Behavior-Aware Anthropometric Scene Generation for Human-Usable 3D Layouts

Semin Jin*
Design Informatics Lab
Hanyang University
Seoul, Republic of Korea
Human-Centered AI Design Institute
Hanyang University
Seoul, Republic of Korea
tpals97@gmail.com

Donghyuk Kim*
Design Informatics Lab
Hanyang University
Seoul, Republic of Korea
Human-Centered AI Design Institute
Hanyang University
Seoul, Republic of Korea
oververitas@gmail.com

Jeongmin Ryu
Design Informatics Lab
Hanyang University
Seoul, Republic of Korea
2002rjm@gmail.com

Kyung Hoon Hyun†
Design Informatics Lab
Hanyang University
Seoul, Republic of Korea
Human-Centered AI Design Institute
Hanyang University
Seoul, Republic of Korea
hoonhello@gmail.com

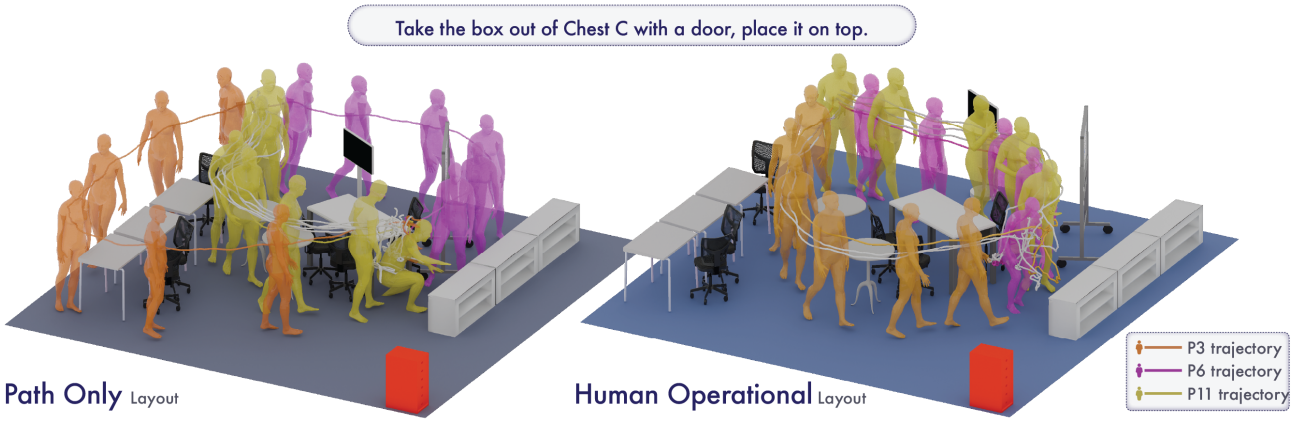

Figure 1: Visualization of Movement Trajectories in Generated Path-Only and Human-Operational Layouts.

Abstract

Well-designed indoor scenes should prioritize how people can act within a space rather than merely what objects to place. However, existing 3D scene generation methods emphasize visual and semantic plausibility, while insufficiently addressing whether people can comfortably walk, sit, or manipulate objects. To bridge this gap, we present a Behavior-Aware Anthropometric Scene Generation framework. Our approach leverages vision–language models (VLMs) to analyze object–behavior relationships, translating spatial requirements into parametric layout constraints adapted to user-specific anthropometric data. We conducted comparative studies with state-of-the-art models using geometric metrics and a user perception study (**N**=16). We further conducted in-depth human-scale studies (individuals, **N**=20; groups, **N**=18). The results showed improvements in task completion time, trajectory efficiency, and human-object manipulation space. This study contributes a framework that bridges VLM-based interaction reasoning with anthropometric constraints, validated through both technical metrics and real-scale human usability studies.

*Co-first authors.
†Corresponding author.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790341

CCS Concepts

• **Human-centered computing → HCI design and evaluation methods**; • **Computing methodologies → Computer vision**.

Keywords

Human-usable 3D layout, Indoor Scene Generation, Anthropometric Data, VLM.

ACM Reference Format:

Semin Jin, Donghyuk Kim, Jeongmin Ryu, and Kyung Hoon Hyun. 2026. Behavior-Aware Anthropometric Scene Generation for Human-Usable 3D Layouts. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3772318.3790341

1 Introduction

Physical environments fundamentally shape human movements and behavior [25]. In the real world, layouts are rarely static; users naturally adjust their surroundings—shifting a chair back to create legroom or pulling a table closer to reach an object—to fit their specific body dimensions and movements. A usable layout is then one already optimized for its occupants' anthropometrics and behaviors. Recent progress in 3D scene generation has enabled automatic creation of plausible layouts [4, 20, 29, 35, 42]. However, a critical limitation is that these models generate layouts based on dataset statistics [20, 29, 42] or large language model (LLM) priors [4, 35], lacking any explicit reasoning about human dimensions or behavioral patterns.

Consider an office scenario: a generative model might place desks back-to-back based on visual symmetry found in training dataset. However, because the model ignores the anthropometric clearance required to actually push a chair back and stand up, the resulting layout creates an immediate conflict zone—a functional failure that a human user would have instinctively avoided by adjusting the furniture distance. This problem becomes even more critical in extended reality (XR) applications requiring embodied interaction fidelity, such as training simulations, telepresence, or digital twins for ergonomic assessment. For example, users may repeatedly reposition themselves to reach targets outside their movement range, or experience visual clipping where virtual hands penetrate surfaces—issues arising because the layout was not optimized for user-specific body dimensions. Although established design standards address functional constraints such as drawer clearances and circulation widths, applying them to layout generation is challenging: the guidelines specify recommended ranges that cannot be reduced to fixed values without specific contexts, and generating a usable layout requires holistic consideration of object types, dimensions, room configurations, and user body characteristics. Our framework addresses these challenges by leveraging VLM reasoning to infer object-specific spatial constraints and grounding them in individualized anthropometric data.

To address this limitation, we propose **Behavior-Aware Anthropometric Scene Generation**, a vision language model (VLM)-based framework that augments text-based layout generation using *behavioral reasoning* and *anthropometric grounding*. We leverage VLMs to infer object functions from visual cues and reason about potential human interactions based on scene type and layout criteria. Our framework uses layout criteria and assets as inputs and proceeds in two stages. First, it constructs behavior-aware relational representations that integrate object semantics, human–object interaction patterns, and group-level spatial relations. Second, it performs constraint-based layout generation by inferring anthropometrically grounded spatial constraints from these relations and encoding them as differentiable penalty terms for gradient-based optimization. By parameterizing these constraints with personalized anthropometric profiles, our approach ensures that the resulting scenes support functional connectivity, adequate operational clearances, and efficient circulation paths tailored to individual users.

We evaluated our framework through technical validation and user studies on human-operational usability. First, we conducted performance comparisons with state-of-the-art LLM-based scene generation methods, LayoutVLM [35], to measure object collisions, floor-plan violations, and professionals' perceptions of generated layouts (N = 16). Second, we implemented the generated layouts at a 1:1 scale in physical office and lounge environments and compared three conditions: **LayoutVLM (Baseline), Passage-Only (PO)**, which ensures minimal navigable passages using static body dimensions; and **Human-Operational (HO)**, which guarantees sufficient human-operational space based on movement-related anthropometric envelopes. For PO and HO conditions, we instantiated participant-specific anthropometric profiles from each participant's Skinned Multi-Person Linear (SMPL) [22] model, parameterizing the constraints (e.g., passage widths, reaching envelopes, and viewing requirements) to each participant's actual body dimensions. Our user studies on human-operational usability consist of two parts: an individual study (N = 20) using structured *object–action–target* tasks [1, 44] to validate the anthropometrically grounded layout support task performance, and a group study (N = 18, six teams of three) using naturalistic collaborative scenarios to examine whether layouts minimize circulation conflicts during shared tasks [23, 24].

The key contributions of our research include:

- We propose a behavior-aware anthropometric scene generation framework that leverages VLM reasoning to infer spatial constraints from human-object interactions and grounds them in individual anthropometric data.
- We instantiated these constraints within a differentiable layout optimization process, operationalizing passage widths, operational clearances, and interaction-space occupancy to generate human-operational layouts.
- We conducted an evaluation combining technical validation, professional perception study, and user studies on human-operational usability in real-scale environments.

2 Related Works

Recent 3D scene generation methods synthesize plausible layouts but rely on generic spatial constraints that overlook individual body dimensions and behavioral patterns. Generating human-usable layouts requires anthropometric design principles, yet existing evaluations rarely assess whether generated layouts support actual human operation. To address these gaps, we review prior work in three areas: LLM-based scene generation, anthropometric design

in virtual and physical spaces, and evaluation methodologies for 3D layouts.

2.1 LLM-based Scene Generation

Recent advances in 3D indoor scene generation have leveraged Transformer architectures [29, 37, 42], diffusion models [20, 38, 48], and LLMs [4, 35, 47] to generate semantically plausible layouts. Among these approaches, LayoutVLM [35] represents the current state-of-the-art in LLM-based scene generation by formulating layout creation as a constraint-satisfaction problem. It leverages LLMs to translate natural-language instructions into spatial constraints (e.g., distance, orientation, and alignment) and then optimizes object positions through gradient-based methods using visual language model feedback. This approach demonstrates strong performance in generating semantically coherent layouts that satisfy user-specified relationships. However, LayoutVLM's reliance on LLM common sense for distance parameters results in generic, one-size-fits-all solutions that fail to consider individual body dimensions or behavioral requirements.

A separate line of research explores LLMs for human behavior simulation within existing scenarios [11, 30, 34, 45]. These approaches excel at generating human motions and planning action sequences in pre-defined spatial layouts but differ from our objective: they adapt human behavior to fit existing spaces, whereas we generate spaces to fit human-operational needs. Our work builds upon LayoutVLM's constraint optimization framework but extends it by integrating anthropometric data directly into the constraint quantification process. Rather than relying on generic distances from LLM common sense, we compute person-specific operational requirements based on individual body measurements and intended interactions–advancing scene generation from semantically plausible to human-operational.

2.2 Anthropometric-Driven Design in Virtual and Physical Scenes

Foundational theories in architecture and human factors establish that spatial design is not merely visual but fundamentally interaction-oriented. In architectural theory, the relationship between the human body and space is central, where operational space is defined by the dynamic range required for human activity [6, 8, 28]. Similarly, ergonomics literature distinguishes between structural anthropometry—static body dimensions—and functional anthropometry, which describes the dynamic range of motion and clearance required for tasks, emphasizing that true usability depends on accommodating the latter [41]. Aligning with these theoretical frameworks, research in ergonomics and XR demonstrates the critical importance of anthropometric considerations in the functional space. Research on personalized furniture design has shown that incorporating individual body measurements—such as reach envelope and joint range of motion—directly improves task performance [31]. XR environments have further embraced this approach, in which virtual objects dynamically adapt to user-body dimensions for optimal interaction [5, 18, 19]. These systems adjust shelf heights based on arm reach, scale workspaces to accommodate sitting eye height, and position controls within comfortable manipulation zones. However, these approaches primarily evaluate the design process itself, measuring whether the generated furniture dimensions match body measurements, rather than assessing actual human behavior in the resulting spaces. Consequently, the gap between anthropometric specifications and real-world use remains largely unknown, and the cascading effects on human-scene and human-object interactions have not been systematically studied in the context of scene generation.

The anthropometric design approach extends beyond individual furniture pieces to encompass operational spaces and the volume required for humans to use objects effectively. The established guidelines [10, 28] define the clearances for drawer operations, circulation paths around desks, and viewing distances. However, current scene generation methods treat these as fixed constants. Our human-operational approach bridges this gap by translating anthropometric measurements into spatial constraints that ensure adequate operational space for each individual's body dimensions. This shift from generic clearances to personalized operational volumes represents an advancement in making generated scenes truly usable, rather than merely plausible. The importance of this personalization becomes evident when considering human diversity; a 5th percentile female and a 95th percentile male require substantially different operational spaces [28]; however, current methods apply uniform constraints that may be insufficient for larger individuals or wastefully spacious for smaller ones. By integrating anthropometric data directly into the generation process, we ensure that layouts accommodate specific individuals who will inhabit these spaces.

2.3 3D Scene Evaluation Methods

Existing 3D scene generation research has primarily relied on visual plausibility metrics such as Fréchet Inception Distance and Kernel Inception Distance [20, 29, 38] to measure the distributional similarity to training data, or user ratings [4, 34, 35] to assess semantic coherence. Although these metrics effectively validate visual quality and learning performance, they fail to capture the behavioral and operational aspects that determine the actual usability. In contrast, architecture and interior design fields employ a comprehensive post-occupation evaluation that assesses cognitive comfort, wayfinding efficiency, spatial satisfaction, and long-term behavioral adaptation [15, 16, 27]; however, these require extended observation periods and are impractical for evaluating generated scenes at scale. Recent simulation-based approaches have attempted to bridge this gap by evaluating furniture mobility and walkability using embodied artificial intelligence (AI) agents [46]. However, these methods cannot capture human behavioral flexibility, as people dynamically adapt by stepping over obstacles, crouching to pass through tight spaces, or creating unexpected optimal paths. Motion generation research [43, 44] models human-scene interactions more realistically, but requires extensive training data and struggles with behavioral improvisation.

