WORLDBRIDGE: BRIDGING INDOOR AND OUTDOOR SPACES FOR UNIFIED WORLD GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robots operating in the real world often need to complete continuous tasks across indoor and outdoor spaces, such as moving from a living room through a porch onto the street. However, existing methods typically treat indoor and outdoor scene generation as separate tasks, making it difficult to generate unified and spatially continuous open environments. This stems from two core challenges: (1) Continuous indoor-outdoor scene data is scarce, so reliable joint generation priors are hard to obtain; (2) Indoor layouts and outdoor connections need to be coordinated, while post-hoc stitching often yields unnatural transitions. To address these challenges, we propose WorldBridge, a procedural framework that bridges indoor and outdoor spaces to generate a unified open world. WorldBridge consists of three stages. First, *Spatial Contract Coordination* derives domain-specific contracts for indoor spaces, building envelopes, and outdoor environments, and reconciles them into a shared spatial contract. Second, *Contract-Conditioned 3D World Realization* jointly generates scene structure and assets under the shared spatial contract, producing a coherent 3D world. Finally, *Physics-Aware Dynamic Extension* attaches object and environmental dynamics to the structured scene, yielding an executable 4D world. Experiments demonstrate that WorldBridge reliably generates diverse, visually compelling, and spatially coherent indoor–outdoor worlds. Project page: `https://anonymous.4open.science/w/worldbridge-4F42/`.

1 INTRODUCTION

Embodied agents in real-world tasks often need to operate across indoor and outdoor spaces (Zhao et al., 2026; Gong et al., 2026). For example, a delivery robot may leave a living room, pass through an entrance, and reach the street. To support such cross-space behavior, simulated environments should model indoor spaces, transition areas, and outdoor spaces as connected parts of a unified world, with coherent geometry, semantics, and navigability.

However, existing generation paradigms struggle to produce unified and continuous scenes that span building interiors and exteriors. As illustrated in Figure 1(a), existing 3D scene generation methods typically treat indoor and outdoor environments as separate domains (Wen et al., 2025; Strader et al., 2024; Liu et al., 2026a), focusing respectively on boundary-constrained room layouts (Fang et al., 2026; Zhong et al., 2026; Xia et al., 2026) and large-scale exterior structures (Liu et al., 2026b; Yang et al., 2026b; Lee et al., 2025). Even methods covering both domains often rely on post-hoc assembly (Hong et al., 2026), which can introduce envelope inconsistency, entrance misregistration, and topological disconnection, as shown in Figure 1(b). Table 1 further details the limitations of existing methods for unified world generation.

Moving beyond this separation faces two key challenges. First, real-world 3D data for unified indoor-outdoor scenes remains scarce (Díaz Vilariño et al., 2022). Existing data rarely captures indoor structures, building exteriors, entrance transitions, and outdoor surroundings together, making it difficult to learn reliable joint generative priors. Second, indoor and outdoor regions follow different spatial priors and scales, yet their geometry and connectivity should remain consistent within a shared spatial frame. Jointly satisfying these heterogeneous constraints while preserving local scene plausibility and global connectivity is challenging.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

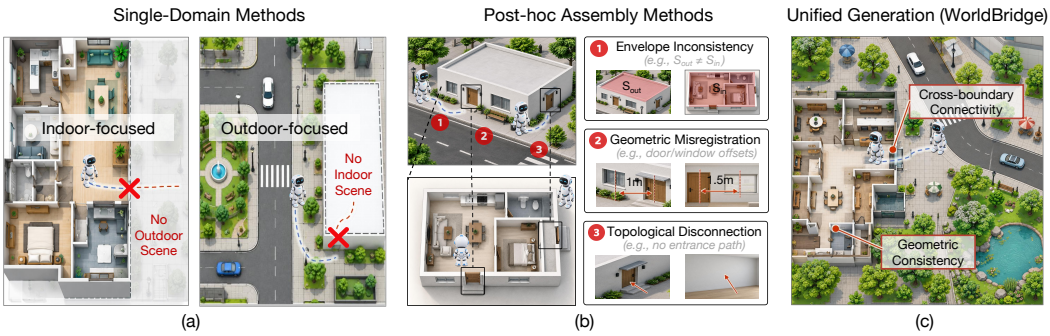

Figure 1: **Comparison of scene-generation paradigms.** (a) Single-domain methods generate either indoor or outdoor scenes, preventing cross-domain interactions. (b) Post-hoc assembly combines independently generated regions, which can introduce envelope inconsistency, geometric misregistration, and topological disconnection. (c) WorldBridge jointly generates indoor and outdoor spaces as a unified scene, improving geometric consistency and cross-boundary connectivity.

To address these challenges, we propose WorldBridge, a procedural framework for generating unified indoor–outdoor worlds. WorldBridge leverages procedural priors to avoid dependence on scarce joint data and uses a global spatial contract to reconcile heterogeneous indoor–outdoor constraints. Specifically, as shown in Figure 2, WorldBridge proceeds in three stages. First, *Spatial Contract Coordination* reconciles indoor, building-envelope, and outdoor requirements into a shared world contract, addressing their heterogeneous spatial constraints before geometry generation. Second, *Contract-Conditioned 3D World Realization* generates procedural assets and jointly realizes structure and asset placement under the shared contract, producing a structured 3D world. Finally, *Physics-Aware Dynamic Extension* attaches object and environmental dynamics to the structured scene, yielding an executable 4D world. Extensive experiments demonstrate that, compared with existing advanced approaches (Wang et al., 2026a; Engstler et al., 2026; Team et al., 2026; Zeng et al., 2026; Team et al., 2023; OpenAI, 2026), WorldBridge achieves stronger image-level appearance quality, scene-level 3D usability, and boundary-level indoor–outdoor integration, while further supporting dynamic world generation. The contributions of this paper are threefold:

- **Unified Indoor–Outdoor World Generation.** We formulate unified indoor–outdoor world generation, requiring indoor spaces, transition regions, and outdoor environments to maintain consistent geometry and navigability within a connected world. This extends scene generation beyond isolated domains and post-hoc assembly.
- **Global-Contract Procedural Framework.** We propose WorldBridge, a procedural framework that uses a global spatial contract to coordinate indoor, building-envelope, and outdoor layouts. These coordinated requirements guide the joint generation of scene structure and semantically controlled assets, producing connected 3D worlds with independently represented assets.
- **Scalable and Diverse 4D World Generation.** WorldBridge supports scalable generation of diverse indoor–outdoor worlds by composing and randomizing its procedural assets. It further equips structured 3D worlds with object and environmental dynamics, extending them into dynamic 4D worlds.

2 Related Work

2.1 3D Scene Datasets

3D scene datasets have played a crucial role in advancing embodied Artificial Intelligence (AI) (Yang et al., 2024; Zhu et al., 2025), with existing data collection and construction typically categorized into indoor (Wang et al., 2026d) and outdoor (Liu et al., 2025) domains. Indoor datasets such as ScanNet (Dai et al., 2017) and Matterport3D (Chang et al., 2017) provide rich room-level 3D scenes through real-world reconstruction, while procedural datasets such as 3D-FRONT (Fu et al., 2021) and ProcTHOR (Deitke et al., 2022) improve diversity and interactivity with large-scale synthetic layouts. These scenes remain bounded within building interiors, with limited connections

Table 1: **Comparison of 3D scene generation methods.** Existing methods mainly focus on indoor or outdoor scenes separately, while our method jointly generates indoor spaces, transition regions, and outdoor environments within a unified, navigable, and simulation-ready scene. † denotes the combined capabilities of the Infinigen series.

Previous Work	Category	Indoor Scene	Outdoor Natural	Outdoor Urban	Transition	Unified	Navigable	Object Level	Simulation Ready
Infinigen† (Raistrick et al., 2024)	Procedural	✓	✓	✗	✗	✗	✓	✓	✓
MetaUrban (Wu et al., 2025)	Procedural	✗	✗	✓	✗	✗	✓	✓	✓
SpatialGen (Fang et al., 2026)	Layout-guided	✓	✗	✗	✗	✗	✗	✗	✗
SceneWeaver (Yang et al., 2026a)	Layout-guided	✓	✗	✗	✗	✗	✓	✓	✓
MajutsuCity (Huang et al., 2026c)	Layout-guided	✗	✗	✓	✗	✗	✓	✓	✓
Scene Splatter (Zhang et al., 2025)	Distillation-based	✓	✓	✓	✗	✗	✗	✗	✗
MagicDrive3D (Gao et al., 2026)	Distillation-based	✗	✗	✓	✗	✗	✗	✗	✗
SynCity-3K (Engstler et al., 2026)	Native 3D	✓	✓	✓	✗	✗	✓	✗	✗
HY-World 2.0 (Team et al., 2026)	Native 3D	✓	✓	✓	✗	✗	✓	✗	✓
Gen3R (Huang et al., 2026a)	Native 3D	✓	✓	✓	✗	✗	✗	✗	✗
Puffin-World (Liao et al., 2026)	Native 3D	✓	✓	✓	✗	✗	✗	✗	✗
WorldGen (Wang et al., 2026a)	Hybrid	✓	✓	✓	✗	✗	✓	✗	✗
WorldBridge (Ours)	Procedural	✓	✓	✓	✓	✓	✓	✓	✓

to the external environment. Outdoor datasets primarily cover urban and natural environments, including street-level scenes (Nguyen et al., 2026; Wu et al., 2025) with roads, vehicles, and building facades, as well as landscapes (Zhang et al., 2026; Vidanapathirana et al., 2025) such as mountains, rivers, and forests. They typically lack internal building structures as well as continuous connections between indoor and outdoor spaces. In summary, this domain-separated data construction limits traditional data-driven methods' ability to construct unified scenes. Consequently, existing simulation worlds fail to provide continuous indoor-outdoor transitions, hindering the development of cross-space embodied AI tasks.

