rigorous mathematical and physical constraint matrices, Gestalt-Solver completely eliminates the development costs associated with manual 3D interface layout for developers. The system functions like a gravitational field of absolute order; no matter how complex or chaotic the intentions generated by large language models may be, it can collapse them within milliseconds into a spatial architecture that aligns perfectly with human physiological intuition and business logic

4. INTERACTION PARADIGM: DIMENSIONAL PROBES

The Spatial Solver described earlier addresses the systemic challenge of "macro-level layout and situational awareness" in generative interfaces. However, when users need to make precise adjustments to the logical topology generated by AI, purely three-dimensional direct manipulation inevitably triggers the "fine-tuning paradox." To address this, Gestalt proposes the core interaction paradigm of this system: Dimensional Probes. Dimensional Probes do not simply flatten 3D models; rather, they constitute a multimodal intervention mechanism that spans system dimensions and encompasses the entire lifecycle. This mechanism is divided into four consecutive execution phases:

4.1 Trigger & Spatial Recession

When a user identifies anomalous nodes within a macro-level 3D business network, they use eye-tracking to lock onto the target node group and execute a physical "pinch-and-pull" micro-gesture. The moment this operation is triggered, the Gestalt system activates the Spatial Recession mechanism. At this point, the vast global 3D topological network does not disappear (to maintain the user's awareness of the global business context), but rather

recedes as a whole to create depth of field, with the global material's emission intensity reduced by 40%. Simultaneously, the solver applies a high physical damping coefficient (Viscous Drag) to unselected nodes, freezing their high-frequency oscillations to ensure absolute stability in the fine-tuning environment.

4.2 Dynamic Flattening

The selected group of 3D nodes is detached from the force-driven physical network, and the complex abstract parameters contained within them (recorded in the Spatial-AST) are instantly "decompiled." The system generates a high-contrast 2D interactive workbench (2D Workbench). Under the hood, we have built a Native Window Bridge that allows standard web-based or native system form components (such as precision numeric sliders, Boolean toggles, and drop-down menus) to run directly on this 2D canvas, restoring the pixel-level input precision of traditional SaaS systems.

4.3 Lazy-Follow Elastic Anchoring

In early spatial computing prototypes, if a 2D configuration panel was rigidly locked in a head-up display (HUD) mode directly in front of the camera, the panel's rigid tracking would trigger severe vestibular mismatch and motion sickness (VR Sickness). To address this issue, Gestalt's probe mechanism introduced the "Lazy-Follow Elastic Anchoring" algorithm. The downsampled 2D panel is rendered in world space, with the panel's center of mass $\overrightarrow{P_{panel}}$ connected to the target center of the user's visual cone $\overrightarrow{P_{target}}$ via an invisible damped spring. Its force model is defined as:

$$\mathbf{F_{anchor}} = -k_{follow}\left(\overrightarrow{P_{panel}} - \overrightarrow{P_{target}}\right) - c_{damp} \cdot \overrightarrow{v_{panel}}$$

Here, k_{follow} is the low-stiffness elastic coefficient, and c_{damp} is the critical damping coefficient. Interaction effect: When the user turns their head to view the receding 3D business network in the background, the 2D workbench in front of them slowly drifts to follow with an extremely smooth, physics-based inertial delay (approximately 150ms). This not only completely eliminates motion sickness but also greatly enhances the workbench's sense of spatial presence and depth cues through this "parallax delay."

4.4 Live Parameter Streaming & Bi-directional Closure

The probe mechanism is not a discrete form submission, but rather a hot-reloading pipeline that maintains two-way data binding. When a user continuously drags a parameter slider on the 2D workbench (such as adjusting the token release rate in real time), the underlying Spatial-AST streams updates at a frequency of 60Hz, even if the probe has not yet been released. These minute, high-frequency parameter fluctuations are injected into the physics engine in real time, allowing users to directly observe synchronous morphing of the receding 3D funding pipelines in the background as they drag the sliders with their fingers. When the adjustment meets expectations and the probe is released, the 2D panel folds back, completing the final solidification of the topological state.

5. APPLICATION WALKTHROUGH: LBS NODE MANAGEMENT

To validate the industrial-grade feasibility of the Gestalt architecture in addressing enterprise-scale complex computing, we built a sandbox scenario featuring a complete business workflow: a location-based (LBS) urban marketing and treasure hunt network scheduling backend. This scenario demonstrates how human architects can regain control when AI generates high-density, non-deterministic business operations.