To address this evaluation gap between visual metrics and real-world usability, we developed behavior-grounded metrics specifically designed for the generated 3D layouts. These metrics measure how humans navigate and utilize space. Our approach captures observable behavioral patterns: trajectory variability reveals layout intuitiveness, action sequences expose compensatory movements,

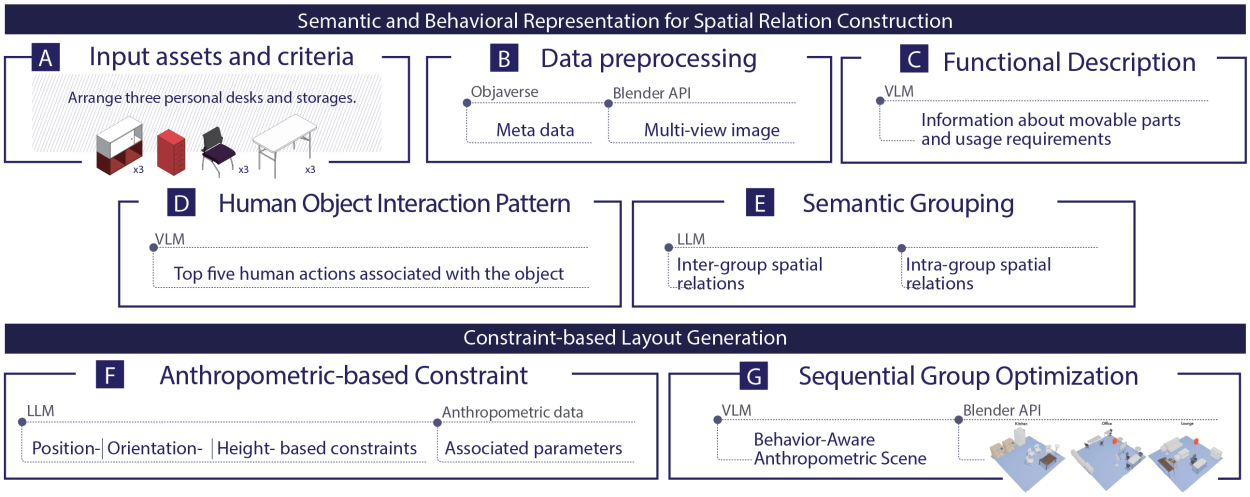

Figure 2: Overview of the Behavior-Aware Anthropometric Scene Generation. The framework proceeds in two phases: Semantic and Behavioral Representation (Stages A–E) constructs spatial relations, and Anthropometric Constraint-based Layout Generation (Stages F–G) optimizes the final layout using anthropometric constraints.

and occupancy ratios validate operational space adequacy. This evaluation framework provides empirical evidence that anthropometric-aware generation improves functional performance, establishing a new paradigm for assessing generated scenes based on human-operational criteria.

3 Behavior-Aware Anthropometric Scene Generation

We present **Behavior-Aware Anthropometric Scene Generation**, an approach that augments language-based layout generation using behavioral reasoning and anthropometric grounding. The goal is to generate spatial layouts that align with both furniture function and human-object interaction. Our approach enables the system to infer spatial constraints by explicitly referencing behavioral context and anthropometric data. As shown in Figure 2, our framework uses scene instructions and assets as inputs and proceeds through two main stages:

(1) **Semantic and Behavioral Representation for Spatial Relation Construction**: Constructing behavior-aware relational representations that integrate object semantics, human-object interaction patterns, and group-level spatial relations (Figure 2A–E).
(2) **Constraint-based Layout Generation**: Inferring anthropometrically grounded constraint representations suitable for differentiable spatial optimization (Figure 2F–G).

Complete prompt templates are provided in Appendix A for reproducibility.

3.1 Semantic and Behavioral Representation for Spatial Relation Construction

The **[A-E] stages** interpret raw 3D assets and layout criteria to produce behavior-aware relational representations that link object geometry, function, and human interaction. These representations establish the contextual foundation required for constraint inference by modeling how objects are used, accessed, and coordinated within a scene. The following sections describe how we organize the asset data, extract behavioral features, and semantically group objects to prepare for constraint inference.

3.1.1 Data and Input. Given user-specified layout criteria (e.g., room type, required furniture), we retrieved 3D assets from open-universe datasets and placed them into initial scene layouts: **[A] Input assets and criteria** (Figure 2A). For each input asset, we leverage OpenShape [21], a large-scale multimodal model that retrieves 3D assets by aligning point clouds with text and image descriptions, to extract candidate objects from the open-universe dataset Objaverse [9]. Once assets are determined, we extract each object's 3D bounding box, metadata (category, width, depth, height), and generate multi-view renderings of the 3D model using a scripted Blender pipeline: **[B] Data Preprocessing** (Figure 3B). Local coordinates were standardized (+X: right, +Y: forward, +Z: upward). We employed a VLM (GPT-4o) to interpret visual and textual cues jointly through prompts describing an object's geometry, parts, and possible interactions. To provide the VLM with visual grounding, we implemented an automated rendering pipeline using Blender. Each object was rendered from four orthogonal viewpoints (0°, 90°, 180°, and 270°) to capture comprehensive geometric details. The multi-view approach captures fine-grained features (e.g., casters, hinges, or door knobs) that dictate functionality but are often absent from text descriptions. We paired these multi-view images with structured metadata, enabling the VLM to cross-reference visual evidence with textual descriptions. The combined input enables the system to infer the functional properties of an object, such as movement axes, articulation points, and kinematic constraints, which are not evident in static category labels.

3.1.2 Human-Object Interaction-based Feature Extraction. Object geometry alone does not specify how furniture should be accessed

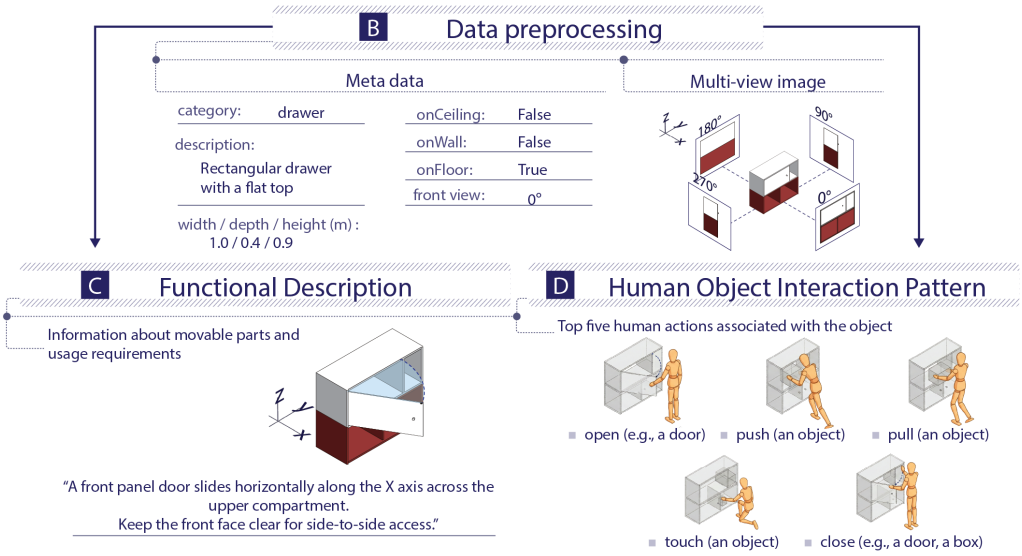

Figure 3: Semantic and Behavioral Representation (Stages B–D). For every 3D asset used in the scene, [B] preprocesses metadata and multi-view renderings, [C] infers a functional description, and [D] extracts a human–object interaction pattern.

or operated–a cabinet may open outward or slide laterally, and a chair may swivel or remain fixed. This stage infers each object's functional and behavioral properties: how it operates and how humans can interact with it. The VLM takes the paired multi-view image set and metadata as input and produces two complementary outputs: **[C] Functional Description** (Figure 3C), capturing qualitative information about movable parts and usage requirements, and **[D] Human-Object Interaction Pattern** (Figure 3D), identifying the top five human actions associated with the object based on atomic visual actions (e.g., sit, open, pull) [13]. The resulting vision-to-text process produces a structured JSON representation that links the objects to their inferred functions and interaction semantics. These representations provide a behavioral foundation for constraint inference (detailed prompts in Figure A1; output example in Figure A2).

3.1.3 Semantic Grouping. The **[E] Semantic Grouping** (Figure 4E) stage organizes objects into functional groups that reflect how humans interact with them in a scene. While individual objects can be interpreted in isolation, meaningful spatial reasoning emerges when they are considered as part of behavioral configurations; for example, chairs around a desk forming *a workspace* or sofas arranged around a coffee table creating *a lounge area.* Semantic grouping has two purposes. First, it reinterprets the atomic human-object interactions inferred in Stage [D] within a group context to extract spatial definitions. Specifically, the system identifies both **intra-group spatial relations** (internal arrangement within a functional unit) and **inter-group spatial relations** (connectivity between distinct groups), which are subsequently converted into a structured symbolic program in Stage [F]. For instance, an atomic 'open' or 'pull (an object)' action becomes the higher-level relation 'organize' when understood within a multi-cabinet storage group. Second, it establishes a structural abstraction that reduces the optimization complexity in Stage [G] by treating each group as an independent unit with a behavioral priority determined by functional significance and object scale (e.g., scene-defining elements such as beds or desks are placed first). We performed semantic asset grouping using the system in Appendix Figure A3, defining each group as a set of objects linked by functional relation, geometric proximity, and shared human action (output example in Figure A4).

3.2 Constraint-based Layout Generation

The **[F-G]** stages infer anthropometrically grounded constraint representations from natural-language spatial relations. They convert these relations, derived from the object geometry, function, and human-object interaction, into executable, differentiable constraints required for spatial optimization. Unlike prior layout approaches that rely on LLM common sense [12, 35, 47], our framework infers constraint specifications by explicitly referencing behavioral semantics and anthropometric rationale.

3.2.1 Spatial Constraint Definition. To define the spatial relationships between objects, we constructed an extended taxonomy adapted from geometric relation formulations used in prior layouts and scene generation methods [29, 35, 47]. We then reinterpreted the constraint semantics and parameterization in a behaviorally and anthropometrically grounded manner. Prior work defines constraints in geometric terms, focusing on the distances and angles between assets, independent of how people interact with them. By contrast, our formulation explicitly encodes how people need to reach, access, and operate furniture. The taxonomy organizes natural-language relations into learnable constraint types—positional, orientational, and height-based—such as *chair against wall, table aligned with sofa,* or *lamp on top of desk.* Table 1 lists each constraint type and the corresponding constraint names.

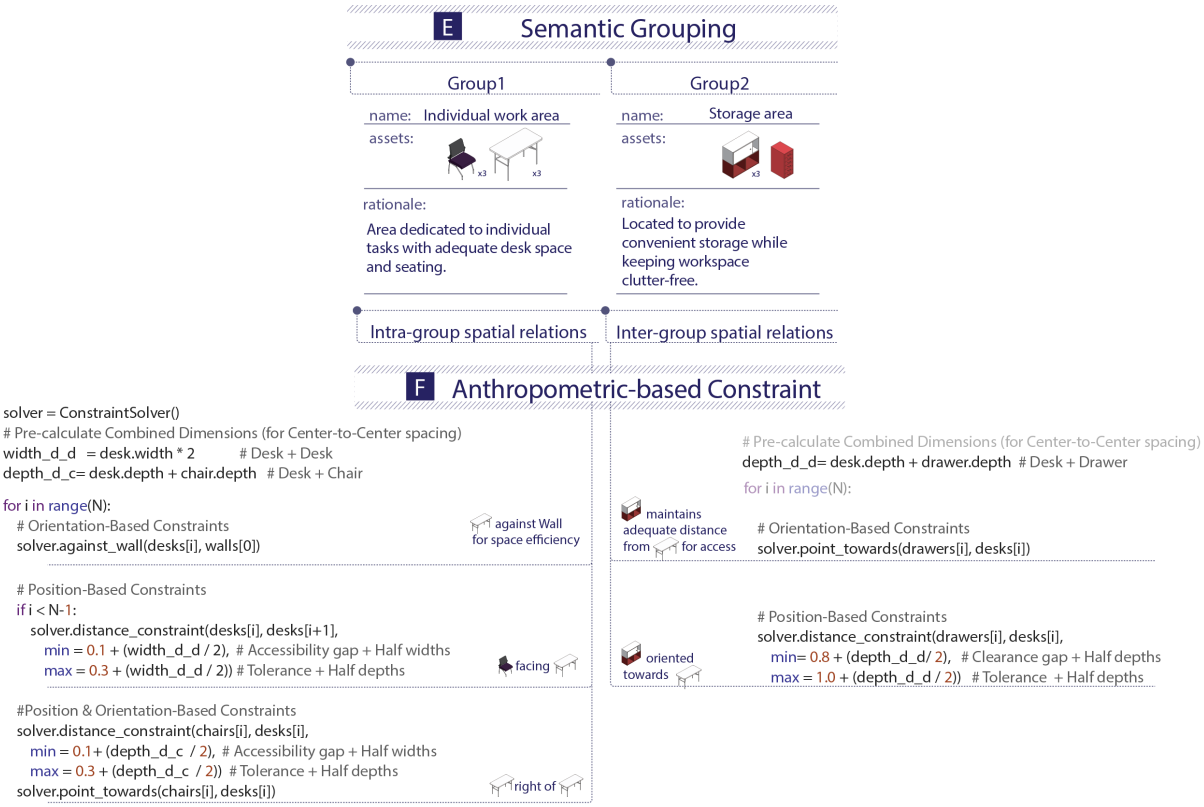

Figure 4: [E] Semantic Grouping and [F] Anthropometric-based Constraint Inference.