2.2 World Generation

Current 3D scene generation methods target large-scale, editable, and simulation-ready environments (Pfaff et al., 2026; Wang et al., 2026c), categorized into procedural and data-driven approaches. Procedural methods (Raistrick et al., 2024; Zhou et al., 2025; Zhang et al., 2024) synthesize scenes via explicit rules and asset composition, offering high controllability and scalability. In contrast, data-driven methods (Meng et al., 2026; Ling et al., 2026) learn generative priors from 3D data or pretrained models, spanning three categories. Layout-guided generation (Yang et al., 2026a; Pfaff et al., 2026) first predicts top-level layouts and then assembles the scene via asset retrieval or generation. They are interpretable and easy to edit, but are limited by the coverage of the asset library. Distillation-based generation (Zhang et al., 2025; Gao et al., 2026) leverage massive 2D diffusion models to lift visual priors into 3D through multi-view optimization. While enabling zero-shot generalization, they often face geometric ambiguity and multi-view artifacts. Native 3D generation (Wang et al., 2026b) directly trains generative models on 3D representations (e.g., point clouds, NeRF, or 3DGS) to learn spatial distributions. It improves geometric fidelity, but remains highly dependent on the scale and diversity of available 3D training data. Both procedural and data-driven paradigms still separate indoor and outdoor domains, lacking unified modeling for continuous, simulation-ready scenes.

3 Method

3.1 Overview

WorldBridge is a procedural framework for generating connected, navigable, and simulation-ready 4D indoor–outdoor worlds. As illustrated in Figure 2, the framework consists of three stages. First, *Spatial Contract Coordination* derives indoor, building-envelope, and outdoor contracts from the world specification and reconciles them into an agreed world contract. Second, *Contract-Conditioned 3D World Realization* jointly generates scene structure and procedural assets under this contract, producing a structured 3D world. Finally, *Physics-Aware Dynamic Extension* augments the 3D world with object and environmental dynamics to produce an executable 4D world.

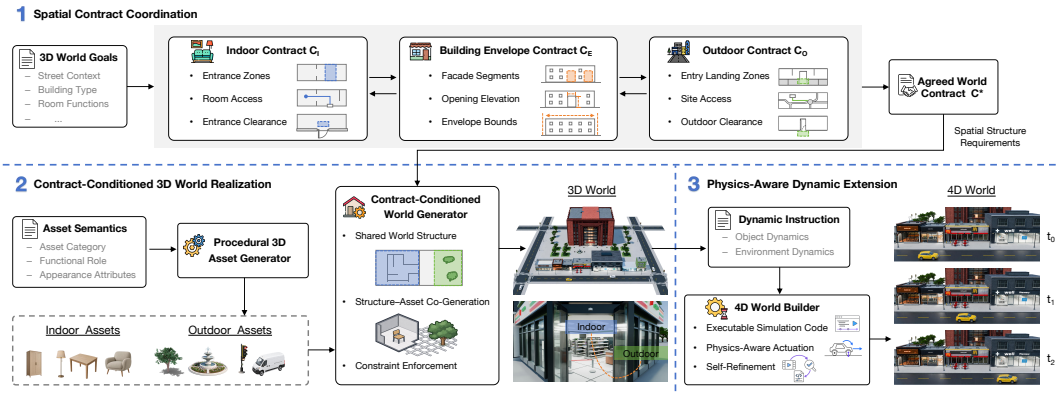

Figure 2: **Overview of WorldBridge.** (1) *Spatial Contract Coordination* establishes an agreed world contract that aligns indoor and outdoor spatial requirements for consistent world generation. (2) During the *Contract-Conditioned 3D World Realization* stage, a procedural 3D asset generator produces indoor and outdoor objects based on semantics. Guided by the spatial structure requirements, a contract-conditioned world generator then performs structure–asset co-generation to produce a coherent 3D world. (3) *Physics-Aware Dynamic Extension* uses a 4D world builder to realize the object and environment dynamics, extending the 3D world into a 4D world.

3.2 Spatial Contract Coordination

Indoor and outdoor regions follow different spatial priors. Indoor generation is organized around rooms, circulation, and object access, whereas outdoor generation is organized around parcels, roads, open space, and site access. Independent generation can therefore produce incompatible building footprints, misaligned entrances, or insufficient transition clearance. WorldBridge resolves these conflicts before geometry is instantiated.

Domain-Specific Contracts. We describe the spatial requirements of the indoor region, building envelope, and outdoor region using three complementary contracts:

$$C_I = (\mathcal{Z}_I, \mathcal{G}_I, \kappa_I), \qquad C_E = (\mathcal{B}_E, \mathcal{F}_E, \mathcal{P}_E), \qquad C_O = (\mathcal{L}_O, \mathcal{G}_O, \kappa_O). \tag{1}$$

The indoor contract C_I contains the entrance zones $\mathcal{Z}_I$, room-access graph $\mathcal{G}_I$, and required indoor-side clearance κ_I. The building-envelope contract C_E specifies the envelope bounds $\mathcal{B}_E$, facade segmentation $\mathcal{F}_E$, and entrance descriptors $\mathcal{P}_E$, including the pose, size, and elevation of each opening. The outdoor contract C_O contains the entry landing zones $\mathcal{L}_O$, site-access graph $\mathcal{G}_O$, and required outdoor-side clearance κ_O. All quantities are expressed in the same metric coordinate frame. These contracts describe complementary views of the same building boundary, but their domain-specific proposals may still be incompatible and must therefore be reconciled before geometry generation.

Contract agreement. The building-envelope contract provides the common reference for reconciling the indoor and outdoor requirements. We write this deterministic coordination process as

$$C^\star = \mathrm{Agree}(C_I, C_E, C_O) \quad \text{s.t.} \quad \begin{cases} \Delta_{\mathrm{env}}(C^\star) \le \epsilon_{\mathrm{env}}, \\ \Delta_{\mathrm{portal}}(C^\star) \le \epsilon_{\mathrm{portal}}, \\ \kappa(C^\star) \ge \kappa_{\min}, \\ \mathrm{Conn}_{\mathrm{plan}}(C^\star) = 1. \end{cases} \tag{2}$$

Here, Δ_{env} measures the discrepancy between the indoor boundary and the building envelope, and Δ_{portal} measures the disagreement between paired entrance descriptors, including position, orientation, size, and elevation. The function κ returns the planned clearance across the transition region, while $\mathrm{Conn}_{\mathrm{plan}}$ indicates connectivity in the symbolic access graph. These conditions are evaluated on contract descriptors rather than final mesh geometry. Constraint violations trigger regeneration or adjustment of the conflicting local proposal. The resulting $C^\star$ provides a shared spatial interface for the next stage.

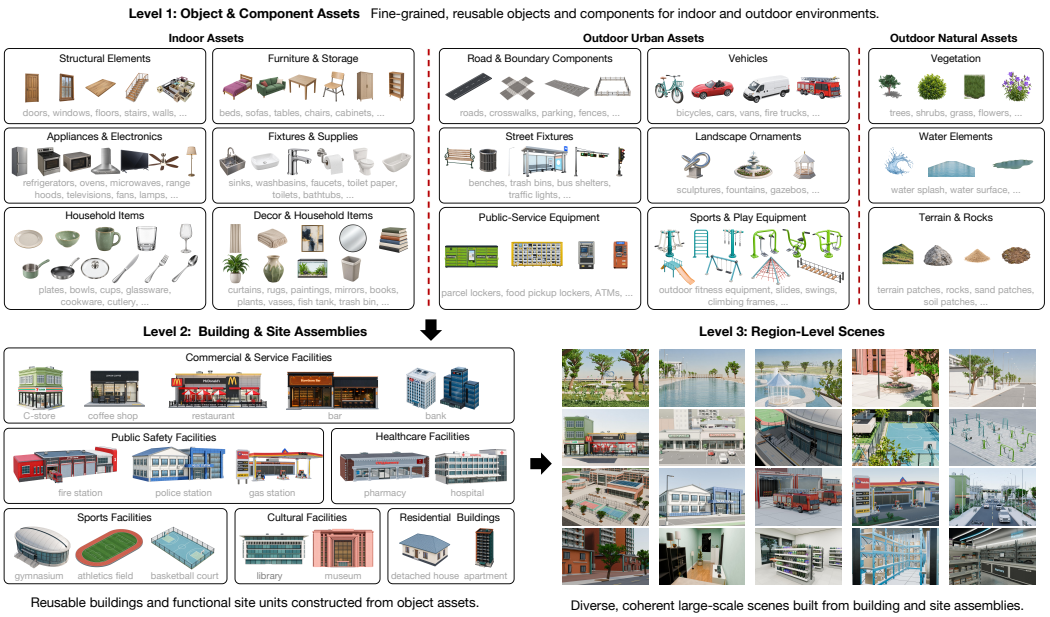

Figure 3: **Procedural assets in WorldBridge.** Level 1 comprises reusable procedural assets spanning indoor objects, outdoor urban components, and natural elements. Level 2 assembles them into functional buildings and sites across multiple categories, while Level 3 composes these assemblies into diverse and coherent region-level scenes.

3.3 Contract-Conditioned 3D World Realization

Given the agreed contract $C^\star$, this stage contains two modules: a procedural 3D asset generator and a contract-conditioned world generator. The former supplies semantically controllable building blocks, while the latter jointly realizes world structure and asset placement under the shared contract.

Procedural 3D Asset Generation. Each asset is described by semantic attributes $s_i = (c_i, r_i, \alpha_i)$, corresponding to its category, functional role, and appearance attributes. The asset generator maps s_i and a randomization variable ξ_i to a parameterized asset:

$$a_i = \mathcal{G}_{\text{asset}}(s_i; \xi_i), \qquad \mathcal{A} = \{a_i\}_{i=1}^{N}. \tag{3}$$

The random variable ξ_i controls geometry, materials, and object variants while preserving the semantics encoded by s_i. As shown in Figure 3, reusable objects and components are assembled into functional buildings and sites and then into region-level scenes. This hierarchical organization allows the same semantic specification to produce diverse but functionally consistent assets at multiple spatial scales.

Contract-Conditioned World Generation. The world generator takes $C^\star$ and $\mathcal{A}$ as input. It first instantiates a shared structure S, including the building envelope, indoor partitions, entrance transitions, outdoor parcels, roads, and terrain. It then places each asset using a transform T_i constrained to a feasible region Ω_i:

$$S = \mathcal{G}_{\text{str}}(C^\star; \xi_S), \quad T_i \in \Omega_i(S, C^\star), \quad W_0 = \mathcal{G}_{\text{world}}\big(S, \{(a_i, T_i)\}_{i=1}^{N}\big). \tag{4}$$

Here, ξ_S controls structural variations not fixed by the contract, while Ω_i encodes the compatible room or site, support surface, functional access, and reserved clearance for asset a_i. Structure and placement are updated jointly whenever an asset affects circulation or transition clearance.