Step 1: Macro Generation & Topology Perception

- The user puts on a headset and enters the following intent command: "Generate a grid-based treasure hunt marketing route covering the core commercial district, highlighting high-budget flow nodes." Gestalt's multimodal engine parses the command, and the LLM outputs a Spatial-AST containing over 50 commercial district nodes. Instantly, the spatial solver arranges an intricate three-dimensional pipeline network around the user. The thickness of the pipelines represents the speed of capital flow, while the height of the nodes

represents permission levels. Thanks to the Barnes-Hut algorithm and the golden reachable region constraint $(E_{ergonomics})$, the system not only avoids any model penetration but also ensures that the three most critical capital distribution hubs are perfectly suspended within the user's optimal operational radius. With a single glance, the user gains extensive parallel spatial situational awareness.

Step 2: Target Identification & Probe Extraction Through situational awareness,

- The user detects a core conversion zone node at the network's edge glowing with a red warning aura. This is caused by the AI's probabilistically generated high-value token release probability being set too high; without intervention, the budget will be maliciously depleted within ten minutes. The user locks their gaze on the node and performs a "pinch-and-pull" motion with their right hand. Instantly, the vast 3D urban topological network dims and recedes. The red anomaly node is isolated, floating before the user's chest and dynamically reduced to a crystal-clear two-dimensional configuration form.

Step 3: Micro-Adjustment in 2D

- On this 2D workspace, the frustrating friction of depth alignment in three-dimensional space is completely eliminated. With the precision of traditional GUI interaction, the user smoothly drags the slider to lower the "GPS anti-cheat drift tolerance" from 15m to 5m and precisely enters the token conversion threshold as 0.45. When turning to view parameters of other nodes in the background, the Lazy-Follow mechanism keeps the 2D panel in perfect sync, with no sense of motion sickness.

Step 4: Real-time Re-compilation

- After confirming the modifications, the user releases the probe. The 2D panel collapses and re-embeds into the 3D network. The underlying Spatial-AST is updated, triggering a global recalculation by the physics engine. The user can visually observe that the high-risk 3D pipeline representing capital outflow instantly contracts and narrows, the red warning is lifted, and the surrounding sub-task flows dependent on this node automatically adjust their spatial distances. The crisis is completely resolved through an elegant cross-dimensional dimensionality reduction maneuver.

6. USER EVALUATION

To rigorously quantify the actual effectiveness of cross-dimensional compilation architectures in enhancing user agency, we conducted a system-level within-subjects study.

6.1 Participants and Task Design

We recruited 18 participants (mean age M=28.5, standard deviation SD=3.2), all of whom had at least three years of experience in enterprise-level SaaS system architecture design or advanced spatial interaction (XR UX) design.

Experimental Sandbox: We replicated the urban LBS topological network described in Chapter 5. The system initially generated a 3D business network comprising 40 interconnected nodes using a large language model.

Core Tasks: Task Isolation and Variable Control: To prevent time savings from automatic layout algorithms from interfering with the evaluation of interaction paradigms, this experiment strictly isolated the "generation phase" from the "editing phase." Before the timer started, the initial 3D networks for both control groups had already been laid out using an optimization solver and presented to the users. The start of the timer was set to the moment the user began searching for and executing the "Micro-Adjustment Task." Participants had to locate three anomalous flow control nodes within the floating network and modify their floating-point values with extreme precision. This design uniquely examines and isolates the differences in execution accuracy between cross-dimensional interactive interventions and pure 3D direct manipulation.

6.2 Experimental Conditions

- Baseline: Represents the optimal solution currently

found in systems from major manufacturers (e.g., visionOS/Horizon OS). Users must use a 6-DoF trackpad or their hands to directly pinch and drag 3D parameter sliders suspended in mid-air within a 3D space.

- Gestalt: Enables the constraint-driven layout engine (Solver) and dimensional probes. Users extract nodes into 2D panels via "pinch-and-drag" to modify them

.

6.3 Quantitative Results: Kinematic Friction and Efficiency

This study broke away from the tradition of relying solely on subjective questionnaires by directly extracting the spatial kinematic trajectories of the subject's right hand from the underlying hardware to calculate the absolute friction of the operation.

- Invalid Translation Distance: In the Baseline group, due to the excessive degrees of freedom of the 3D slider, the subject's hand was highly prone to minor depth (Z-axis) deviations, causing parameters to frequently drift away from target values. To complete the fine-tuning of the three nodes, the subject's right hand performed an average of 4.84 meters (SD=1.21) of unnecessary probing and alignment movements in mid-air. In contrast, in the Gestalt group, because the dimension-reduction probe instantly collapsed the manipulation into absolute coordinates on a two-dimensional plane, the total hand movement trajectory plummeted to 0.91 meters (SD=0.18) Kinematic friction was remarkably reduced by 81.2%.