Among these, *distance constraint* requires additional clarification, because it directly governs human accessibility and functional clearance. We represent each distance constraint as a center-to-center distance range $[d_{\text{min}}, d_{\text{max}}]$, explicitly derived from the manner in which people reach, stand, and move around objects. Here, $d_{\text{accessibility}}$ represents the minimum distance required to reach or access an object, while $d_{\text{clearance}}$ represents the space needed for operational movements (e.g., pulling out a chair). We infer these bounds based on the interaction semantics: for accessibility-focused relations, $d_{\text{min}} = d_{\text{accessibility}}$ and $d_{\text{max}} = d_{\text{min}} + \tau$; for clearance-focused relations, $d_{\text{min}} = d_{\text{clearance}} - \tau$ and $d_{\text{max}} = d_{\text{clearance}}$, where τ is a tolerance buffer inferred by the VLM based on object function and interaction context.

3.2.2 Anthropometric-based Constraint Inference. In the **[F] Anthropometric based Constraint Inference** stage (Figure 4F), we infer a complete constraint specification, including the constraint type, its parameters, and the associated anthropometric rationale, from natural-language spatial relations. We first parse each relation (e.g., *Office Chair facing Desk* for intra-group relations, *Double Chest maintains adequate distance from Desk* for inter-group relations) and assign it to one of the predefined constraint types. Each type includes a parameter template that specifies the spatial quantities to be inferred, such as distances, angles, or vertical offsets. To infer these parameters, we reference standardized anthropometric datasets using *Human Dimension and Interior Space* [28], which aggregate data from multiple anthropometric sources [8, 26, 40]. These datasets serve as population-level references. We use the 5th–95th percentile ranges for key horizontal dimensions, including forward reach, lateral reach, body breadth, and body depth, to represent the natural variability across the population. This enables our inferred constraints to remain valid across diverse body sizes and movement capabilities, without assuming a single average individual.

We then mapped each constraint type to a specific anthropometric rationale: reach-related relations use arm-reach measures, clearance and circulation relations use body breadth or depth, adjacency relations use minimal offsets required for co-functioning objects, and visual/orientation constraints use eye position and preferred viewing directions to align object fronts toward the primary interaction or viewing area. By integrating the relation type, the relevant anthropometric ranges, and the interactions, we infer the constraint type and its associated parameters such as `distance constraints` $[d_{min}, d_{max}]$, `align_0°`, and `align_90°`. Finally, we encode this specification in an executable optimization schema that records the constraint type, object pair, parameter values, and anthropometric rationale. This structured representation is passed directly to the differentiable optimization module in Stage [G] (system prompt in Appendix Figure A5; output example in Figure A6).

Table 1: Spatial Constraint Taxonomy for Behavior-Aware Anthropometric Scene Generation, comparing our behavior-aware, anthropometrically grounded constraints with the geometric relations [35].

Constraint Type	Constraint Name	Method	Description
Position-based	$L_{\text{distance}}(p_i, p_j, d_{\min}, d_{\max})$	LayoutVLM [35]	Distance between the two assets should fall within the range $[d_{\min}, d_{\max}]$.
		Ours	Distance between two objects to $[d_{\min}, d_{\max}]$, where bounds are inferred from reach and clearance requirements based on anthropometric data.
	$L_{\text{against wall}}(p_i, w_j, b_i)$	LayoutVLM [35]	Place an asset against wall w_j.
		Ours	Places the object against a specific wall while considering accessibility and clearance requirements for nearby interactions.
Orientation-based	$L_{\text{align with}}(p_i, p_j, \Theta)$	LayoutVLM [35]	Align two assets at a specified angle Θ.
		Ours	Aligns the rotations of two objects; the angle parameter Θ reflects task-oriented alignment (e.g., parallel or perpendicular configurations for joint use).
	$L_{\text{point towards}}(p_i, p_j, \Theta)$	LayoutVLM [35]	Orient one asset to face another with an offset angle Θ.
		Ours	Adjusts orientation so that an object's front faces the target, with Θ encoding preferred viewing or interaction directions (e.g., facing a desk, or seating area).
Height-based	$L_{\text{on top of}}(p_i, p_j, h)$	LayoutVLM [35]	Position one asset on top of another.
		Ours	Defines a vertical stacking relationship for placing smaller objects on surfaces while keeping sufficient interaction area.

Notation: p_i, p_j object poses; w_j wall index; b_i object bounding box; $d_{\min}, d_{\max}$ minimum/maximum center distance; Θ relative angle; h height offset.

3.2.3 VLM-based Sequential Group Optimization. After the previous stages, the **[G] Sequential Group Optimization** stage compiles a constraint program that encodes the layout criteria and intra- and inter-group spatial relations as differentiable, anthropometrically grounded expressions. Each constraint incorporates the functional rationale and human behavioral conditions defined in the earlier stages, allowing the optimization process to preserve both semantic plausibility and operational validity. We represent all constraints as differentiable functions that enable the use of gradient-based optimization. Following the general optimization structure of LayoutVLM [35], we formulated each violation as a continuous penalty for the object poses. However, unlike LayoutVLM, which derives its losses primarily from VLM-inferred spatial relations and physics-based collision terms, our violation terms are grounded in anthropometric reachability, functional clearance, and behavior-grounded spatial semantics. Specifically, we convert each inferred constraint into a differentiable penalty function:

- **Distance constraints** produce violations when the center-to-center distance deviates from the inferred $[d_{\min}, d_{\max}]$ range.
- **Orientation constraints** generate violations when object orientations differ from interaction-inferred alignment angles.

The optimization process minimizes the aggregated violation, $L = \sum_i w_i \cdot \text{violation}_i(\theta)$, where i enumerates all behavior-aware, anthropometrically grounded constraints, w_i controls the relative influence of each constraint in the optimization, and $\text{violation}_i(\theta)$ computes a continuous penalty based on the deviation of object poses θ from the constraint's target distance, orientation, or group-level configuration. We employ adaptive weighting where w_i prioritizes collision avoidance when objects overlap significantly (> 50% bounding box overlap) while allowing flexible refinement of behavioral constraints. Optimization proceeds sequentially by group. For each group, we first minimize the intra-group constraint violations. As subsequent groups are optimized, inter-group constraints are applied to maintain behavioral relationships between previously placed groups and the current group. The optimizer runs for 400 gradient-based iterations, and the final layout is rendered in Blender using optimized object poses and floor-plan coordinates in meters.

4 Technical Validation

The following subsections describe the validation procedure, geometry based evaluation, user perception study, and the results that assess the layout quality of our approach relative to prior LLM-based methods.

4.1 Validation Procedure

In technical validation, our goal is to assess how the proposed scene generation framework, grounded in human-object interaction reasoning and anthropometric data, affects the layout quality relative to prior LLM-based approaches. We adopted a validation framework commonly used in 3D scene generation research, consisting of two components: **geometry-based metrics** (*object-to-object* collision score and *object-to-floor* in-boundary score) [4, 12, 14, 35, 36] and a **user perception study** [35, 36].

Baseline. We compared our method with a state-of-the-art system, LayoutVLM [35]. We selected the baseline because it is a representative LLM-based layout method that frames scene synthesis as a differentiable constraint optimization problem. The optimization structure enables a controlled comparison, in which the only major difference lies in how the constraints are obtained. The baseline relies on LLM common sense, whereas our method infers constraints

from human-object interaction and anthropometric reference data. We did not include alternative LLM-based layout methods [4, 47] because constraining them with fixed assets or deterministic placement would undermine their core generative process, precluding a fair comparison in controlled human-subject experiments where treatment consistency is necessary.

Setup. The validation covered five room types (Bedroom, Lounge, Office, Kitchen, and Dining) with two scenes each, totaling ten test scenarios with an average of 9.9 assets per room. All scenes were generated within a standardized 5.5 × 5.5 × 2.5m floor plan. For furniture assets, we selected 3D assets from the verified and preprocessed Objaverse [9] for each room type following a prior methodology [4, 35]. For each method and scene, we generated five candidate layouts using different random seeds and selected the layout with the highest collision-free score for analysis. This procedure yielded a matched set of 20 layouts (10 scenes × 2 methods), which we used for both the geometry-based evaluation in this section and the expert perception assessment.

Anthropometric data. Since our approach incorporates human dimensions into the constraint inference process, we utilized synthetic anthropometric profiles to validate this capability across diverse user bodies. First, we identified the 5^{th}–95^{th} percentile ranges for six key body dimensions (e.g., forward reach, lateral reach, body breadth, and body depth). We then provided the LLM with several example records sampled within these ranges and prompt it to generate a plausible combination of body dimensions that remains within the same percentile bounds. A synthetic profile was generated for each scene when the constraints were inferred, enabling technical validation to cover diverse users.

4.2 Geometry-Based Evaluation

To ensure a standardized quantitative comparison of physical validity, we adopted the widely used collision-free score and in-boundary score [4, 12, 14, 35, 36]. These metrics were selected to verify the basic physical validity of the generated scenes, as physical feasibility (i.e., no overlaps, inside walls) is a necessary condition before a layout can be meaningfully evaluated for usability. For both metrics, higher values indicate greater physical plausibility. The **collision-free score** measures the proportion of object pairs that do not overlap:

$$\text{CF} = \frac{1}{N_c} \sum_{(i,j)} \mathbb{I}(d_{ij} > r_i + r_j), \tag{1}$$

where d_{ij} is the distance between the object centers, r_i and r_j are the bounding-circle radii of objects i and j, $\mathbb{I}(\cdot)$ returns 1 if the condition holds and 0 otherwise, and N_c is the number of evaluated object pairs. The **in-boundary score** captures whether all objects remain within the room:

$$\text{IB} = \frac{1}{N_o} \sum_{i} \mathbb{I}(0 \le x_{\min}(i), x_{\max}(i) \le W,\ 0 \le y_{\min}(i), y_{\max}(i) \le D), \tag{2}$$

where $x_{\min}(i), x_{\max}(i), y_{\min}(i)$, and $y_{\max}(i)$ denote the floor-plan bounding box extents of the object, W and D are the room width and depth, and N_o is the number of objects.

4.3 User perception study on the Generated Scene

Participants. We recruited 16 participants (six males and ten females; aged $M = 26.88, SD = 1.78$) who were professionals in interior design and architecture. All participants had completed a university degree in architecture or interior design, ensuring that they were familiar with spatial layout reasoning and furniture arrangements.

Procedure. The study was conducted online. For each layout, the participants were shown two rendered images, a top view and a perspective view of the scene, and rated the layout on five criteria using a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree). To support accurate judgment, the interface included orientation arrows and a metric floor plan (Figure 5). This resulted in 1,600 ratings (16 participants × 20 layouts × 5 criteria).

Measurements. We refined the evaluation criteria based on the baseline study [35] to capture the specific contributions of our method. The baseline originally evaluated layouts using Position, Orientation, and a combined physically grounded semantic alignment score (which aggregated physical plausibility and semantic consistency). However, a combined score limits the ability to distinguish between a layout that is simply logically correct versus one that is practically usable for a human. Since our approach emphasizes human-operational usability, we decomposed the baseline's composite metric into three granular criteria: semantic plausibility, physical plausibility, and functional usability. This separation allows us to isolate and measure how our anthropometric grounded constraints improve the detailed operability of the scene. Participants evaluated their agreement with statements regarding:

- **Position appropriateness:** whether each object is placed at an appropriate location for its function and use within the room (e.g., not unreasonably far from related furniture).
- **Orientation appropriateness:** whether each object faces a direction that matches its function (e.g., a chair facing a desk or TV facing the seating area).
- **Semantic plausibility:** whether the arrangement of furniture is logically consistent with the intended use of the space (e.g., a bed placed meaningfully relative to a bedside table).
- **Physical plausibility:** whether the layout appears physically feasible in terms of basic spatial rules (e.g., no obvious collisions or boundary violations).
- **Functional usability:** whether people seem able to approach, access, and operate each object in a realistic way (e.g., sufficient space to open drawers fully, pull out chairs, or stand and work on a surface).