After assembly, the generator performs a bounded number of local refinement steps using lightweight placement residuals:

$$\mathbf{r}(W_k, C^\star) = \big(r_{\text{pen}}, r_{\text{sup}}, r_{\text{bound}}, r_{\text{clear}}\big), \tag{5}$$

$$W_{k+1} = \mathcal{R}\big(W_k, \mathbf{r}(W_k, C^\star)\big), \quad k = 0, \ldots, K-1, \tag{6}$$

$$W^{3D} = W_K. \tag{7}$$

The residuals are computed from proxy collision geometry, declared support relations, assigned boundaries, and contract clearance. Each refinement updates only the responsible structural or placement parameters. The fixed budget K makes this a best-effort generation procedure rather than a hard acceptance test; full-mesh collision, NavMesh construction, and end-to-end connectivity are evaluated independently after generation. The resulting world is stored as an object-level scene graph $W^{3D} = (\mathcal{V}, \mathcal{E})$, where nodes contain asset geometry, semantics, transforms, and physical metadata, and edges encode containment, support, and functional relations. This representation supports editing, simulator export, and the dynamic extension described next.

3.4 Physics-Aware Dynamic Extension

Given the structured world W^{3D}, the final stage attaches time-varying behavior to existing object nodes and environmental regions. Because dynamics are grounded in the scene graph, they preserve the geometry, semantics, and cross-boundary connectivity established by the previous stages.

Object and Environmental Dynamics. Dynamic instructions are divided into object dynamics D_{obj} and environmental dynamics D_{env}. For a dynamic object i, let q_i^t denote its state and u_i^t its actuation or route command. Let e^t denote the state of an environmental process and ω^t its external input. Their evolution is written as

$$q_i^{t+1} = \mathcal{F}_i(q_i^t, u_i^t; \theta_i), \quad e^{t+1} = \mathcal{H}(e^t, \omega^t; \phi), \quad W^{4D} = \left\{ \left(W^{3D}, Q^t, e^t \right) \right\}_{t=0}^{T}. \tag{8}$$

where $Q^t = \{q_i^t\}$ collects all object states, θ_i contains object-specific physical and behavioral parameters, and ϕ controls environmental effects. This formulation covers object motion such as vehicles following site-access routes, as well as environmental processes such as water emission, wind-driven vegetation, and falling leaves.

Executable 4D World Building. The builder assigns collision geometry and physical properties, binds actuators and trajectories, and generates executable simulation logic. A short preview simulation produces dynamic residuals

$$\mathbf{r}_{4D}^t = \left(r_{\mathrm{pen}}^t, r_{\mathrm{bound}}^t, r_{\mathrm{route}}^t \right), \qquad t = 0, \ldots, T, \tag{9}$$

which summarize penetration, permitted-region, and route-deviation errors. A bounded local refinement adjusts the initial state, trajectory, or dynamic parameters associated with the affected behavior. As in 3D realization, this stage improves physical executability without asserting perfect validity for every generated sequence. The resulting W^{4D} retains the spatial contract and structured representation of W^{3D} while supporting executable object behaviors and environmental dynamics.

4 Experiments

In this section, we first describe the experimental setup (§ 4.1), then present the main results on generation quality, indoor–outdoor connectivity, and dynamic world generation (§ 4.2), followed by ablation studies of the spatial contracts (§ 4.3) and extended results on cross-space tasks (§ 4.4).

4.1 Experiment Settings

Evaluation protocol. For each indoor and outdoor domain, we use 25 scene specifications and four random seeds, yielding 100 planned generations per domain. Indoor–outdoor evaluation uses a separate set of 25 paired specifications with the same seed schedule. Each scene is evaluated using eight anchor views and a 50-frame walkthrough. We group baselines into single-domain and indoor–outdoor methods and evaluate each method only on measurements supported by its native output representation. In particular, object-level collision and support metrics require explicit instance and support metadata.

Metrics. We evaluate the generated worlds at three levels: rendered appearance, 3D usability, and indoor–outdoor integration. (1) At the image level, we measure perceptual quality, layout and prompt agreement, and diversity across generation seeds. (2) At the scene level, we evaluate object-placement validity, geometric consistency, and NavMesh-based navigability. (3) At the boundary level, we assess indoor–outdoor correspondence, entrance alignment and passability, and end-to-end cross-boundary reachability. More details are provided in Appendix A.

Table 2: **Quantitative comparison of image-level appearance quality.** The evaluation covers perceptual quality, image-based scene fidelity, and visual diversity. Single-domain methods are evaluated on their supported domain. For indoor–outdoor methods, a unified scene is jointly generated, while metrics are computed per region and averaged for reporting. Valid Gen. denotes the number of successful scene generations out of 100 runs. Green, yellow, and blue indicate the best, second-best, and third-best results, respectively.

Method	Scene Coverage	Valid Gen.	Perceptual Quality		Scene Quality			Diversity	
			Q-Align ↑	CLIP-IQA+ ↑	Layout Plaus. ↑	Prompt Align. ↑	3D Consist. ↑	Appearance Div. ↑	Layout Div. ↑
Single-domain methods									
Infinigen Indoor (Raistrick et al., 2024)	Indoor	89	3.72	0.520	55.7	52.0	82.1	0.767	0.966
SpatialGen (Fang et al., 2026)	Indoor	100	3.29	0.538	57.4	81.1	83.2	0.348	0.783
SceneWeaver (Yang et al., 2026a)	Indoor	47	3.59	0.574	60.8	68.0	38.3	0.640	0.888
MetaUrban (Wu et al., 2025)	Outdoor	100	3.85	0.629	60.0	75.6	68.1	0.267	0.775
MajutsuCity (Huang et al., 2026c)	Outdoor	86	3.94	0.565	52.1	37.3	61.6	0.673	0.946
Indoor–outdoor methods									
WorldGen (Wang et al., 2026a)	Indoor + Outdoor	98	2.69	0.397	65.0	62.6	71.0	0.672	0.940
SynCity-3K (Engstler et al., 2026)	Indoor + Outdoor	97	3.26	0.527	53.6	75.4	88.5	0.398	0.868
HY-World 2.0 (Team et al., 2026)	Indoor + Outdoor	81	1.67	0.311	89.5	26.8	79.2	0.638	0.921
GLM-5.3 Flash (Zeng et al., 2026)	Indoor + Outdoor	44	3.29	0.548	53.5	74.3	69.5	0.576	0.949
GLM-5.3 (Zeng et al., 2026)	Indoor + Outdoor	42	3.35	0.555	63.9	71.9	62.0	0.550	0.938
Gemini 3.1 Pro (Team et al., 2023)	Indoor + Outdoor	76	3.43	0.537	52.9	71.3	59.7	0.485	0.915
GPT-6 Astra `(high)` (OpenAI, 2026)	Indoor + Outdoor	91	3.89	0.616	73.7	75.5	70.8	0.514	0.869
GPT-6 Astra `(xhigh)` (OpenAI, 2026)	Indoor + Outdoor	82	3.96	0.620	68.5	75.8	81.4	0.560	0.874
WorldBridge (Ours)	Indoor + Outdoor	100	3.78	0.617	82.6	84.2	88.4	0.724	0.975



Figure 4: **Qualitative comparison of indoor–outdoor scene generation.** The top panel shows domain-specific methods for either indoor or outdoor scenes, while the bottom panel compares methods covering both domains. Black boxes and arrows highlight representative artifacts. WorldBridge unifies indoor–outdoor generation, producing detailed, natural, and geometrically coherent scenes.

4.2 Main Results

Image-Level Appearance Quality. WorldBridge provides reliable generation while maintaining strong visual quality and diversity. As shown in Table 2, all 100 planned runs produce valid scenes. WorldBridge ranks first in prompt alignment (84.2) and layout diversity (0.975), while remaining competitive in Q-Align (3.78), CLIP-IQA+ (0.617), layout plausibility (82.6), and appearance diversity (0.724). Together, these results show that WorldBridge generates diverse scenes that faithfully satisfy the input specifications while maintaining competitive perceptual quality. The qualitative comparison in Figure 4 further shows that WorldBridge avoids common failure modes such as invalid object placement, incomplete or distorted geometry, and oversimplified textures, producing coherent indoor and outdoor scenes.

Table 3: **Quantitative comparison of scene-level 3D usability.** Generated scenes are evaluated in terms of structural validity and navigability. Green, yellow, and blue indicate the best, second-best, and third-best results, respectively.

Method	Category		Structural Validity				Navigability		
	Scene Coverage	Track	Collision Rate ↓	Floating Rate ↓	OOB Rate ↓	Support Validity ↑	Navigable Ratio ↑	Connected Ratio ↑	NavMesh Success Rate ↑
Single-domain methods									
Infinigen Indoor (Raistrick et al., 2024)	Indoor	Structured	14.4	14.8	12.8	85.2	59.7	83.8	84.0
SpatialGen (Fang et al., 2026)	Indoor	Surface-only	–	–	–	–	12.5	32.2	46.0
SceneWeaver (Yang et al., 2026a)	Indoor	Structured	50.1	42.0	38.0	43.2	46.5	47.6	50.0
MetaUrban (Wu et al., 2025)	Outdoor	Structured	39.3	9.4	14.1	48.3	95.2	93.7	100.0
MajutsuCity (Huang et al., 2026c)	Outdoor	Structured	24.6	16.9	18.6	78.9	61.7	72.9	85.7
Indoor–outdoor methods									
WorldGen (Wang et al., 2026a)	Indoor + Outdoor	Surface-only	–	–	–	–	24.2	99.0	100.0
SynCity-3K (Engstler et al., 2026)	Indoor + Outdoor	Surface-only	–	–	–	–	2.3	5.1	11.5
HY-World 2.0 (Team et al., 2026)	Indoor + Outdoor	Surface-only	–	–	–	–	5.7	43.2	50.5
Gemini 3.1 Pro (Team et al., 2023)	Indoor + Outdoor	Structured	34.0	35.5	32.6	64.5	66.1	54.9	72.5
GLM-5.3 Flash (Zeng et al., 2026)	Indoor + Outdoor	Structured	68.2	76.9	67.6	23.1	32.8	28.8	35.0
GLM-5.3 (Zeng et al., 2026)	Indoor + Outdoor	Structured	76.3	81.7	74.8	18.3	22.9	24.2	26.0
GPT-6 Astra (`high`) (OpenAI, 2026)	Indoor + Outdoor	Structured	11.1	26.6	13.7	73.4	60.4	62.0	68.0
GPT-6 Astra (`xhigh`) (OpenAI, 2026)	Indoor + Outdoor	Structured	10.2	24.8	11.7	75.2	56.6	57.5	67.5
WorldBridge (Ours)	Indoor + Outdoor	Structured	8.7	5.8	6.5	92.6	96.7	97.8	98.0

Table 4: **Quantitative comparison of boundary-level indoor–outdoor integration.** Generated scenes are evaluated in terms of indoor–outdoor correspondence, transition quality, and cross-domain connectivity. Green, yellow, and blue indicate the best, second-best, and third-best results, respectively. *HoloWorld results are reported from the original paper.