- Task Completion Time (TCT): Benefiting from the elimination of kinematic friction, the average TCT for the Gestalt group was only 34.2 seconds (SD=4.1), whereas the Baseline group's TCT was as high as 89.5 seconds (SD=12.3). A paired t-test revealed a highly significant difference ($p < 0.001$).

6.4 Qualitative Results: Cognitive Load and Agency

After completing the task, participants filled out the NASA-TLX cognitive load scale and our custom System Agency Scale (1–7 Likert scale).

- Cognitive Load Plummets: The Gestalt system reduced "Mental Demand" by 62% compared to baseline and saw a dramatic 75% drop in "Frustration." Several participants remarked: "Adjusting parameters in pure 3D is like trying to pick up tofu with long chopsticks, but the probe mechanism makes me feel like I've got a scalpel back in my hand."

- Restored Control: Users' sense of control over system-generated results soared from a baseline score of 3.1 to 6.4. This demonstrates that granting users the privilege to reduce dimensionality at will is the shortest path to building human trust in strong AI systems.

7. DISCUSSION

The rise of large language models once fostered a technological utopian fantasy within the industry: the belief that natural language (LUI) would eventually consume everything, and that the ultimate goal of computing was a completely invisible AI agent. This study powerfully refutes this view.

When dealing with enterprise-level computing and workflows characterized by highly complex dependencies, there exists an irreconcilable cognitive mismatch between the "one-dimensionality" of language and the "three-dimensionality" of spatial computation. Pure language leads to a state black box, while pure three-dimensionality leads to operational chaos.

The Gestalt architecture demonstrates that human-computer interaction in the era of strong AI is by no means about humans relinquishing control, but rather a redistribution of responsibilities at the system protocol layer. We delegate full authority for generating logical topologies to large models (Spatial-AST), entrust the management of physically demanding spatial layouts to real-time physics engines (Spatial Solver), and firmly retain the "Dimensional Probes"—the privilege of traversing between "high-dimensional situational awareness" and "low-dimensional panel control"—as the ultimate control surface, firmly in human hands.

8. Conclusion

This paper addresses the "fine-tuning paradox" arising

from Generative Spatial Computing (Gen-SUI) in complex system management by proposing a revolutionary Gestalt compilation architecture. By constructing the Spatial-AST protocol, a constraint-driven multi-objective spatial solver, and a flexible, time-delayed dimensional probe mechanism, we have successfully built a bridge that seamlessly switches between three-dimensional high-level overview and two-dimensional high-precision control.

Empirical data overwhelmingly demonstrates the exceptional efficacy of cross-dimensional intervention in eliminating spatial friction and restoring human control. The foundation of future spatial operating systems will no longer consist of stacked static 3D windows, but rather a fluid spatiotemporal engine capable of sensing human physiological limits and dynamically compiling AI business logic. Gestalt establishes a rigorous theoretical framework and engineering benchmark for this impending revolution in computational interaction.

Architect's Closing Remarks

REFERENCES

[1] Sangho Bae et al. 2024. Generative Spatial Interfaces: Bridging Natural Language and 3D Layouts via Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, New York, NY, USA.

[2] Josh Barnes and Piet Hut. 1986. A hierarchical O(N log N) force-calculation algorithm. *Nature* 324, 6096 (1986), 446–449.

[3] Doug A. Bowman, Ernst Kruijff, Joseph J. LaViola Jr., and Ivan Poupyrev. 2004. *3D User Interfaces: Theory and Practice*. Addison-Wesley Professional, Boston, MA.

[4] Juan David Hincapié-Ramos, Xiang Guo, Paymahn Moghadasian, and Pourang Irani. 2014. Consumed endurance: a metric to quantify arm fatigue of mid-air interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 1063–1072.

[5] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*, Vol. 52. North-Holland, 139–183.

[6] Jens Grubert, Lukas Witzani, Eyal Ofek, Michel Pahud, Matthias Kranz, and Per Ola Kristensson. 2018. Text entry in immersive head-mounted display-based virtual reality using standard keyboards. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 159–166.

[7] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[8] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and Andrew Fitzgibbon. 2011. KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology (UIST '11)*. ACM, New York, NY, USA, 559–568.

[9] Thomas M. J. Fruchterman and Edward M. Reingold. 1991. Graph drawing by force-directed placement. *Software: Practice and Experience* 21, 11 (1991), 1129–1164.

[10] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic interaction: coupling cognition and computation through usable spatial representations. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2779–2788.