4.4 Results of Technical Validation

Geometry-Based Evaluation Results. Regarding the comparison of physical validity, our method achieved a collision-free score of 90.4 and an in-boundary score of 73.4, compared to the baseline's scores of 92.2 and 63.2. Although our framework showed a slightly lower collision-free score (−1.8%), it demonstrated substantially improved boundary adherence (+10.2%). Collision-free and in-boundary score metrics provide objective, geometry-based indicators of physical

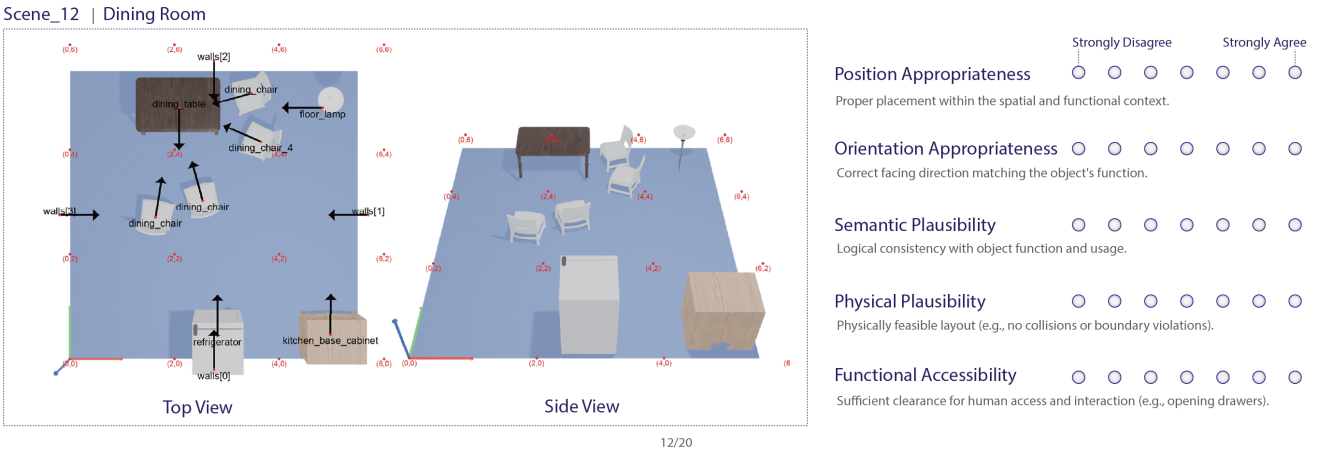

Figure 5: Interface for User perception study. Red markers indicate floor-plan coordinates in meters, and arrows denote furniture orientations. Participants rated criteria on a 7-point Likert scale.

Table 2: Results of the user perception study. Values represent the median, with the interquartile range [1st quartile, 3rd quartile] shown in brackets. Statistical significance was determined using the Wilcoxon Signed-Rank test with Holm-Bonferroni correction ($^*p < 0.01$).

Measurements	Baseline	Ours	p_{adj}	r
Position appropriateness	3.00 [2.25, 4.00]	**5.00** [4.38, 6.00]	<0.01	0.85
Orientation appropriateness	3.00 [2.00, 3.25]	**5.00** [4.50, 6.13]	<0.01	0.88
Semantic plausibility	3.00 [2.75, 3.63]	**5.00** [4.38, 6.00]	<0.01	0.86
Physical plausibility	3.00 [2.00, 4.13]	**5.00** [5.00, 6.00]	<0.01	0.78
Functional usability	3.00 [2.00, 4.00]	**5.00** [4.50, 6.00]	<0.01	0.77

plausibility by measuring whether objects overlap and whether assets remain within room boundaries.

User Perception Study Results. When examining the results of the user perception study, our method consistently outperformed the baseline across all five evaluation criteria (Table 2). Our method achieved a median of 5.0 in all metrics, whereas the baseline remained at a median of 3.0. Looking at the distribution, our method achieved 1st quartile values ranging from 4.375 to 5.0 and 3rd quartile values from 6.0 to 6.125. In contrast, the baseline showed significantly lower performance, with 1st quartile values between 2.0 and 2.75, and 3rd quartile values between 3.25 and 4.125. Notably, while the baseline achieved a higher collision-free score, this difference did not translate into improved user perception. Statistical analysis using the Wilcoxon Signed-Rank test with Holm-Bonferroni correction confirmed that these differences were statistically significant across all five criteria (adjusted $p < 0.01$). Furthermore, the effect sizes were large ($r > 0.77$). These results indicate that experts recognize our layouts as substantially more suitable for human operation than the baseline.

Limitations and Motivation for Usability Study. Geometry-based evaluation metrics have limitations when assessing human usability. First, they evaluate scenes under fixed-object assumptions (e.g., non-articulated drawers), and therefore cannot consider dynamic object functionality or actual human behavioral patterns. Second, while the user perception study complements these metrics by incorporating expert judgment, it relies solely on visual estimation from static renderings. Evaluating a layout visually does not guarantee that it supports actual physical operations—such as comfortable reaching or unhindered movement—especially for users with diverse body dimensions. This limitation motivated the human-operational usability study, which is a task-based user study to evaluate the generated scenes in realistic usage scenarios.

5 User Studies on Human-Operational Usability

To investigate whether our behavior-aware and anthropometric approach translates into actual operational benefits, we conducted a human-centered usability evaluation based on layouts parameterized by each participant's measured anthropometric data. We designed two complementary user studies to examine how participants physically moved and acted within these scenes from different perspectives. The experimental procedure flow for both sessions is illustrated in Figure 6.

- **Individual sessions** ($N = 20$): Provides controlled, quantitative measures of how anthropometrically grounded, behavior-aware layouts support task performance. Through structured single-user tasks, we captured movement behaviors, quantified task efficiency and interaction-space usage, and obtained data that enabled the statistical assessment of our approach.
- **Group sessions** ($N = 18$, six teams of three): Evaluates the same layouts in realistic multi-user scenarios where multiple occupants naturally share space. By analyzing cumulative movement trajectories and assessing post-task interview

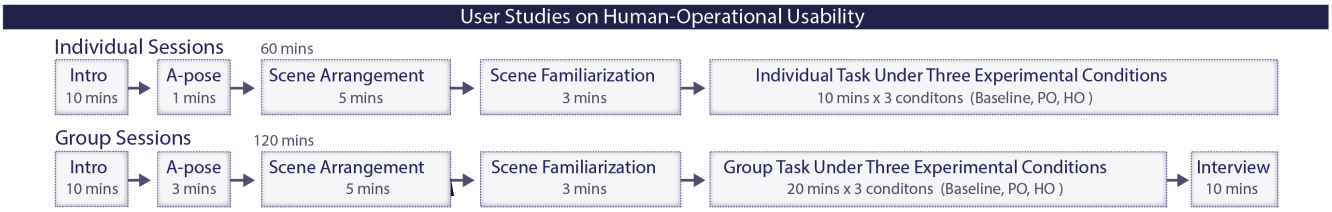

Figure 6: Experimental Procedure Flowcharts for Individual and Group Sessions.

feedback, we captured how layouts support spatial negotiation and collaborative activities, which are conditions that reflect real-world shared environments.

5.1 Setup for Individual and Group Studies

The following experimental conditions, environments, apparatus, and data collection procedures were applied consistently across both the individual and group studies.

5.1.1 Experimental Conditions. We evaluated three conditions: a geometry-based **baseline [35]** and two versions of our framework, **Passage-Only (PO)** and **Human-Operational (HO)**, which vary only in the type of anthropometric parameters injected into spatial constraints (Figure 7).

- **Baseline:** Layouts generated without any personalized anthropometric information; spatial relations are determined solely by generic geometric constraints and collision checks.
- **PO:** Our framework with *static anthropometric dimensions (e.g., body width, torso depth)* applied to guarantee minimal passage width and basic navigability.
- **HO:** Our framework with *movement-based anthropometric dimensions (e.g., extended arm reach, forward reach, seating buttock-to-toe length)* applied to ensure adequate operational space for human-object interactions.

PO and HO share the same spatial relationship and constraint types; only the minimum and maximum distances differ: static body sizes in PO versus movement envelopes in HO. This design enables two key comparisons: comparing PO and HO isolates the effect of dynamic, task-aware anthropometrics, because any usability differences directly reflect the added value of modeling movement-based operational space. Comparing baseline and HO assesses the full benefit of our framework, which integrates semantic scene understanding, spatial relation inference, and personalized anthropometric parameterization. This comparison highlights how behavior-aware layouts improve upon geometry-based methods by aligning spatial constraints with actual human movement.

5.1.2 Evaluation Environments. We selected office and lounge environments as representative indoor spaces, where diverse object arrangements and human activities occur naturally, supporting work, collaboration, storage, navigation, and rest tasks, while maintaining sufficient experimental control. All three conditions (Baseline, PO, HO) used the same 5.5 × 5.5 × 2.5m floor plan and were generated following the same procedure used in the technical validation. For the user studies, the PO and HO were instantiated on a per-participant basis by substituting each participant's measured anthropometric profile into the distance constraints, whereas the baseline remained purely geometry-based. Across all conditions, the office and lounge scenes shared the same furniture sets and overall scene topology (Figure 7).

5.1.3 Motion Capture and Apparatus. We employed a markerless motion capture setup using EasyMocap [33] to obtain absolute human coordinates and body measurements for both scene generation and human-scene interaction evaluation. We installed eight RGB cameras (Logitech webcams) in a $9 \times 9 \times 2.5$m indoor space and performed intrinsic and extrinsic calibration using a 4×7 checkerboard [33]. The calibrated cameras recorded synchronized multi-view videos at 4K/60fps within a capture volume of $5.5 \times 5.5 \times 2.5$m. To reconstruct the 3D body motion, the recorded data were processed frame-by-frame through (1) 2D keypoint extraction, (2) 3D triangulation, and (3) SMPL fitting. The experimental setup, camera placement, and captured data types are shown in Figure 8.

5.1.4 Anthropometric Data Collection. Participant-specific anthropometric measurements were extracted from the fitted SMPL models to parameterize the distance constraints used in the PO and HO conditions. Accordingly, we adopted the pose-independent 3D anthropometry method of Bojanić et al. [3], which predicts standardized body measurements from sparse 3D landmarks. We selected this method because it accurately reconstructs canonical A-pose dimensions regardless of the participant's actual pose during capture. This eliminates the need for strict A-pose scans while maintaining accuracy comparable to dense-geometry approaches. From each participant's reconstructed SMPL body, we first extracted the 3D coordinates of the 70 body landmarks, and passed them through the pre-trained prediction model to obtain a set of standardized anthropometric dimensions. For additional task-specific operational dimensions not directly provided by the Bojanić model, such as specific reach distances or joint-to-joint separations relevant to our interaction scenarios, we computed the Euclidean distances between the selected pairs of SMPL landmarks. The resulting combined profile served as the quantitative basis for instantiating personalized object–object distance constraints.

6 User Study 1: Individual Sessions

Individual sessions examined how the three scene conditions–Baseline, PO, and HO–affect the task-level usability for a single user in personalized layouts. Building on anthropometric personalization and motion capture framework, this study focuses on a controlled, quantitative evaluation of human–operational performance. Specifically, we compared three conditions in terms of

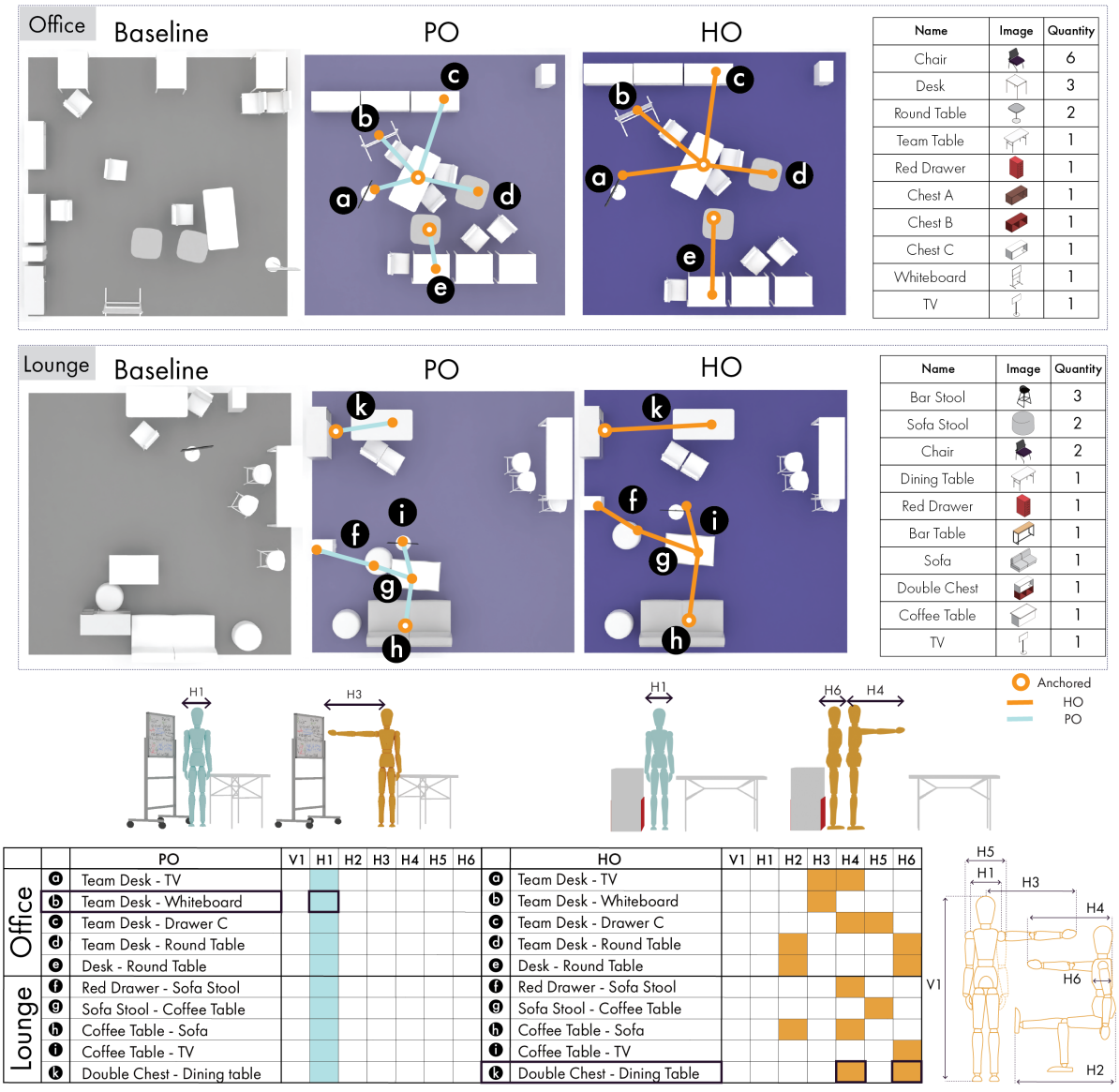

Figure 7: Baseline, PO, and HO layouts across office and lounge environments.