Method	Indoor–Outdoor Correspondence				Transition Quality			Cross-domain Connectivity	
	Functional AQS ↑	Visual AQS ↑	Spatial AQS ↑	Shape IoU ↑	Entrance Align. ↑	Entrance Pass. ↑	Transition Collision ↓	I–O Connect. Rate ↑	Cross-bound. Reach. ↑
HoloWorld* (Huang et al., 2026b)	7.47	8.21	8.77	0.997	–	–	–	–	–
HY-World 2.0 (Team et al., 2026)	3.21	3.57	3.85	0.420	44.0	29.0	27.0	13.0	9.0
WorldGen (Wang et al., 2026a)	6.77	6.95	6.05	0.640	67.0	52.0	16.5	37.0	30.0
MajutsuCity (Huang et al., 2026c)	5.90	6.35	5.30	0.600	64.0	55.0	13.5	48.0	40.0
SynCity-3K (Engstler et al., 2026)	6.13	6.18	5.25	0.560	52.0	28.0	24.0	12.0	8.0
GLM-5.3 Flash (Zeng et al., 2026)	4.64	4.33	4.65	0.690	78.0	73.0	10.5	64.0	55.0
GLM-5.3 (Zeng et al., 2026)	5.13	5.17	4.90	0.670	76.0	70.0	12.0	60.0	51.0
Gemini 3.1 Pro (Team et al., 2023)	4.33	4.05	4.75	0.770	86.0	82.0	7.2	76.0	68.0
GPT-6 Astra (`high`) (OpenAI, 2026)	3.52	4.01	4.20	0.780	87.0	79.0	8.0	68.0	59.0
GPT-6 Astra (`xhigh`) (OpenAI, 2026)	2.97	3.49	3.55	0.790	86.0	79.0	8.3	70.0	62.0
WorldBridge (Ours)	8.10	8.45	9.05	0.992	98.0	96.0	1.8	97.0	94.0



Figure 5: **Qualitative comparison of indoor–outdoor connectivity.** Exterior (top) and interior (bottom) views are shown for each method, with blue arrows indicating navigation paths.

Scene-Level 3D Usability. WorldBridge consistently produces structurally valid and navigable 3D scenes. As shown in Table 3, it records the lowest collision (8.7%), floating (5.8%), and out-of-bounds (6.5%) rates among structured methods, while achieving the highest support validity (92.6%) and navigable area ratio (96.7%). The resulting free space is also well connected, with a connected area ratio of 97.8% and a NavMesh success rate of 98.0%. Together with the 3D consistency score of 88.4 in Table 2, these results show that WorldBridge preserves coherent geometry across views and produces simulation-ready spaces rather than merely plausible renderings.

Boundary-Level Indoor–Outdoor Integration. WorldBridge achieves coherent indoor–outdoor integration from regional correspondence to global reachability. As shown in Table 4, at the region level, it obtains functional, visual, and spatial AQS scores of 8.10, 8.45, and 9.05, respectively, together with a Shape IoU of 0.992. At the entrance level, WorldBridge reaches 98% alignment and 96% passability, while reducing transition collision to 1.8%. At the world level, these properties yield 97% indoor–outdoor connectivity and 94% end-to-end cross-boundary reachability. The qualitative comparison in Figure 5 further confirms this progression. Although competing methods may form indoor–outdoor paths, obstacles, geometric misalignment, or insufficient clearance around the



Figure 6: Visualizations of generated 4D worlds with object and environmental dynamics.

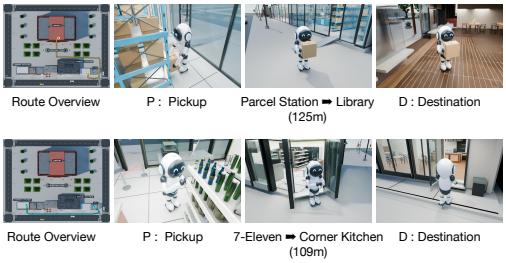

Figure 7: Executions of cross-space robot tasks spanning indoor and outdoor environments.

entrance can impede traversal. WorldBridge instead maintains aligned entrance geometry, sufficient free space, and a clear path across the boundary.

Dynamic World Generation. WorldBridge extends static 3D scenes into executable 4D worlds with both environmental and object dynamics. Figure 6 visualizes three representative sequences: water emitted by a fountain, foliage swaying and leaves falling under wind, and vehicles entering a gas station. These examples span continuous environmental effects, natural motion, and object-level behaviors, showing that dynamic processes can be instantiated within the generated world while preserving its spatial structure.

4.3 Ablation Study

To isolate the role of each spatial contract, we remove C_I, C_O, or C_E while retaining the remaining contracts and using identical generation seeds. Table 5 presents the ablation results. Without C_E, entrance alignment drops by 31 percentage points, from 98.0% to 67.0%, while passability and cross-boundary reachability decrease to 70.0% and 54.0%, respectively. This confirms that the envelope constraints are critical for coordinating building boundaries and entrance openings. Removing C_I or C_O has a smaller effect on alignment, but reduces entrance passability by 18 and 15 percentage points and cross-boundary reachability by 29 and 32 percentage points, respectively. Thus, the indoor and outdoor contracts primarily ensure access and clearance on the two sides of the boundary. The full model with $C^\star$ performs best across all three metrics, confirming that the contracts play distinct yet complementary roles in cross-boundary integration.

Table 5: Ablation study of spatial contracts.

Method	Entrance Align.↑	Entrance Pass.↑	Cross-Bound. Reach.↑
w/o C_I	94.0	78.0	65.0
w/o C_O	93.0	81.0	62.0
w/o C_E	67.0	70.0	54.0
w/ $C^\star$	**98.0**	**96.0**	**94.0**

4.4 Cross-Space Embodied Task Execution

To assess downstream utility, we execute long-horizon pickup-and-delivery tasks in the generated worlds. Figure 7 shows two representative episodes: a 125 m route from a parcel station to a library and a 109 m route from a convenience store to a corner kitchen. Each episode requires object pickup, outdoor navigation, entrance traversal, and delivery to an indoor destination. These executions provide qualitative evidence that WorldBridge supports continuous navigation and interaction across spatial domains. They validate the generated environment rather than compare robot policies.

5 Conclusion

We presented WorldBridge, a procedural framework for generating unified and connected indoor–outdoor worlds. By coordinating indoor layouts, building envelopes, and outdoor environments through a shared spatial contract, WorldBridge jointly constructs structured 3D scenes and extends them with executable object and environmental dynamics. Extensive experiments demonstrate that WorldBridge produces diverse and visually compelling worlds with strong structural validity, navigability, and cross-boundary connectivity. The resulting environments further support continuous embodied navigation and interaction across indoor and outdoor spaces.

AI USE STATEMENT

We used generative AI tools primarily to assist with language polishing and LLM-assisted code debugging. All AI-assisted text and code were reviewed and revised by the authors, and all experimental outputs were independently verified. Generative models used as scene-generation baselines or automated evaluators are identified in the experimental setup and appendix and served only in their stated experimental roles. We take responsibility for the final content, including the text, claims, results, and code.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we release the implementation and evaluation materials at `https://anonymous.4open.science/w/worldbridge-4F42/`. The project includes (i) the source code for WorldBridge, (ii) the code, configurations, and evaluation scripts used to run the compared methods and quantitative experiments, and (iii) visual comparisons of the generated results. The appendix complements these artifacts by specifying the scene inputs and random seeds, rendering protocol, metric definitions, and aggregation rules. Together, the released artifacts and documented protocol enable reproduction of the generation pipeline, quantitative evaluations, and qualitative comparisons reported in this paper.

REFERENCES

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.

Lucía Díaz Vilariño, Ha Tran, Ernesto Frías Nores, Jesús Balado Frías, Kourosh Khoshelham, et al. 3d mapping of indoor and outdoor environments using apple smart devices. *ISPRS-International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences*, 2022.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 27713–27724, 2025.

Paul Engstler, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Syncity 3000 bootstrapping scene-scale 3d diffusion. In *European Conference on Computer Vision*, pp. 281–299. Springer, 2026.

Chuan Fang, Heng Li, Yixun Liang, Jia Zheng, Yongsen Mao, Yuan Liu, Rui Tang, Zihan Zhou, and Ping Tan. Spatialgen: Layout-guided 3d indoor scene generation. In *2026 International Conference on 3D Vision (3DV)*, pp. 1392–1402. IEEE, 2026.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10933–10942, 2021.

Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5724–5733. IEEE, 2026.

Ruiyan Gong, Meisheng Zhang, Yuxiang Zhao, Mingchao Sun, Yanfen Shen, Zedong Chu, Zhining Gu, Wei Guo, Xiaolong Cheng, Qiming Li, et al. Poinav: Benchmarking and enhancing final-meters arrival in real-world vision-language navigation. *arXiv preprint arXiv:2605.28237*, 2026.

Yining Hong, Rui Sun, Bingxuan Li, Xingcheng Yao, Maxine Wu, Alexander Chien, Da Yin, Ying Nian Wu, Zhecan Wang, and Kai-Wei Chang. Embodied web agents: Bridging physical-digital realms for integrated agent intelligence. *Advances in Neural Information Processing Systems*, 38, 2026.

Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3r: 3d scene generation meets feed-forward reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25358–25369, 2026a.

Xiaobin Huang, Zilong Huang, Yang Luo, Hongchao Fan, Yiping Chen, and Ting Han. To see a world in a living context: Unified indoor-outdoor urban world generation. *arXiv preprint arXiv:2608.05879*, 2026b.

Zilong Huang, Jun He, Xiaobin Huang, Ziyi Xiong, Yang Luo, Junyan Ye, Weijia Li, Yiping Chen, and Ting Han. Majutsucity: Language-driven aesthetic-adaptive city generation with controllable 3d assets and layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31608–31618, 2026c.

Han-Hung Lee, Qinghong Han, and Angel X Chang. Nuiscene: Exploring efficient generation of unbounded outdoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26509–26518, 2025.

Kang Liao, Yihang Luo, Xiao-Ming Wu, Linyi Jin, Size Wu, Chunyu Lin, Yao Zhao, Fei Wang, Wei Li, and Chen Change Loy. Puffin-world: Scaling a unified multimodal model with native 3d world states. *arXiv preprint arXiv:2609.04196*, 2026.

Lu Ling, Yunhao Ge, Yichen Sheng, and Aniket Bera. I-scene: 3d instance models are implicit generalizable spatial learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26974–26983, 2026.

Jiahang Liu, Tianyu Xu, Jiawei Chen, Lu Yue, Jiazhao Zhang, Zhiyong Wang, Minghan Li, Qisheng Zhao, Anqi Li, Qi Su, et al. Span-nav: Generalized spatial awareness for versatile vision-language navigation. *arXiv preprint arXiv:2603.09163*, 2026a.

Xinhao Liu, Jiaqi Li, Youming Deng, Ruxin Chen, Yingjia Zhang, Yifei Ma, Li Guo, Yiming Li, Jing Zhang, and Chen Feng. Wanderland: Geometrically grounded simulation for open-world embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1041–1052, 2026b.

Yuheng Liu, Xinke Li, Yuning Zhang, Lu Qi, Xin Li, Wenping Wang, Chongshou Li, Xueting Li, and Ming-Hsuan Yang. Controllable 3d outdoor scene generation via scene graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28052–28062, 2025.

Yanxu Meng, Haoning Wu, Ya Zhang, and Weidi Xie. Scenegen: Single-image 3d scene generation in one feedforward pass. In *2026 International Conference on 3D Vision (3DV)*, pp. 543–553. IEEE, 2026.

Duc Nguyen, Yan-Ling Lai, Qilin Zhang, Prabin Gyawali, Fan Wang, Benedikt Schwab, Olaf Wysocki, and Thomas H Kolbe. Truecity: Real and simulated urban data for cross-domain 3d scene understanding. In *2026 International Conference on 3D Vision (3DV)*, pp. 1905–1915. IEEE, 2026.

OpenAI. GPT-6 Astra System Card. `https://deploymentsafety.openai.com/gpt-6-astra`, September 2026. Published September 3, 2026.

Nicholas Pfaff, Thomas Cohn, Sergey Zakharov, Rick Cory, and Russ Tedrake. Scenesmith: Agentic generation of simulation-ready indoor scenes. *arXiv preprint arXiv:2602.09153*, 2026.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, et al. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21783–21794, 2024.

Jared Strader, Nathan Hughes, William Chen, Alberto Speranzon, and Luca Carlone. Indoor and outdoor 3d scene graph generation via language-enabled spatial ontologies. *IEEE Robotics and Automation Letters*, 9(6):4886–4893, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

HY Team, Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, Yuning Gong, Yang Liu, Bo Yuan, et al. Hy-world 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds. *arXiv e-prints*, pp. arXiv–2604, 2026.

Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

Kavisha Vidanapathirana, Joshua Knights, Stephen Hausler, Mark Cox, Milad Ramezani, Jason Jooste, Ethan Griffiths, Shaheer Mohamed, Sridha Sridharan, Clinton Fookes, et al. Wildscenes: A benchmark for 2d and 3d semantic segmentation in large-scale natural environments. *The International Journal of Robotics Research*, 44(4):532–549, 2025.

Dilin Wang, Hyunyoung Jung, Tom Monnier, Kihyuk Sohn, Chuhang Zou, Xiaoyu Xiang, Yu-Ying Yeh, Di Liu, Zixuan Huang, Thu Nguyen-Phuoc, et al. Worldgen: From text to traversable and interactive 3d worlds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27124–27135, 2026a.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

Pengfei Wang, Liyi Chen, Zhiyuan Ma, Yanjun Guo, Guowen Zhang, and Lei Zhang. One2scene: Geometric consistent explorable 3d scene generation from a single image. *arXiv preprint arXiv:2602.19766*, 2026b.

Puyi Wang, Yuhao Wang, Linjie Li, Zhengyuan Yang, Kevin Qinghong Lin, Yangguang Li, and Yu Cheng. Scenecode: Executable world programs for editable indoor scenes with articulated objects. *arXiv preprint arXiv:2605.19587*, 2026c.

Xinyi Wang, Xun Yang, Yanlong Xu, Yuchen Wu, Zhen Li, and Na Zhao. Affordbot: 3d fine-grained embodied reasoning via multimodal large language models. *Advances in Neural Information Processing Systems*, 38:140367–140387, 2026d.

Beichen Wen, Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, and Ziwei Liu. 3d scene generation: A survey. *arXiv preprint arXiv:2505.05474*, 2025.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

Wayne Wu, Honglin He, Jack He, Yiran Wang, Chenda Duan, Zhizheng Liu, Quanyi Li, and Bolei Zhou. Metaurban: An embodied ai simulation platform for urban micromobility. International Conference on Learning Representations (ICLR) 2025, 2025.

Hongchi Xia, Xuan Li, Zhaoshuo Li, Qianli Ma, Jiashu Xu, Ming-Yu Liu, Yin Cui, Tsung-Yi Lin, Wei-Chiu Ma, Shenlong Wang, et al. Sage: Scalable agentic 3d scene generation for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22358–22368, 2026.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

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

Yandan Yang, Baoxiong Jia, Shujie Zhang, and Siyuan Huang. Sceneweaver: All-in-one 3d scene synthesis with an extensible and self-reflective agent. *Advances in neural information processing systems*, 38:140319–140351, 2026a.

Yu Yang, Alan Liang, Jianbiao Mei, Yukai Ma, Yong Liu, and Gim Hee Lee. X-scene: Large-scale driving scene generation with high fidelity and flexible controllability. *Advances in Neural Information Processing Systems*, 38:104415–104451, 2026b.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Shengjun Zhang, Jinzhao Li, Xin Fei, Hao Liu, and Yueqi Duan. Scene splatter: Momentum 3d scene generation from single image with video diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6089–6098. IEEE, 2025.

Shougao Zhang, Mengqi Zhou, Yuxi Wang, Chuanchen Luo, Rongyu Wang, Yiwei Li, Zhaoxiang Zhang, and Junran Peng. Cityx: Controllable procedural content generation for unbounded 3d cities. *arXiv preprint arXiv:2407.17572*, 2024.

Yi Zhang, Yunshuang Wang, Zeyu Zhang, and Hao Tang. Code2worlds: Empowering coding llms for 4d world generation. *International Conference on Machine Learning*, 2026.

Yuxiang Zhao, Yirong Yang, Yanqing Zhu, Yanfen Shen, Chiyu Wang, Zhining Gu, Pei Shi, Wei Guo, and Mu Xu. Bridging the indoor-outdoor gap: Vision-centric instruction-guided embodied navigation for the last meters. *arXiv preprint arXiv:2602.06427*, 2026.

Weipeng Zhong, Peizhou Cao, Yichen Jin, Luo Li, Wenzhe Cai, Jingli Lin, Hanqing Wang, Zhaoyang Lyu, Tai Wang, Xudong Xu, et al. Internscenes: A large-scale simulatable indoor scene dataset with realistic layouts. *Advances in Neural Information Processing Systems*, 38, 2026.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5122–5130. IEEE, 2017.

Mengqi Zhou, Yuxi Wang, Jun Hou, Shougao Zhang, Yiwei Li, Chuanchen Luo, Junran Peng, and Zhaoxiang Zhang. Scenex: Procedural controllable large-scale scene generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10806–10814, 2025.

Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, et al. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8120–8132, 2025.

APPENDIX

A EVALUATION PROTOCOL AND METRIC DEFINITIONS

Our evaluation characterizes both the appearance of generated worlds and their usability as connected three-dimensional environments. We first assess rendered observations for perceptual quality, semantic agreement, and variation across generations. We then evaluate geometric constraints and navigation queries on the declared scene representation. This distinction determines which measurements are supported by each output format and how their results should be interpreted.

A.1 EVALUATION INSTANCES AND OBSERVATIONS

Let p denote an input specification and $s = (p, k)$ its generated output under seed k. The evaluation manifest contains $P = 25$ specifications per indoor or outdoor domain and $K = 4$ seeds, $k \in \{0, 1, 2, 3\}$. Indoor–outdoor evaluation uses a separate set of 25 building–interior specifications with the same seed schedule. Metrics are aggregated within their respective evaluation populations.

For each scene, the reference observation protocol specifies $V = 8$ RGB views at 1280×720 resolution and a 50-frame walkthrough at 512×512 resolution and 10 frames per second. The anchor views correspond to trajectory frames $1, 8, 15, 22, 29, 36, 43, 50$. The reference camera height is $1.55\,\mathrm{m}$ and its horizontal field of view is $70°$; method-specific native rendering and scale conversions follow the recorded evaluation configuration. We write $I_{s,v}$ for anchor view v. Corresponding view indices are retained across seeds when computing diversity. Building–interior assessment uses four exterior and four interior views, supplemented by a fixed-scale plan overlay for spatial assessment.

A.2 ASSESSMENT FROM RENDERED OBSERVATIONS

We measure perceptual quality with the image-quality head of Q-Align (Wu et al., 2023) and the learned-prompt RN50 variant of CLIP-IQA+ (Wang et al., 2023), using fixed pretrained checkpoints. Q-Align assigns probabilities $p_{s,v,\ell}$ to five ordered quality levels, whereas CLIP-IQA+ compares positive and negative quality logits $\lambda^{+}_{s,v}$ and $\lambda^{-}_{s,v}$, including the model's logit scale. Their scene scores are

$$Q_s = \frac{1}{V} \sum_{v=1}^{V} \sum_{\ell=1}^{5} \ell \, p_{s,v,\ell}, \qquad Q_s \in [1, 5], \tag{10}$$

$$C_s = \frac{1}{V} \sum_{v=1}^{V} \frac{\exp(\lambda^{+}_{s,v})}{\exp(\lambda^{+}_{s,v}) + \exp(\lambda^{-}_{s,v})}, \qquad C_s \in [0, 1]. \tag{11}$$

These scores summarize perceived image quality; semantic requirements are evaluated separately using a fixed rubric and a specification-level fact checklist.