task completion time, trajectory efficiency during navigation, and interaction-space utilization around key objects.

6.1 Experimental Setup

Participants. The individual sessions involved 20 participants: 10 in the office space (four males and six females; aged $M = 24, SD = 1.89$) and 10 in the lounge space (five males and five females; aged $M = 24.3, SD = 2.63$).

Procedure and Task. At the beginning of each session, we explained the study purpose and procedure and then captured each participant in an A-pose in an empty space using the motion capture setup. We fitted the captured data to the SMPL models, computed anthropometric measurements, and used these measurements to instantiate the PO and HO layouts for each participant. Baseline layouts were fixed for each environment. While one researcher provided task instructions and a short practice session, the others prepared the physical environment by arranging furniture according to the designated layout. The order of the three layout conditions (Baseline, PO, HO) was counterbalanced by using a Latin square. The participants received safety instructions and an explanation of furniture naming conventions before the main trials. For each assigned environment (office or lounge), we defined ten tasks that reproduced everyday situations, such as navigating between furniture, sitting, and opening drawers, to ensure that all furniture items were used at least once (Table 3). Task instructions

Table 3: Sequential task list for office and lounge environments.

Task Sequence for Office	Task Sequence for Lounge
O1. Take box out of Chest C, place it on top.	L1. Move to Sofa Stool 1, sit down.
O2. Move box onto Desk C, sit in front of it.	L2. Take cup from Coffee Table, move to Sofa Stool 2, sit down.
O3. Stand up, take pen from third compartment on Red Drawer.	L3. Holding cup, move to Bar Table, sit down.
O4. Move to Round Table 2, use remote control to turn on TV.	L4. Holding cup, walk to Double Chest, place cup inside compartment.
O5. On Round Table 1, write two content names from TV onto paper.	L5. Find remote control in Red Drawer, move to Sofa, sit down.
O6. Holding paper, move to Chest B, take out board marker.	L6. From Sofa, turn on TV, find content, read title aloud.
O7. Holding paper and marker, move to Team Table, sit down.	L7. Place remote control on Coffee Table.
O8. Move to Whiteboard, copy the two words you wrote.	L8. Move to Double Chest, take out pen and paper.
O9. Sit at Team Table, read Whiteboard words backward from right.	L9. Move to Dining Table, sit, write content title on paper.
O10. Move to Desk A, sit down, write remembered words onto paper.	L10. Stand up, place paper and pen in lower compartment of Double Chest.

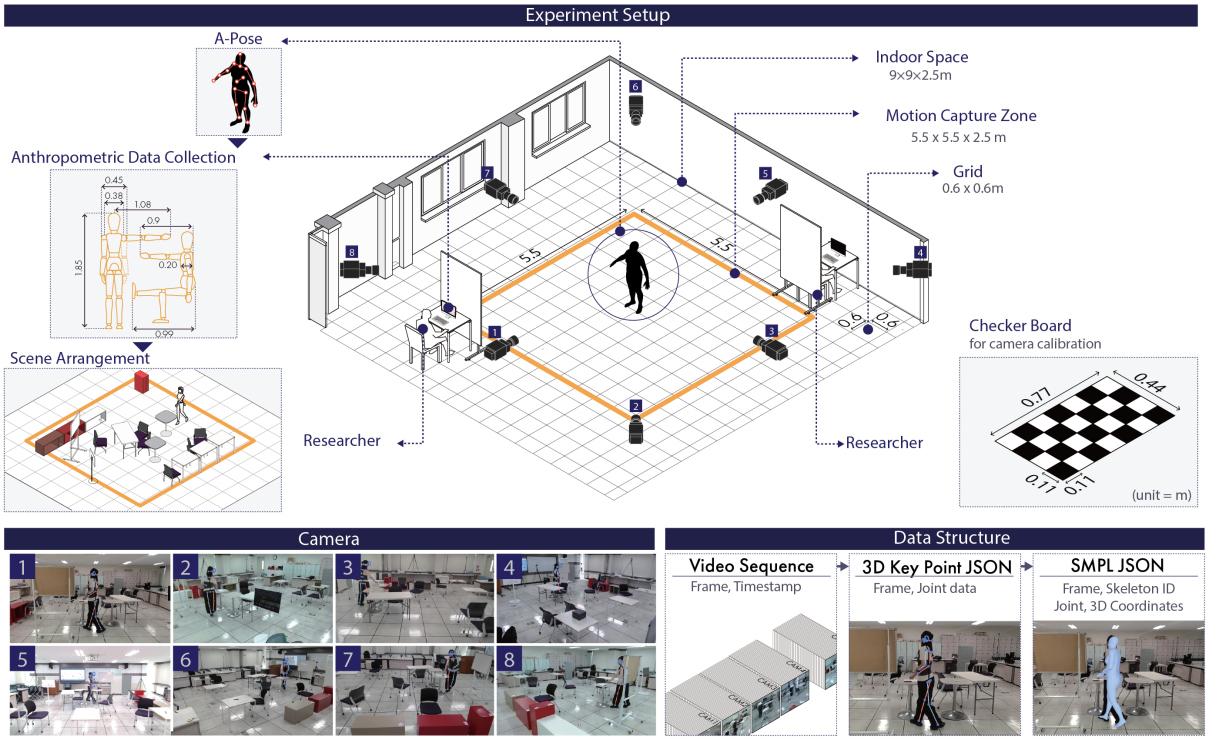

Figure 8: Experimental environment, motion capture setup, and data processing pipeline. The illustration depicts the physical configuration of the 8-camera array and capture zone, alongside the sequential data flow from raw video to 3D SMPL model reconstruction. All spatial measurements are in meters.

followed the *object–action–target* format [1, 44]. To preserve the initial layout, the participants were instructed not to move furniture except for chairs. The tasks were connected sequentially; the endpoint of one task served as the starting point for the next task. Each participant executed ten tasks in their assigned space under all three conditions, resulting in 60 recorded interaction sequences (20 participants × 3 conditions), with each session lasting approximately 60 minutes.

6.2 Measurements and Data Analysis

Our measurement set was designed to evaluate whether the proposed anthropometric constraints yield measurable usability improvements. Each metric links the imposed anthropometric constraints to a specific aspect of use: task performance, trajectory efficiency, or interaction-space utilization. We focused on four metrics:

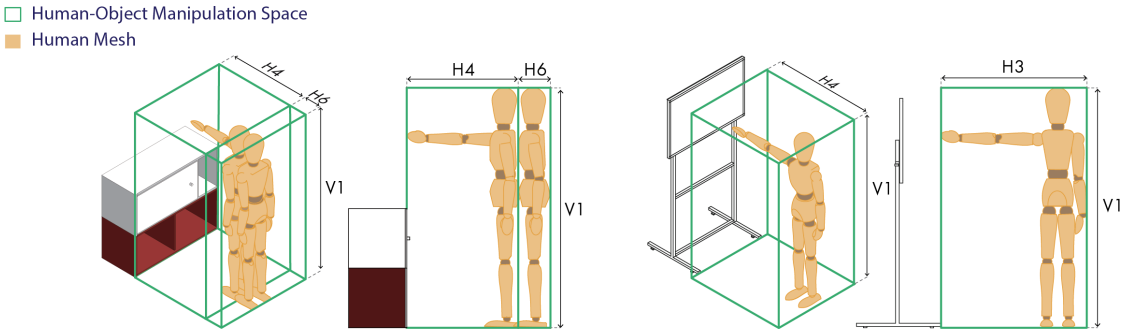

Figure 9: Definition of Human-Object Manipulation Space. The green wireframe defines the Human-Object Manipulation Space, a bounding volume parameterized by the participant's specific anthropometric dimensions (e.g., vertical height V1, horizontal reach H3, H4).

Task Completion Time. We measured the task completion time from the moment the participants began to act after receiving the full task instructions to the moment they verbally indicated completion. Three researchers independently reviewed all recordings at the frame level to annotate precise start and end points, and cross-checked their annotations to ensure consistency.

Trajectory Counting. To analyze how layouts shape movement paths, we projected the SMPL pelvis joint onto a 2D plane to generate time-series trajectories and then counted the number of distinct paths taken. Trajectories were treated as distinct when they involved different detours around furniture or obstacles, thereby quantifying the number of different paths that participants used to complete identical tasks.

Sequence Action Labeling. To capture the action-level differences in movement behavior, we used MMAction2 [7], a spatio-temporal action detection model, to automatically label the participant action sequences. We divided each video into short 6-frame segments and detected the actions for each segment. For every 6-frame segment in each of the three camera views, we extracted the top two action labels with confidence scores greater than 0.6. Then, we combined the labels from the three views and weighed them by confidence to obtain the final action sequence for each task.

Volumetric Occupancy Ratio. This metric measures the extent to which the intended interaction-space around an object is used by the participant. We first defined a human-object manipulation space as a 3D bounding box around each target object based on object-specific interaction distances derived from each participant's anthropometric profile. The participant's height was set to the vertical extent, and the anthropometric distance constraints determined the width and depth (Figure 9). We focus on objects with movable parts that require sufficient clearance: a double chest, a single chest, a whiteboard in the office, and a red drawer in the lounge. Using Blender and a Python script, we computed (1) the accumulated 3D volume occupied by the participant's body during task execution (the temporal union of human mesh volumes across all frames) and (2) its intersection with the human-object manipulation space.

M denotes the cumulative human mesh volume (temporal union across frames) and B denotes the defined Human-Object Manipulation Space (bounding box). The ratio is defined as:

$$\text{Volumetric Occupancy Ratio} = \frac{\text{Vol}(M \cap B)}{\text{Vol}(B)}$$

where $\text{Vol}(\cdot)$ represents the volume operator and $\cap$ denotes the intersection. We computed this ratio for each task and averaged the results to obtain a final metric per participant.

Experimental outcomes were analyzed using mixed design analyses of variance (ANOVAs) with layout (Baseline, PO, HO) as a within-subjects factor and environment (office, lounge) as a between-subjects factor. For each metric, we first tested the layout × environment interaction to verify consistent effects across environments, then examined the Layout main effect. Given the modest sample size, follow-up pairwise comparisons used Wilcoxon signed-rank tests with Bonferroni correction ($\alpha = 0.017$) as a conservative approach.

6.3 Results of User Study 1: Individual Sessions

Mixed-design ANOVAs revealed significant main effects of layout for task completion time ($F(2,36) = 4.38$, $p = 0.020$), trajectory counting ($F(2,36) = 7.45$, $p = 0.002$), sequence action labeling ($F(2,36) = 32.49$, $p < 0.001$), and volumetric occupancy ratio ($F(2,36) = 66.12$, $p < 0.001$). No significant layout × environment interactions were observed (all ps > 0.05), indicating consistent effects across both environments. Below, we report detailed comparisons for each metric.