The layout rubric covers boundary and collision plausibility, support and placement, scale and density, and functional access and circulation. Let $L_{s,r,d} \in \{1, \ldots, 5\}$ be the score for dimension d under scoring profile r. For the F_p facts extracted from specification p, let $y_{s,r,f} = 1$ when fact f is judged visibly satisfied, and zero for either a negative or a not-visible judgment. Layout Plausibility and Prompt Alignment are computed as

$$\mathrm{LP}_s = 25 \left(\frac{1}{4R} \sum_{r=1}^{R} \sum_{d=1}^{4} L_{s,r,d} - 1 \right), \tag{12}$$

$$\mathrm{PA}_s = \frac{100}{R F_p} \sum_{r=1}^{R} \sum_{f=1}^{F_p} y_{s,r,f}. \tag{13}$$

Both lie in $[0, 100]$. The checklist includes object presence, counts, attributes, and spatial relations, and is fixed before inspecting generated outputs. Consequently, Prompt Alignment measures requirements verified from the available views and also depends on observation coverage.

The current annotation source uses $R = 3$ synthetic scoring profiles derived from a shared model judgment. For the layout rubric, one profile retains the neutral scores and the other two decrease or

increase one registered dimension by one point, clipped to $[1, 5]$. Clear positive and negative facts retain their labels across profiles; ambiguous facts receive negative, not-visible, and positive labels, respectively. These profiles are dependent proxy annotations. Their average defines the reported score and does not constitute a human evaluation.

For building–interior pairs, we adopt the functional, visual, and spatial correspondence dimensions of HoloWorld's Absolute Quantitative Scoring (AQS) (Huang et al., 2026b). Functional correspondence asks whether the interior supports the building's intended use. Visual correspondence concerns compatible materials, architectural style, and appearance across the boundary. Spatial correspondence concerns the interior's organization relative to the building footprint, scale, orientation, and entrance arrangement. With a fixed rubric on a 1–10 scale, the score for dimension d is

$$\mathrm{AQS}_{s,d} = \frac{1}{J}\sum_{j=1}^{J} a_{s,j,d}, \qquad a_{s,j,d} \in [1, 10], \quad d \in \{\mathrm{fun}, \mathrm{vis}, \mathrm{spa}\}. \tag{14}$$

The local evaluator is Qwen3-VL-8B-Instruct. We average $J = 3$ passes with seeds 102, 203, and 304, temperature 0.2, and top-p 0.9. These results are identified as AQS-VLM. The evaluator receives the paired views and, for the spatial dimension, the plan overlay described above. This measures correspondence between the two regions, complementing their individual perceptual quality scores.

A.3 Variation across generation seeds

Diversity is evaluated conditionally on the same input specification. Let $\mathcal{P}_p$ contain the unordered pairs of successful seeds for specification p. For a distance function d, the conditional score is

$$D_p^{\mathrm{valid}}(d) = \frac{1}{V|\mathcal{P}_p|}\sum_{(a,b)\in\mathcal{P}_p}\sum_{v=1}^{V} d(I_{(p,a),v}, I_{(p,b),v}). \tag{15}$$

We use LPIPS (Zhang et al., 2018) with AlexNet features and version 0.1 calibration to measure Appearance Diversity. Its raw distance is retained; we do not impose a unit upper bound.

Layout Diversity applies the same averaging to a semantic-mask distance. We obtain masks with a fixed SegFormer-B2 model (Xie et al., 2021) trained on ADE20K (Zhou et al., 2017). For two corresponding views, let M_c and M'_c be their pixel sets for class c, and let $\mathcal{C}$ contain the registered domain classes appearing in either mask. We use

$$d_{\mathrm{sem}}(I, I') = 1 - \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\frac{|M_c \cap M'_c|}{|M_c \cup M'_c|}. \tag{16}$$

Classes absent from both views are excluded, and an empty $\mathcal{C}$ gives distance zero. This score captures changes in projected semantic arrangement, with sensitivity to the prescribed views and segmentation model. Both diversity scores are interpreted jointly with quality and prompt satisfaction, since variation alone does not establish valid generation. The adjustment for failed seeds is specified in Section A.6.

A.4 Geometric consistency and validity

We assess consistency across rendered views using the static 3D consistency component of WorldScore (Duan et al., 2025). DROID-SLAM (Teed & Deng, 2021) processes the 50-frame walkthrough. For its mean valid reprojection error e_s, the fixed WorldScore normalization gives

$$\mathrm{C3D}_s = 100\left[1 - \mathrm{clip}_{[0,1]}\Big(\frac{e_s}{1.0719}\Big)\right]. \tag{17}$$

Tracking failure or the absence of valid reprojection measurements receives zero. This observation-based geometric proxy depends on tracking conditions, including texture and camera motion. Object placement and boundary agreement are therefore evaluated directly on the declared geometry as follows.

Object placement and support. All geometric quantities use metric units in a right-handed, Z-up world frame with fixed coordinate conversions. Let $\mathcal{O}_s$ be the eligible placed-object instances, $\mathcal{Q}_s \subseteq \mathcal{O}_s$ those requiring support or attachment, and $\mathcal{E}_s$ the expected child–support relations. Structural floors, walls, and terrain provide obstacles and support surfaces but do not enter the placed-object denominator. Define c_i, f_i, and o_i as indicators of unintended collision, missing support, and an assigned-boundary violation; v_{ij} indicates that expected support relation (i, j) passes its geometric checks. We report

$$\mathrm{CR}_s = \frac{100}{|\mathcal{O}_s|} \sum_{i \in \mathcal{O}_s} c_i, \qquad \mathrm{FR}_s = \frac{100}{|\mathcal{Q}_s|} \sum_{i \in \mathcal{Q}_s} f_i, \tag{18}$$

$$\mathrm{OOB}_s = \frac{100}{|\mathcal{O}_s|} \sum_{i \in \mathcal{O}_s} o_i, \qquad \mathrm{SV}_s = \frac{100}{|\mathcal{E}_s|} \sum_{(i,j) \in \mathcal{E}_s} v_{ij}. \tag{19}$$

Collision Rate counts each affected object once, irrespective of how many other objects it intersects. Intended contacts and registered attachment exceptions are excluded. A collision requires a penetration estimate above $0.01\,\mathrm{m}$, or an overlap estimate exceeding both $10^{-5}\,\mathrm{m}^3$ and 0.5% of the smaller object's volume. The implementation combines oriented-box penetration and axis-aligned overlap-volume proxies with mesh-intersection checks.

Floating Rate flags a required support gap above $0.02\,\mathrm{m}$ or the absence of a legitimate support or attachment; wall- and ceiling-mounted instances follow their mounting rules. An object is out of bounds if its center lies outside its assigned room or parcel, or if more than 1% of its projected area crosses that boundary. Support Validity checks each expected relation using a signed contact gap in $[-0.01, 0.02]\,\mathrm{m}$, at least 5% projected support coverage, a compatible mounting category, and a projected object center within the permitted support region. These predicates assess geometric placement and support consistency. They do not estimate dynamic stability from mass distributions. The denominators distinguish object-level defects from failures of expected support relations; unavailable object identities or support metadata make the affected measurements inapplicable.

Agreement across the indoor–outdoor boundary. Let F_s^{out} be the exterior building's ground-floor footprint and F_s^{in} the interior shell's outer envelope in the same world frame. Their planar correspondence is

$$\mathrm{ShapeIoU}_s = \frac{\mathrm{Area}(F_s^{\mathrm{out}} \cap F_s^{\mathrm{in}})}{\mathrm{Area}(F_s^{\mathrm{out}} \cup F_s^{\mathrm{in}})}. \tag{20}$$

The polygons describe the enclosing footprint, excluding furniture occupancy and internal free-space partitions. Coordinate transforms are fixed before evaluation; per-sample alignment or scale optimization is excluded. Shape IoU thus measures envelope agreement in the generated configuration.

Entrance Alignment additionally tests the registered exterior and interior descriptors of the primary portal. An entrance passes only when its normal separation is at most $0.10\,\mathrm{m}$, tangential center offset at most $0.20\,\mathrm{m}$, bottom-height difference at most $0.05\,\mathrm{m}$, outward normal angle at most 15°, and opening IoU at least 0.80. The score is the percentage passing this joint predicate. Portal descriptors are evaluated in the common frame; the geometry provenance specified in Section A.6 determines whether this tests native entrances or adapter-defined openings.

A.5 Navigable space and cross-boundary access

We construct navigation meshes with Recast/Detour using an agent of height $1.70\,\mathrm{m}$ and radius $0.30\,\mathrm{m}$, maximum climb $0.20\,\mathrm{m}$, and maximum slope 45°. Cell size and cell height are both $0.05\,\mathrm{m}$. Postprocessing preserves all connected components without joining disconnected regions. Let A_s^{ref} be the reference walkable NavMesh area, and $A_{s,j}$ the area of component j in the evaluated scene, with $A_s^{\mathrm{nav}} = \sum_j A_{s,j}$. Then

$$\mathrm{NAR}_s = 100\,\mathrm{clip}_{[0,1]}\left(\frac{A_s^{\mathrm{nav}}}{A_s^{\mathrm{ref}}}\right), \tag{21}$$

$$\mathrm{CAR}_s = 100\frac{\max_j A_{s,j}}{A_s^{\mathrm{nav}}}. \tag{22}$$

For structured scenes, the reference is the declared empty structural shell. Surface-only evaluation uses a floor or pedestrian region fixed by the specification. Navigable Area Ratio measures retained traversable area, while Connected Area Ratio measures its fragmentation. A small surviving island can have high CAR, so these quantities are interpreted together. An empty or failed navigation mesh receives zero for both.

NavMesh Success Rate records the fraction of planned runs that produce a minimally usable navigation representation. A run must pass geometry and unit validation, yield a finite, loadable NavMesh with area at least $0.01A_s^{\mathrm{ref}}$, and project a registered spawn onto it within $0.30\,\mathrm{m}$ horizontally and $0.50\,\mathrm{m}$ vertically. Specific entrance and destination queries then test accessibility beyond this construction criterion.

For local entrance traversal, we place probes $1.5\,\mathrm{m}$ outside and inside the primary portal in its declared open state. Entrance Passability is the percentage of entrances admitting a local path between these probes under the same agent-clearance and step constraints. We also sample the fixed exterior–portal–interior centerline at $0.05\,\mathrm{m}$ intervals. With M_s samples and capsule-collision indicator $b_{s,t}$, Transition Collision is

$$\mathrm{TC}_s = \frac{100}{M_s} \sum_{t=1}^{M_s} b_{s,t}. \tag{23}$$

Permitted floor and threshold contacts are excluded. The path-existence test and the collision fraction measure distinct properties: a feasible path may depart from the fixed centerline, and navigation can also fail because of clearance or step constraints.

To test access through the combined outdoor and indoor environment, let $\mathcal{U}_s$ be the registered public outdoor starts, h_s the foyer waypoint, and $\mathcal{G}_s$ the required interior destinations. These endpoints are specified independently of navigation outcomes. Define $r_s(u, g) = 1$ if both endpoints project successfully onto the combined NavMesh and belong to the same connected component, and zero otherwise. The two global access scores are

$$\mathrm{IOC}_s = 100\,\mathbf{1}\{\exists u \in \mathcal{U}_s : r_s(u, h_s) = 1\}\,, \tag{24}$$

$$\mathrm{CBR}_s = \frac{100}{|\mathcal{U}_s||\mathcal{G}_s|} \sum_{u \in \mathcal{U}_s} \sum_{g \in \mathcal{G}_s} r_s(u, g). \tag{25}$$

I–O Connectivity averages IOC_s over planned pairs and tests whether entry from public outdoor space reaches the foyer. Cross-boundary Reach measures access to the required interior destinations from all registered starts. Missing or unprojectable required targets remain in the denominator and contribute zero. Together, these queries distinguish a locally passable doorway from an accessible entrance and a connected interior. They evaluate graph reachability under the specified agent geometry.

A.6 Aggregation, failures, and measurement scope

For a scene-level metric m, observations or evaluator passes are first averaged within the scene. We then average seeds within specifications and give each specification equal weight:

$$\overline{m} = \frac{1}{P} \sum_{p=1}^{P} \frac{1}{K} \sum_{k=1}^{K} \widetilde{m}_{p,k}, \qquad \widetilde{m}_{p,k} = \begin{cases} m_{p,k}, & \text{if a valid measurement is obtained,} \\ b_m, & \text{if the supported run fails.} \end{cases} \tag{26}$$

Failed supported runs remain in the planned denominator. The prescribed failure value b_m is 1 for Q-Align and each AQS dimension, 100 for Collision, Floating, OOB, and Transition Collision, and zero for the remaining non-diversity metrics. Unsupported measurements are marked N/A with their missing capability; incomplete evaluations retain a pending status.

Diversity is aggregated at the specification level. If n_p of the K seeds succeed, the implemented failure adjustment is

$$\widetilde{D}_p(d) = \begin{cases} (n_p/K)^2 D_p^{\mathrm{valid}}(d), & n_p \geq 2, \\ 0, & n_p < 2, \end{cases} \qquad \overline{D}(d) = \frac{1}{P} \sum_{p=1}^{P} \widetilde{D}_p(d). \tag{27}$$

Table 6: Methods are evaluated on their supported indoor and urban outdoor domains under the same protocol. For methods supporting both domains, *Full* reports the average indoor–outdoor performance. Valid Gen. denotes the number of successful scene generations out of 100 runs.

Method	Valid Gen.	Scene Coverage	Perceptual Quality		Scene Quality			Diversity	
			Q-Align ↑	CLIP-IQA+ ↑	Layout Plaus. ↑	Prompt Align. ↑	3D Consistency ↑	Appearance Div. ↑	Layout Div. ↑
Single-domain methods									
Infinigen Indoor (Raistrick et al., 2024)	89	Indoor	3.72	0.520	55.7	52.0	82.1	0.767	0.966
SpatialGen (Fang et al., 2026)	100	Indoor	3.29	0.538	57.4	81.1	83.2	0.348	0.783
SceneWeaver (Yang et al., 2026a)	47	Indoor	3.59	0.574	60.8	68.0	38.3	0.640	0.888
MetaUrban (Wu et al., 2025)	100	Outdoor	3.85	0.629	60.0	75.6	68.1	0.267	0.775
Indoor–outdoor methods									
WorldGen (Wang et al., 2026a)	98	Indoor	2.84	0.388	71.2	73.0	80.9	0.683	0.946
		Outdoor	2.54	0.405	58.7	52.2	61.0	0.660	0.933
		Full	2.69	0.397	65.0	62.6	71.0	0.672	0.940
SynCity-3K (Engstler et al., 2026)	97	Indoor	3.05	0.496	44.9	73.5	87.7	0.416	0.956
		Outdoor	3.47	0.558	62.3	77.3	89.3	0.379	0.779
		Full	3.26	0.527	53.6	75.4	88.5	0.398	0.868
HY-World 2.0 (Team et al., 2026)	81	Indoor	1.62	0.279	86.3	15.2	81.7	0.623	0.878
		Outdoor	1.72	0.342	92.6	38.3	76.7	0.652	0.964
		Full	1.67	0.311	89.5	26.8	79.2	0.638	0.921
GLM-5.3 Flash (Zeng et al., 2026)	44	Indoor	3.14	0.505	48.8	69.6	66.2	0.512	0.935
		Outdoor	3.44	0.591	58.1	78.9	72.8	0.640	0.963
		Full	3.29	0.548	53.5	74.3	69.5	0.576	0.949
GLM-5.3 (Zeng et al., 2026)	42	Indoor	3.36	0.518	55.8	67.9	72.8	0.505	0.935
		Outdoor	3.34	0.592	71.9	75.9	51.1	0.594	0.940
		Full	3.35	0.555	63.9	71.9	62.0	0.550	0.938
Gemini 3.1 Pro (Team et al., 2023)	76	Indoor	3.45	0.504	53.9	57.9	67.5	0.392	0.857
		Outdoor	3.41	0.570	51.9	84.7	51.9	0.578	0.973
		Full	3.43	0.537	52.9	71.3	59.7	0.485	0.915
GPT-6 Astra *(low)* (OpenAI, 2026)	97	Indoor	3.92	0.566	86.1	71.7	80.5	0.439	0.836
		Outdoor	3.74	0.639	88.3	80.3	54.8	0.445	0.821
		Full	3.83	0.603	87.2	76.0	67.7	0.442	0.829
GPT-6 Astra *(medium)* (OpenAI, 2026)	88	Indoor	3.95	0.574	60.1	64.8	78.5	0.475	0.858
		Outdoor	3.76	0.637	71.5	78.1	54.8	0.486	0.853
		Full	3.86	0.606	65.8	71.5	66.7	0.481	0.856
GPT-6 Astra *(high)* (OpenAI, 2026)	91	Indoor	3.95	0.588	73.4	70.2	82.9	0.514	0.887
		Outdoor	3.83	0.643	73.9	80.7	58.6	0.513	0.851
		Full	3.89	0.616	73.7	75.5	70.8	0.514	0.869
GPT-6 Astra *(xhigh)* (OpenAI, 2026)	82	Indoor	4.02	0.599	66.6	67.1	84.0	0.550	0.902
		Outdoor	3.89	0.641	70.3	84.4	78.8	0.569	0.845
		Full	3.96	0.620	68.5	75.8	81.4	0.560	0.874

This is an explicit success-rate-squared penalty. The fraction of valid unordered seed pairs would instead be $\binom{n_p}{2}/\binom{K}{2}$, so the two weighting rules are not interchangeable. We compute 95% percentile confidence intervals using 10,000 bootstrap resamples of specifications, retaining all seeds and associated observations within each sampled specification. This captures variation over evaluation inputs; synthetic profiles are not treated as independent annotators. Indoor, outdoor, and paired-scene populations are reported separately.

Each geometric score refers to the representation supplied to the evaluator. Computing collision and support rates per instance requires object identities and support metadata; reconstructed surfaces may support navigation while lacking the metadata for these tests. The fixed indoor–outdoor adapters used with WorldGen (Wang et al., 2026a) and HY-World (Team et al., 2026) provide a synthetic shell, portal, and walkway. Their native generated content is retained for visual assessment but excluded from the adapter collision geometry. Accordingly, geometric and navigation measurements for these variants characterize the adapter geometry; their spatial AQS also depends on an adapter-derived overlay. Native-world access requires evaluation on the corresponding generated geometry. AQS values taken from prior work retain their original evaluation provenance and are distinguished from scores obtained with this local VLM protocol.

B Additional Results

B.1 Quantitative Results

Generation quality. Table 6 shows that single-domain methods can achieve strong performance within their specialized domains, but they do not cover both indoor and outdoor environments. For example, Infinigen Indoor (Raistrick et al., 2024) produces diverse indoor scenes, whereas MetaUr-

Table 7: Quantitative comparison of 3D structural validity and navigability. Every method is shown with Indoor, Outdoor, and *Full* rows. For indoor–outdoor methods, *Full* is the unweighted average of the two domain-level results. Object-level structural metrics are unavailable for strict surface-only outputs. N/A denotes an unsupported domain, and N/R denotes a result not reported in the available evaluation.