6.3.1 Task Completion Time. A lower task completion time indicates a more efficient layout. When participants perform the same sequence of tasks, faster completion implies that the layout allows shorter or less obstructed routes and easier access to target objects. Therefore, we compared the total time required to complete ten sequential tasks in an individual study, where each task endpoint served as the starting point for the next task. In the office environment, HO achieved a mean completion time of 14.87 seconds, compared with 17.11 seconds for the baseline and 17.22 seconds for PO (Figure 10a). The difference between baseline and HO was

statistically significant ($p < 0.01$). HO achieved the shortest completion time, requiring 13.1% less time than baseline. Similar patterns were observed in the lounge environment, where HO required 9.14 seconds, PO required 10.66 seconds, and baseline required 11.59 seconds (Figure 10b). The task completion time differences in the lounge showed significant differences between baseline and HO ($p < 0.01$) with 21.1% reduction compared to baseline.

6.3.2 Trajectory Efficiency. We measured the layout trajectory efficiency using trajectory counting and sequence action labeling to assess whether the layouts induced intuitive path selection and to analyze the economic efficiency through the number of trajectories used.

Trajectory Counting. A lower trajectory count indicates a more intuitive and efficient layout (Figure 11). If the count is 1, all participants who experienced that layout followed the same path, suggesting that the scene allowed a clear and natural route. The mean trajectory counts across all sequences showed a consistent pattern of HO < baseline < PO in both office and lounge layouts, with HO demonstrating the lowest trajectory count. In the office, HO generated an average of 2.5 distinct paths per task across all participants, the baseline generated 2.6 paths, and PO generated 3.1 paths, although these differences were not statistically significant (Figure 10c). In the lounge, HO showed 1.6 paths, baseline showed 1.9 paths, and PO generated 3.0 paths (Figure 10d). HO required fewer paths compared to PO ($p < 0.01$) with a 46.7% reduction, and PO showed more paths than baseline ($p < 0.05$).

Sequence Action Labeling. We counted frame-by-frame labeled actions to determine the action sequences required for task completion, and calculated the mean action counts per task. A lower action count indicates a more efficient trajectory; additional actions typically arise from unnecessary behaviors, such as moving obstacles or making corrective adjustments. For example, in task L4, participants were required to stand up from the bar table while holding a cup, walk to the double chest, and place the cup inside a compartment. In the HO lounge layout, P12 completed the task using four actions {sit-hold-walk-stand}. In the baseline layout, an extra 'bend' action was required to move an obstructing stool, resulting in {sit-hold-walk-stand-bend-stand}. Therefore, fewer actions reflected a more direct and efficient trajectory for task completion. In office layouts, HO required an average of 4.79 actions, whereas baseline required 6.99 actions and PO required 8.95 actions (Figure 10e). HO required significantly fewer actions than both baseline ($p < 0.05$) with 31.5% reduction and PO ($p < 0.01$) with 46.5% reduction. In lounge layouts, HO required an average of 2.95 actions, baseline required 4.18 actions, and PO required 5.26 actions (Figure 10f). HO showed significantly lower action counts than both baseline ($p < 0.001$) with 29.4% reduction and PO ($p < 0.001$) with 43.9% reduction.

6.3.3 Interaction-Space Utilization. For the Volumetric Occupancy Ratio, higher values indicated better interaction-space utilization; participants occupied more of the intended human-object manipulation space and could operate objects more smoothly. Qualitative examples illustrate this pattern (Figure 12). In the PO office layout, P4 could only use a part of the Whiteboard because the Team Table was placed too close; therefore, the accumulated body volume within the Whiteboard's manipulation box remained low. In the baseline lounge layout, P17 had to work sideways at the Double Chest because a Sofa Stool blocked the front access area, again yielding sparse occupancy within the manipulation space. In contrast, in both scenarios, the HO layouts showed dense, evenly distributed occupancy within the manipulation boxes while participants performed the associated tasks. In the office layouts, HO showed an average occupancy ratio of 0.25, PO showed 0.18, and baseline showed 0.19 (Figure 10g). The lounge layouts showed similar results (Figure 10h). In office layouts, HO showed significantly higher occupancy ratios than both baseline ($p < 0.001$) with 31.6% increase and PO ($p < 0.001$) with 38.9% increase. Lounge layouts showed HO significantly outperforming baseline ($p < 0.001$) with 52.9% increase, and PO showing higher ratios than baseline ($p < 0.01$) with 11.8% increase.

6.3.4 Summary. Across all four metrics, the HO condition consistently provided better task-level usability than the baseline and PO layouts.

- **Task Completion Time:** HO demonstrated the shortest completion times in both the office and lounge environments, whereas PO showed no improvement over baseline.
- **Trajectory Counting:** The mean trajectory counts followed a consistent pattern of HO < baseline < PO, indicating that HO layouts induced more intuitive routes, whereas PO often led to additional detours.
- **Sequence Action Labeling:** HO required the fewest actions per task in both environments, reducing extra behaviors such as moving obstacles or making corrective adjustments compared to the baseline and PO.
- **Volumetric Occupancy Ratio:** HO achieved the highest interaction-space utilization, whereas PO and baseline left more of the intended manipulation space unused.

7 User Study 2: Group Sessions

Although individual sessions provide controlled, single-user measurements of task-level performance, many real-world environments are occupied by multiple people who perform overlapping activities over extended periods. To account for these multi-user dynamics, where layout usability depends not only on individual route efficiency but also on how occupants share space, experience mutual interference, and adapt their movements in response to others, we conducted group sessions in which small teams performed continuous collaborative work in the same baseline, PO, and HO layouts used in the individual study. In addition to the motion capture data, we conducted post-interviews to qualitatively examine perceived crowding, coordination, and workflow support in each layout condition.

7.1 Experimental Setup

Participants. The group sessions involved 18 participants organized into six teams. In the office space, three teams (nine participants: three males and six females; aged $M = 23.56$, $SD = 2.92$) performed tasks, whereas the remaining three teams (nine participants: four males and five females; aged $M = 23.78$, $SD = 2.49$) performed tasks in the lounge space.

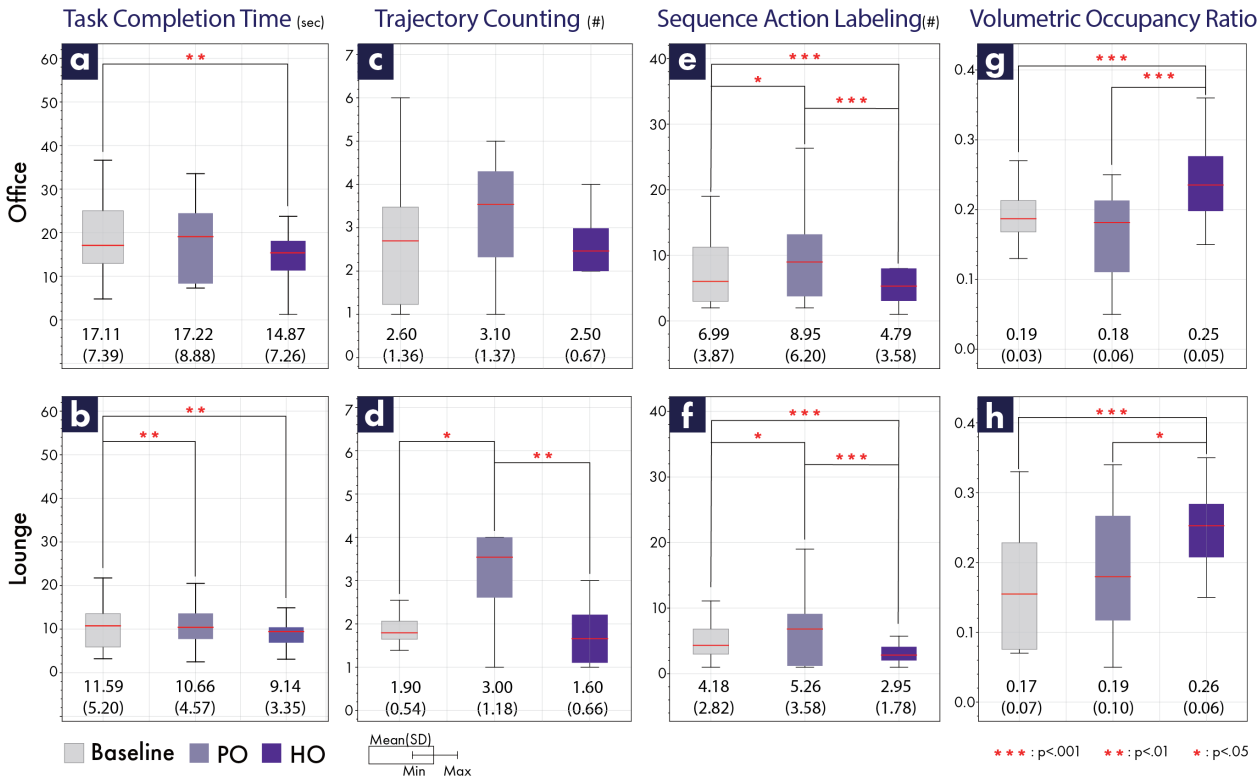

Figure 10: Results of Measurements. (a, b): Task Completion Time. (c, d) : Trajectory Counting. (e, f) : Sequence Action Labeling. (g, h) : Volumetric Occupancy Ratio.

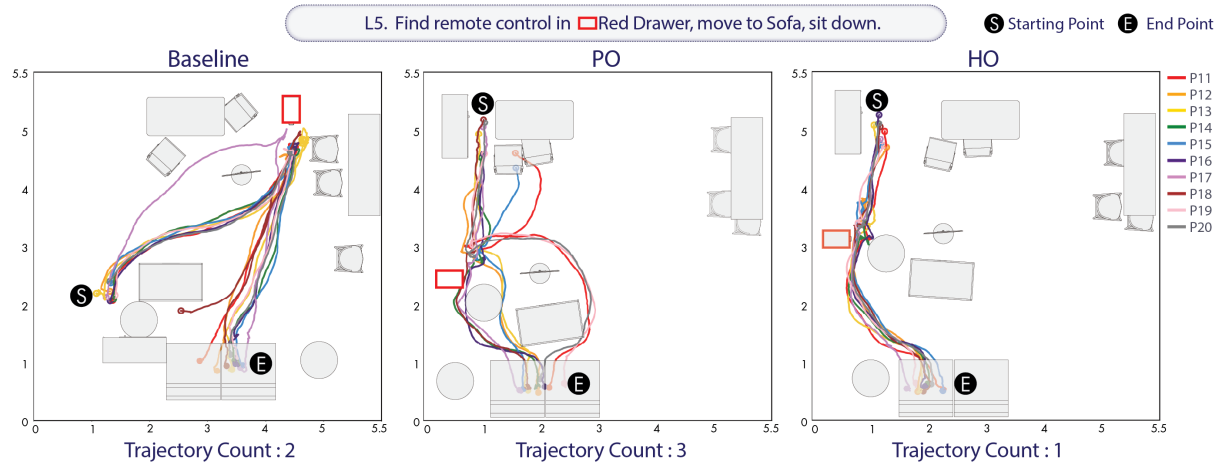

Figure 11: Examples of trajectory counting and participant trajectory visualization for L5.

Procedure and Tasks. At the start of each session, the teams were briefed on the study and recorded using the same motion capture and anthropometric framework. For scene generation, we constructed a team-level anthropometric profile by taking the maximum value among the three team members for each dimension and used this profile to instantiate the PO and HO constraints. After safety instructions and a short practice session, we arranged the furniture for the current condition and asked teams to perform a series of collaborative coloring and sorting tasks in each of the three layouts (Baseline, PO, HO), following spatial-interaction protocols

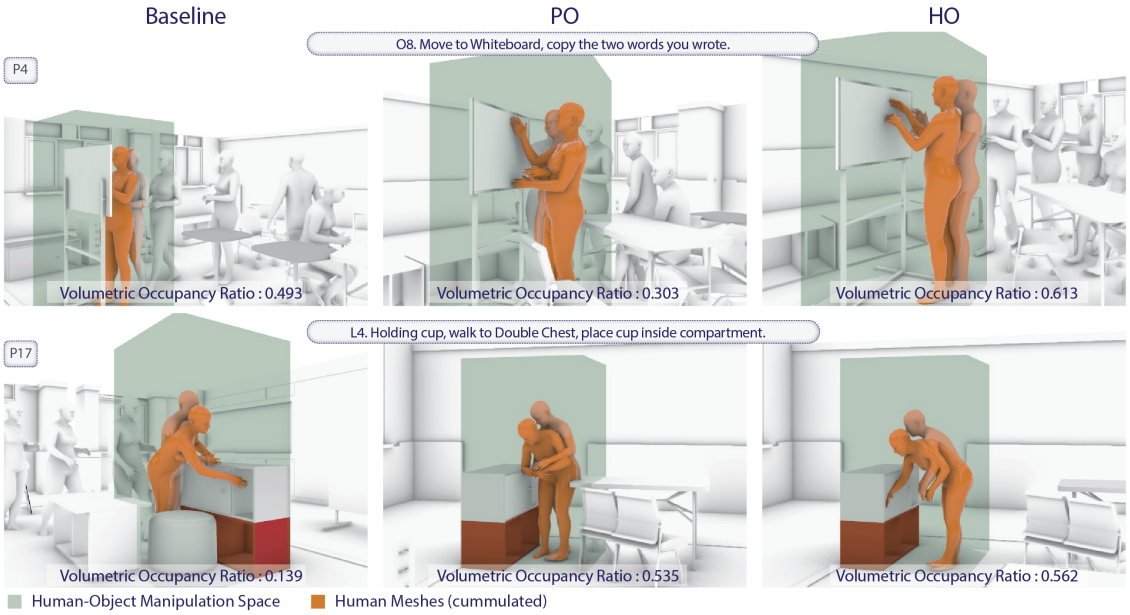

Figure 12: Graphical Examples of Human-Object Manipulation Space (green box) and SMPL meshes for P4 (O8) and P17 (L4).

adapted from [23, 24]. In each layout, participants viewed templates on the TV, retrieved materials from storage units, worked at individual desks using hints from the shared space, reconvened to discuss the sorting criteria shown on the TV, and attached the final templates to the whiteboard. The task materials were distributed across the space to encourage resource exchange and movement, naturally inducing circulation, path intersections, and clustering. After all three conditions, we conducted individual post-interviews in a separate room. Each team session lasted approximately 120 minutes, yielding 18 group interaction sequences in total (six teams × three conditions).