	Category		Structural Validity				Navigability		
Method	Scene Coverage	Track	Collision Rate ↓	Floating Rate ↓	OOB Rate ↓	Support Validity ↑	Navigable Area Ratio ↑	Connected Area Ratio ↑	NavMesh Success Rate ↑
Single-domain methods									
	Indoor		14.4	14.8	12.8	85.2	59.7	83.8	84.0
Infinigen Indoor (Raistrick et al., 2024)	Outdoor	Structured	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Full		N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Indoor		–	–	–	–	12.5	32.2	46.0
SpatialGen (Fang et al., 2026)	Outdoor	Surface-only	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Full		N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Indoor		50.1	42.0	38.0	43.2	46.5	47.6	50.0
SceneWeaver (Yang et al., 2026a)	Outdoor	Structured	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Full		N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Indoor		N/A	N/A	N/A	N/A	N/A	N/A	N/A
MetaUrban (Wu et al., 2025)	Outdoor	Structured	39.3	9.4	14.1	48.3	95.2	93.7	100.0
	Full		N/A	N/A	N/A	N/A	N/A	N/A	N/A
Indoor–outdoor methods									
	Indoor		–	–	–	–	24.2	99.0	100.0
WorldGen (Wang et al., 2026a)	Outdoor	Surface-only	–	–	–	–	N/R	N/R	N/R
	Full		–	–	–	–	N/R	N/R	N/R
	Indoor		–	–	–	–	4.6	10.1	23.0
SynCity-3K (Engstler et al., 2026)	Outdoor	Surface-only	–	–	–	–	0.0	0.0	0.0
	Full		–	–	–	–	2.3	5.1	11.5
	Indoor		–	–	–	–	10.8	77.5	84.0
HY-World 2.0 (Team et al., 2026)	Outdoor	Surface-only	–	–	–	–	0.5	8.9	17.0
	Full		–	–	–	–	5.7	43.2	50.5
	Indoor		61.3	74.2	60.6	25.8	38.4	43.4	43.0
GLM-5.3 Flash (Zeng et al., 2026)	Outdoor	Explicit mesh	75.1	79.6	74.6	20.4	27.1	14.2	27.0
	Full		68.2	76.9	67.6	23.1	32.8	28.8	35.0
	Indoor		65.5	71.4	60.7	28.6	33.1	40.5	40.0
GLM-5.3 (Zeng et al., 2026)	Outdoor	Explicit mesh	87.0	92.0	89.0	8.0	12.7	7.9	12.0
	Full		76.3	81.7	74.8	18.3	22.9	24.2	26.0
	Indoor		34.7	37.9	26.1	62.1	63.9	75.9	77.0
Gemini 3.1 Pro (Team et al., 2023)	Outdoor	Explicit mesh	33.3	33.2	39.1	66.8	68.2	34.0	68.0
	Full		34.0	35.5	32.6	64.5	66.1	54.9	72.5
	Indoor		9.3	23.6	3.0	76.4	66.2	95.7	93.0
GPT-6 Astra (*low*) (OpenAI, 2026)	Outdoor	Explicit mesh	7.1	20.7	13.3	79.3	63.2	44.7	51.0
	Full		8.2	22.2	8.2	77.8	64.7	70.2	72.0
	Indoor		17.5	30.1	12.0	69.9	59.9	86.8	84.0
GPT-6 Astra (*medium*) (OpenAI, 2026)	Outdoor	Explicit mesh	5.3	20.8	14.5	79.2	68.6	36.3	55.0
	Full		11.4	25.5	13.2	74.5	64.3	61.5	69.5
	Indoor		15.0	30.8	9.0	69.2	61.7	89.7	85.0
GPT-6 Astra (*high*) (OpenAI, 2026)	Outdoor	Explicit mesh	7.3	22.5	18.3	77.5	59.1	34.2	51.0
	Full		11.1	26.6	13.7	73.4	60.4	62.0	68.0
	Indoor		13.9	28.2	8.0	71.8	55.3	81.8	80.0
GPT-6 Astra (*xhigh*) (OpenAI, 2026)	Outdoor	Explicit mesh	6.6	21.3	15.4	78.7	58.0	33.1	55.0
	Full		10.2	24.8	11.7	75.2	56.6	57.5	67.5

ban (Wu et al., 2025) achieves strong perceptual quality in outdoor environments. In contrast, methods supporting both domains must balance generation reliability, perceptual quality, 3D consistency, and diversity under a substantially more demanding setting.

Among the indoor–outdoor methods, the GPT-6 Astra variants (OpenAI, 2026) provide the strongest perceptual quality. In particular, GPT-6 Astra (*xhigh*) (OpenAI, 2026) achieves the best Full Q-Align and CLIP-IQA+ scores of 3.96 and 0.620, respectively. Compared with the strongest non-GPT indoor–outdoor baselines, this corresponds to an absolute improvement of 0.53 in Q-Align over Gemini 3.1 Pro (Team et al., 2023) and 0.065 in CLIP-IQA+ over GLM-5.3 (Zeng et al., 2026). The same configuration also obtains a 3D consistency score of 81.4, second only to the 88.5 achieved by SynCity-3K (Engstler et al., 2026). Thus, its improvement in perceptual quality does not come at the cost of severely degraded cross-view geometric consistency.

The different GPT-6 Astra (OpenAI, 2026) reasoning levels reveal a clear, but non-uniform, trade-off. Increasing the reasoning level from *low* to *xhigh* consistently improves Q-Align from 3.83 to 3.96, CLIP-IQA+ from 0.603 to 0.620, appearance diversity from 0.442 to 0.560, and layout diversity from 0.829 to 0.874. These results suggest that a larger reasoning budget produces visually stronger and more varied scenes. However, the valid-generation count decreases from 97 for the *low* configuration to 82 for *xhigh*, indicating that the increased generation complexity can reduce execution reliability. Therefore, additional reasoning primarily benefits perceptual quality and content richness rather than uniformly improving every evaluation dimension.

Other methods retain advantages on individual metrics. HY-World 2.0 (Team et al., 2026) achieves the highest Full layout plausibility of 89.5, but its prompt-alignment score is only 26.8. SynCity-3K (Engstler et al., 2026) obtains the highest 3D consistency of 88.5, while remaining relatively limited in layout plausibility and appearance diversity. Among the competing indoor–outdoor baselines, WorldGen (Wang et al., 2026a) and GLM-5.3 Flash (Zeng et al., 2026) achieve the strongest appearance diversity and layout diversity, with scores of 0.672 and 0.949, respectively. These results demonstrate that no individual metric is sufficient to characterize world-generation quality: structured procedural priors often benefit layout or diversity, whereas stronger general-purpose generative models tend to provide better perceptual quality and semantic alignment.

Structural validity and navigability. Table 7 further reveals substantial differences among scene representations. Object-level collision, floating, out-of-bounds, and support metrics cannot be reliably computed for strict surface-only outputs because they do not expose an explicit object decomposition. These missing values should therefore not be interpreted as zero error or superior structural validity; surface-only methods are primarily comparable through their navigability results.

Among indoor–outdoor methods producing explicit meshes, GPT-6 Astra (*low*) (OpenAI, 2026) provides the strongest overall structural validity. Its Full collision, floating, and out-of-bounds rates are 8.2, 22.2, and 8.2, respectively, and it achieves the highest support validity of 77.8. It also obtains the highest connected-area ratio among explicit-mesh methods, reaching 70.2. Gemini 3.1 Pro (Team et al., 2023) performs slightly better in navigable-area ratio and NavMesh success rate, with scores of 66.1 and 72.5, compared with 64.7 and 72.0 for GPT-6 Astra (*low*) (OpenAI, 2026). Overall, the *low* configuration provides a particularly balanced trade-off among structural correctness, spatial connectivity, and successful navigation.

Notably, increasing the reasoning level does not monotonically improve geometry or navigability. Although GPT-6 Astra (*xhigh*) (OpenAI, 2026) produces stronger perceptual quality, richer scenes, and higher 3D consistency, its Full navigable-area and connected-area ratios decrease to 56.6 and 57.5, respectively. One possible explanation is that additional visual and geometric complexity introduces more obstacles and small structural elements, making collision-free geometry and NavMesh construction more difficult. Consequently, the *xhigh* configuration is preferable when perceptual quality and scene diversity are the primary objectives, whereas the *low* configuration provides a more reliable structural basis for navigation, physical interaction, and large-scale generation.

Several single-domain methods remain highly competitive within their supported environments. Infinigen Indoor (Raistrick et al., 2024) achieves a support validity of 85.2 and a NavMesh success rate of 84.0 on indoor scenes, while MetaUrban (Wu et al., 2025) obtains a navigable-area ratio of 95.2 and a NavMesh success rate of 100.0 outdoors. These results, however, rely on domain-specific priors and do not directly measure unified indoor–outdoor generation. Taken together, the two tables expose a clear Pareto frontier among perceptual quality, structural validity, navigability, diversity, and generation success, with no method dominating all evaluation dimensions.

B.2 Additional Qualitative Results

Additional qualitative results and videos are available on our anonymous project webpage.[1] The webpage first presents four complete generated worlds: a civic fountain district, a riverside industrial district, a lakeside medical district, and a greenway service station. Each world is shown through a scene-level video together with complementary outdoor and indoor views, covering diverse functional spaces such as libraries, cafes, restaurants, fire stations, hospitals, pharmacies, markets, and service facilities.

The webpage also presents six spatial-continuity examples. Each example includes a scene overview, a continuous outdoor-to-indoor camera trajectory, and the corresponding indoor-to-outdoor trajectory. These results demonstrate that entrances, indoor layouts, building envelopes, and surrounding roads or public spaces are represented within a shared geometric frame. The entrance locations and functional semantics remain coherent across views: for example, hospital facades connect to reception and pharmacy spaces, while the vehicle-scale infrastructure of a service station connects to its market and neighboring restaurant interiors. The primary qualitative benefit is there-

[1] https://anonymous.4open.science/w/worldbridge-4F42/

fore not limited to the appearance of an individual frame, but also includes topological continuity and functional consistency across building boundaries.

Beyond static generation, the webpage demonstrates dynamic fountain water and ripples, wind-driven vegetation and falling leaves, and vehicle arrival and departure in a service district. These examples show that the generated worlds can be configured as executable environments with dynamic state. Four cross-space robot demonstrations further involve picking up and transporting a snack bag, coffee cup, apple, and parcel between different indoor spaces through the shared outdoor environment. Together, these examples illustrate support for continuous navigation and object interaction across indoor and outdoor regions.

Finally, the webpage provides image and video comparisons covering three indoor-only methods, two outdoor-only methods, and ten methods or configurations that produce results in both domains. The examples highlight differences in perceptual quality, scene complexity, geometric completeness, and motion stability. Because the scene prompts and camera trajectories are not strictly matched across methods, these comparisons should be interpreted as representative qualitative examples rather than a replacement for the quantitative evaluation. Moreover, "indoor and outdoor" in the comparison section only indicates that a method produces outputs in both domains; it does not necessarily imply that those outputs form a connected indoor–outdoor world. The robot trajectories are authored motion and interaction sequences, and are intended to demonstrate scene executability rather than learned robot-policy performance.