7.2 Data Analysis and Visualization

Spatial Usage Visualization. To analyze group movement patterns, we developed a heatmap visualization of the mean speed. For each grid cell (i, j) in a 1024×1024 discretizations of the space, the average movement velocity was calculated as follows:

$$\bar{v}_{i,j} = \frac{\sum_{p=1}^{P} \sum_{t=1}^{T} v_{p,t} \cdot \delta_{i,j}(x_{p,t}) \cdot \Delta t}{\sum_{p=1}^{P} \sum_{t=1}^{T} \delta_{i,j}(x_{p,t}) \cdot \Delta t}$$

where $v_{p,t} = \|x_{p,t+1} - x_{p,t}\| \cdot \text{fps}$ is the velocity of participant p at time t, and $\delta_{i,j}$ equals 1 when position $x_{p,t}$ falls within cell (i, j). Gaussian smoothing ($\sigma = 0.01$m) was applied to create continuous representations. Consistent color scaling across conditions enables direct comparison of spatial usage patterns.

7.3 Results of User Study 2: Group Sessions

For each team, we aggregated their movements over the entire task sequence and visualized them as heatmaps. We observed salient differences in the spatial usage patterns. We report the detailed findings below, integrating these observed movement behaviors (Figure 13) with insights from participant interviews.

7.3.1 Functional Visibility and Physical Accessibility. In the baseline condition, geometric placement often failed to support functional visibility and access, leading to unnecessary movement. In the office, the TV orientation forced repeated back-and-forth movements to read the screen, whereas in PO and HO layouts, no additional movements were required (Figure 13). In interviews, 7 of 9 participants (P23–P29) identified the TV as the most inconvenient asset; P28 noted, "*I wanted to rotate the TV toward the table where people sit,*" and P29 remarked, "*The path required just to see the TV screen was too long.*"

Similar issues emerged in the lounge, where participants identified the TV and obstructed Double Chest as major pain points. Three participants (P30, P31, P36) reported the TV was too distant and only visible from a diagonal angle. Furthermore, access to the Double Chest was obstructed by Sofa Stool 1, which blocked frontal access and forced side-only interaction. Four participants (P30–P32, P38) identified this obstruction as the primary source of inconvenience, explicitly noting that the stool hindered both material retrieval and circulation.

7.3.2 Trajectory Continuity and Operational Clearance. Focusing on the distinction between HO and PO, trajectory visualizations revealed that PO layouts often caused trajectory breaks and operational bottlenecks due to cramped configurations (Figure 13).

In the office, HO showed continuous movement, whereas PO exhibited breaks between Chest B and the whiteboard, and between Round Table 1 and Desk 1. As P28 reported about PO, "*Because the Round Table was attached to the Team Table, sitting was uncomfortable, and during standing meetings, I had to walk around excessively.*"

Similar patterns were observed in the lounge. For the Double Chest, HO provided adequate human-object manipulation space, yielding smooth, front-facing trajectories. In PO, the Dining Table encroached on the chest's manipulation space, narrowing the

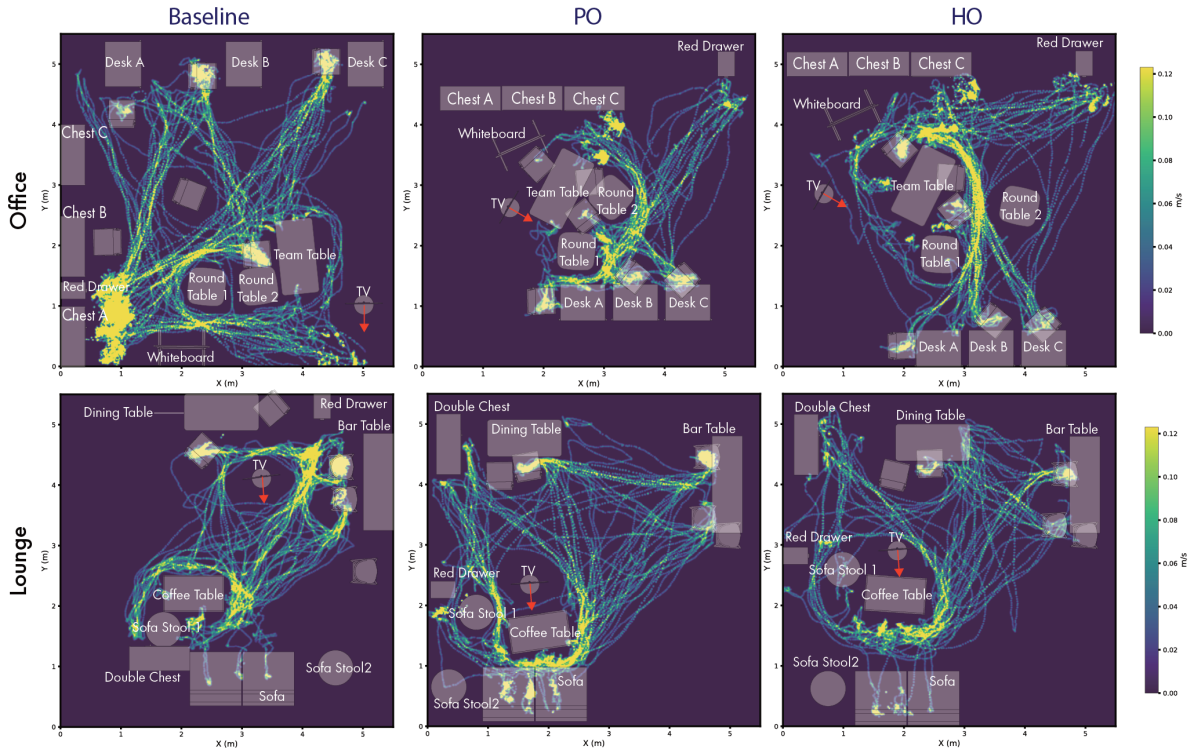

Figure 13: The Visualization of Mean speed heatmap of Team 1: P21-23 (Office; Top) and Team 6 : P36-38 (Lounge; Bottom)

movement corridor; P22 observed, "*The double chest door didn't open fully, and a bottleneck formed in front of it.*" Four participants (P30, P36–P38) described the distance between the upper chest and dining table as too narrow, making it difficult to use the chest and sit in nearby chairs. In contrast, participants described HO as more spacious and better at using the available space, emphasizing that moving between the upper chest and other work areas felt unconstrained. P30 specifically noted that opening the upper chest was improved and rated HO as the most favorable layout.

7.3.3 Summary. Across group sessions, the HO condition consistently provided better group-level usability than both the baseline and PO layouts, which is consistent with both the trajectory visualizations and participant reports.

- **Functional Visibility and Accessibility:** HO ensured that shared resources such as the TV, sofa, and storage furniture were accessible at appropriate distances and directions. In contrast, baseline layouts frequently compromised visibility or physically blocked access, leading to unnecessary movement.
- **Trajectory Continuity and Operational Clearance:** HO supported continuous movement patterns around key work areas, minimizing the bottlenecks observed in PO. Although PO provided basic passage width, it failed to account for operational encroachment (e.g., furniture blocking storage access), resulting in trajectory breaks and user complaints.

These qualitative observations complement our quantitative findings, demonstrating that anthropometrically grounded layouts not only improve individual task performance, but also facilitate more natural and efficient group interactions in shared spaces.

8 Discussion

Our findings demonstrate that behavior-aware anthropometric constraints substantially improve layout usability beyond semantic plausibility. We discuss implications across four key areas: functional relationships in generated layouts, operational clearance requirements, applications to XR environments, and limitations with future research directions.

8.1 Functional Relationships in Behavior-Aware Layouts

The comparison between the semantic-driven baseline and HO demonstrates that semantic plausibility does not guarantee usability. Although the semantic-driven baseline correctly grouped related objects—such as positioning a coffee table near a sofa, it often failed to provide the physical clearance necessary for interaction.

Our two user studies on human-operational usability clearly exposed these limitations, revealing substantial functional disconnections in the baseline layouts. For example, baselines frequently placed TVs at awkward viewing angles or positioned tables blocking whiteboard access—arrangements that were semantically valid but practically unusable. In contrast, our behavior-aware layouts explicitly inferred the operational envelope required for human actions.

HO preserved these functional relationships, allowing participants to maintain natural workflows without repeatedly walking to the TV or viewing it from an awkward diagonal angle. Overall, these results show that semantic-driven approaches alone are insufficient to capture human behavioral patterns or functional relationships between objects and that behavior-aware constraints materially improve the usability of generated spaces.

8.2 Operational Clearance for Manipulation

The comparison between PO and HO highlights a critical distinction: passage clearance is not equivalent to operational clearance. Although both conditions use anthropometric data, they differ in the definition of clearance. PO ensures only passage clearance based on body depth and width, whereas HO additionally enforces operational clearance for manipulation and interaction.

This distinction leads to substantial behavioral differences. Compared to PO, HO significantly reduced compensatory movements and trajectory interruptions, eliminating the need for users to reposition furniture or make detours around obstacles. The higher volumetric occupancy under HO quantitatively confirms that securing space for passing through does not automatically guarantee space for operating. Particularly in the group study, the lack of operational constraints in the PO caused practical bottlenecks: objects were placed too close together, obstructing seating at the round table or preventing chest doors from opening fully due to encroaching furniture. These issues appear as trajectory discontinuities in otherwise collision-free layouts. These comparisons confirm that even with the same anthropometric data, there is a difference between designing for mere passage clearance and designing for operational clearance. These findings suggest practical guidance for constraint selection: operational clearance (HO) provides measurable benefits when layouts involve frequent object manipulation—such as opening drawers, adjusting furniture, or accessing storage—whereas passage-only constraints (PO) may suffice for circulation-focused spaces with minimal interaction requirements. By clarifying when anthropometric-aware operational constraints yield usability gains, our work helps practitioners decide where to invest additional design effort.

8.3 Applications to 3D Environments

Our framework encodes anthropometric considerations as differentiable constraints, which opens opportunities for applications beyond static physical layout generation. We discuss three potential directions where behavior-aware scene generation could provide value, while noting that empirical validation in these domains remains for future work.

Adaptation to Device-Augmented Interactions. In XR environments, users often interact through controllers or perform wide gestural commands, effectively extending their kinematic volume beyond their physical body dimensions [2, 32]. Our framework could be extended to such device-specific operational envelopes, potentially enabling layouts that account for the additional clearance required during controller-based or gestural interactions. This represents an extension of our anthropometric parameterization, though the specific requirements would need to be validated through user studies in XR settings.

Enhancing Embodied Interaction Fidelity in XR. Maintaining immersion in XR depends on minimizing the gap between user intent and virtual constraints [39]. Notably, virtual environments allow instant spatial modifications without physical effort, making real-time anthropometric adaptation feasible—a capability impractical in physical settings. Applying our framework to XR implies that generated environments could help prevent the interaction mismatches—such as users having to overextend to reach a control or seeing their virtual hands penetrate a table surface. By optimizing layouts based on the specific anthropometric profiles of users (or avatars), our framework has the potential to support natural, collision-free workflows in virtual spaces, thereby contributing to higher embodied interaction fidelity without requiring manual post-processing.

Scalable Content Creation for AI agents and Digital Twins. Finally, for scenarios requiring mass production of synthetic scenes—such as AI agent training, robotic navigation, or digital twin simulations—our method may offer a pathway toward improved behavioral realism. Instead of static asset placement, users can specify high-level activities (e.g., *sitting and retrieving objects*). The system then generates layouts that provide the kinematic space required for these actions, potentially reducing the awkward motions or unnatural inverse kinematics often observed in simulations where virtual agents or robots struggle to navigate functionally disconnected spaces.

8.4 Limitations and Future Works

Our study focuses on investigating whether anthropometric and behavioral data can improve usability in computational scene generation. With this scope in mind, we discuss several limitations and future directions below.

First, regarding generalizability and validation scope, our technical validation covered 20 scenes and user studies focused on 6 scenes. Additionally, while our participant pool provided sufficient power to detect usability improvements within these contexts, the sample size limits the ability to make broad claims about diverse populations. Future research should conduct large-scale evaluations and expand validation to complex, highly regulated environments such as kitchens, bathrooms, or retail spaces to establish normative design guidelines. Second, our current framework focuses on horizontal arrangements due to the lack of parametric variability in our asset library. Consequently, vertical interactions—such as shelf heights or standing-to-sitting transitions—were not optimized. Future work should incorporate parametric object generation to enable full 3D optimization, allowing the system to adjust furniture heights based on individual anthropometric profiles. Third, regarding multi-user optimization, our current approach prioritized accessibility by using the maximum body dimensions within a group to ensure clearance for all. However, this revealed an inherent trade-off: ensuring geometric clearance for the largest user often resulted in uncomfortable interaction distances for shorter participants (P22). More broadly, this finding illustrates that the appropriate constraint priority may vary depending on the object relationship and intended interaction—some relationships benefit from closer placement for reach, while others require wider spacing for manipulation. Since our primary focus was to examine how

behavior and anthropometric information affects layout quality, we did not address adaptively balancing such competing constraints. Future research should explore adaptive constraint prioritization that considers the specific interaction context of each object pair, aligning with universal design principles for diverse user populations. Fourth, our VLM-based constraint inference was validated using representative furniture assets from preprocessed datasets. The system reliably inferred functional descriptions and interaction patterns for the tested furniture assets. However, since the VLM relies on visual cues and common sense, affordances that are not visually apparent (e.g., hidden mechanisms or non-standard opening directions) cannot be inferred without supplementary text information. Generalization to unconventional furniture designs—such as diverse forms, or ambiguous multi-functional pieces—remains unexamined. Finally, our current framework was instantiated and evaluated in physical environments; however, the same anthropometric and behavior-aware constraints can be extended to XR settings. As XR technologies become more prevalent in workplaces and residential environments [17, 23, 24], spatial layouts will increasingly need to accommodate physical and virtual interaction zones.

9 Conclusion

We presented Behavior-Aware Anthropometric Scene Generation, an approach that augments language-based layout generation with behavioral reasoning and anthropometric grounding, shifting the focus from visual and semantic plausibility to human-operational usability by examining how behavioral and anthropometric information affects layout quality. Our contributions are threefold: First, we introduce a two-stage scene generation framework that translates behavioral reasoning into differentiable spatial constraints, enabling gradient-based optimization of layout usability. Second, we defined explicit spatial constraints for operational clearance and interaction zones, ensuring that layouts accommodate dynamic human actions—such as reaching or opening drawers—rather than merely avoiding static collisions. Finally, we demonstrated through technical validation and user studies that behavior-centric metrics capture how people actually use space beyond conventional collision and boundary scores. Our behavior-aware anthropometric perspective represents a critical step toward functional operational environments in XR, embodied AI, and digital twin systems, where human movement, reach, and interaction become first-class design constraints.

Acknowledgments

This work was supported by the Technology Innovation Program (RS-2025-02317326, Development of AI-Driven Design Generation Technology Based on Designer Intent) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*. Springer, 422–440. doi:10.1007/978-3-030-58452-8_25

[2] Karan Ahuja, Vivian Shen, Cathy Mengying Fang, Nathan Riopelle, Andy Kong, and Chris Harrison. 2022. Controllerpose: inside-out body capture with VR controller cameras. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–13. doi:10.1145/3491102.3502105

[3] David Bojanić, Stefanie Wuhrer, Tomislav Petković, and Tomislav Pribanić. 2024. Pose-independent 3d anthropometry from sparse data. In *European Conference on Computer Vision*. Springer, 237–256. doi:10.1007/978-3-031-91575-8_15

[4] Ata Çelen, Guo Han, Konrad Schindler, Luc Van Gool, Iro Armeni, Anton Obukhov, and Xi Wang. 2024. I-design: Personalized llm interior designer. In *European Conference on Computer Vision*. Springer, 217–234. doi:10.1007/978-3-031-92387-6_17

[5] Pei Chen, Kexing Wang, Lianyan Liu, Xuanhui Liu, Hongbo Zhang, Zhuyu Teng, and Lingyun Sun. 2025. Exploring the role of Mixed Reality on Design Representations to Enhance User-Involved Co-Design Communication. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–29. doi:10.1145/3710979

[6] Francis DK Ching. 2023. *Architecture: Form, space, and order*. John Wiley & Sons.

[7] MMAction2 Contributors. 2020. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2.

[8] Albert Damon, Howard W Stoudt, and Ross A McFarland. 1966. *The human body in equipment design*. Harvard University Press. doi:10.4159/harvard.9780674491892

[9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13142–13153. doi:10.1109/cvpr52729.2023.01263

[10] Iman Dianat, Johan Molenbroek, and Héctor Ignacio Castellucci. 2018. A review of the methodology and applications of anthropometry in ergonomics and product design. *Ergonomics* 61, 12 (2018), 1696–1720. doi:10.1080/00140139.2018.1502817

[11] Yan Ding, Xiaohan Zhang, Chris Paxton, and Shiqi Zhang. 2023. Task and motion planning with large language models for object rearrangement. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2086–2092. doi:10.1109/iros55552.2023.10342169

[12] Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Xuehai He, S Basu, Xin Eric Wang, and William Yang Wang. 2023. LayoutGPT: Compositional Visual Planning and Generation with Large Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=Xu8aG5Q8M3

[13] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6047–6056. doi:10.1109/cvpr.2018.00633

[14] Siyi Hu, Diego Martin Arroyo, Stephanie Debats, Fabian Manhardt, Luca Carlone, and Federico Tombari. 2024. Mixed Diffusion for 3D Indoor Scene Synthesis. arXiv:2405.21066 [cs.CV] doi:10.48550/arXiv.2405.21066

[15] Begoña Juliá Nehme, Eugenio Rodríguez, and So-Yeon Yoon. 2020. Spatial user experience: A multidisciplinary approach to assessing physical settings. *Journal of Interior Design* 45, 3 (2020), 7–25. doi:10.1111/joid.12177

[16] Begoña Juliá Nehme, David Torres Irribarra, Patricio Cumsille, and So-Yeon Yoon. 2021. Waiting room physical environment and outpatient experience: The spatial user experience model as analytical tool. *Journal of Interior Design* 46, 4 (2021), 27–48. doi:10.1111/joid.12205

[17] Dooyoung Kim, Seonji Kim, Selin Choi, and Woontack Woo. 2024. Spatial Affordance-aware Interactable Subspace Allocation for Mixed Reality Telepresence. arXiv:2408.04297 [cs.ET] doi:10.1109/ismar62088.2024.00142

[18] Bokyung Lee, Minjoo Cho, Joonhee Min, and Daniel Saakes. 2016. Posing and acting as input for personalizing furniture. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. 1–10. doi:10.1145/2971485.2971487

[19] Bokyung Lee, Joongi Shin, Hyoshin Bae, and Daniel Saakes. 2018. Interactive and situated guidelines to help users design a personal desk that fits their bodies. In *Proceedings of the 2018 designing interactive systems conference*. 637–650. doi:10.1145/3196709.3196725

[20] Chenguo Lin and Yadong Mu. 2024. Instructscene: Instruction-driven 3d indoor scene synthesis with semantic graph prior. *arXiv preprint arXiv:2402.04717* (2024). doi:10.48550/arXiv.2402.04717

[21] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. 2023. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems* 36 (2023), 44860–44879. doi:10.48550/arXiv.2305.10764

[22] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* (2015). doi:10.1145/3596711.3596800

[23] Weizhou Luo, Mats Ole Ellenberg, Marc Satkowski, and Raimund Dachselt. 2025. Documents in Your Hands: Exploring Interaction Techniques for Spatial Arrangement of Augmented Reality Documents. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22. doi:10.1145/3706598.3713518

[24] Weizhou Luo, Anke Lehmann, Hjalmar Widengren, and Raimund Dachselt. 2022. Where should we put it? layout and placement strategies of documents in augmented reality for collaborative sensemaking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16. doi:10.1145/3491102.3501946

[25] Weizhou Luo, Zhongyuan Yu, Rufat Rzayev, Marc Satkowski, Stefan Gumhold, Matthew McGinity, and Raimund Dachselt. 2023. Pearl: Physical environment based augmented reality lenses for in-situ human movement analysis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15. doi:10.1145/3544548.3580715

[26] NASA. 1978. *Anthropometric Source Book. Volume 2: A Handbook of Anthropometric Data*. Webb Associates yellow springs oh.

[27] Binh Vinh Duc Nguyen and Andrew Vande Moere. 2024. The adaptive architectural layout: How the control of a semi-autonomous mobile robotic partition was shared to mediate the environmental demands and resources of an open-plan office. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20. doi:10.1145/3613904.3642465

[28] Julius Panero and Martin Zelnik. 1979. *Human Dimension and Interior Space: A Source Book of Design Reference Standards*. Watson-Guptill, New York.

[29] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. 2021. Atiss: Autoregressive transformers for indoor scene synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 12013–12026. doi:10.48550/arXiv.2110.03675

[30] Haoxuan Qu, Ziyan Guo, and Jun Liu. 2024. GPT-Connect: Interaction between Text-Driven Human Motion Generator and 3D Scenes in a Training-free Manner. *arXiv preprint arXiv:2403.14947* (2024). doi:10.48550/arXiv.2403.14947

[31] Raf Ramakers, Danny Leen, Jeeeun Kim, Kris Luyten, Steven Houben, and Tom Veuskens. 2023. Measurement patterns: User-oriented strategies for dealing with measurements and dimensions in making processes. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17. doi:10.1145/3544548.3581157

[32] Anthony Scavarelli and Robert J Teather. 2017. Vr collide! comparing collision-avoidance methods between co-located virtual reality users. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 2915–2921. doi:10.1145/3027063.3053180

[33] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. 2022. Novel View Synthesis of Human Interactions from Sparse Multi-view Videos. In *SIGGRAPH Conference Proceedings*. doi:10.1145/3528233.3530704

[34] Zejia Su, Qingnan Fan, Xuelin Chen, Oliver Van Kaick, Hui Huang, and Ruizhen Hu. 2023. Scene-aware Activity Program Generation with Language Guidance Supplementary Material. *ACM Trans. Graph* 42, 6 (2023). doi:10.1145/3618338

[35] Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, Manling Li, Nick Haber, and Jiajun Wu. 2025. Layoutvlm: Differentiable optimization of 3d layout via vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 29469–29478. doi:10.1109/cvpr52734.2025.02744

[36] Jia-Mu Sun, Jie Yang, Kaichun Mo, Yu-Kun Lai, Leonidas Guibas, and Lin Gao. 2024. Haisor: Human-aware indoor scene optimization via deep reinforcement learning. *ACM Transactions on Graphics* 43, 2 (2024), 1–17. doi:10.1145/3632947

[37] Qi Sun, Hang Zhou, Wengang Zhou, Li Li, and Houqiang Li. 2025. Forest2seq: Revitalizing order prior for sequential indoor scene synthesis. In *European Conference on Computer Vision*. Springer, 251–268. doi:10.1007/978-3-031-72698-9_15

[38] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. 2024. DiffuScene: Denoising Diffusion Models for Generative Indoor Scene Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20507–20518. doi:10.1109/cvpr52733.2024.01938

[39] Yujie Tao, Cheng Yao Wang, Andrew D Wilson, Eyal Ofek, and Mar Gonzalez-Franco. 2023. Embodying physics-aware avatars in virtual reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15. doi:10.1145/3544548.3580979

[40] Harold P Van Cott and Robert G Kinkade. 1972. *Human engineering guide to equipment design*. Department of Defense.

[41] Carlos Viviani, PM Arezes, Sara Braganca, Johan Molenbroek, Iman Dianat, and HI Castellucci. 2018. Accuracy, precision and reliability in anthropometric surveys for ergonomics purposes in adult working populations: A literature review. *International Journal of Industrial Ergonomics* 65 (2018), 1–16. doi:10.1016/j.ergon.2018.01.012

[42] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. 2021. Sceneformer: Indoor scene generation with transformers. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 106–115. doi:10.1109/3dv53792.2021.00021

[43] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. 2024. Move as You Say Interact as You Can: Language-guided Human Motion Generation with Scene Affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 433–444. doi:10.1109/cvpr52733.2024.00049

[44] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. 2022. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems* 35 (2022), 14959–14971. doi:10.48550/arXiv.2210.09729

[45] Zhen Wu, Jiaman Li, Pei Xu, and C Karen Liu. 2025. Human-object interaction from human-level instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11176–11186. doi:10.48550/arXiv.2406.17840

[46] Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. 2024. Physcene: Physically interactable 3d scene synthesis for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16262–16272. doi:10.1109/cvpr52733.2024.01539

[47] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. 2024. Holodeck: Language Guided Generation of 3D Embodied AI Environments. arXiv:2312.09067 [cs.CV] doi:10.1109/cvpr52733.2024.01536

[48] Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. 2023. Commonscenes: Generating commonsense 3d indoor scenes with scene graph diffusion. *Advances in Neural Information Processing Systems* 36 (2023), 30026–30038. doi:10.48550/arXiv.2305.16283