Zero-Shot Generative World Priors for Planning and Object Reaching in Partially Observed Rooms

Hao Lu*, Subhransu S. Bhattacharjee, Dylan Campbell, Rahul Shome
School of Computing, Australian National University, Canberra, ACT 2600, Australia
*Corresponding author: Hao.Lu@anu.edu.au.

***Abstract*—Autonomous robots often view indoor rooms only partially, for example, from outside a doorway, while walls and scene structure hide the geometry and target-object semantics needed for safe navigation and object reaching. We study whether off-the-shelf pretrained generative vision models can provide a usable *world sampler* for planning in such partially observed scenes. From an egocentric RGB crop and a queried object category, our pipeline samples plausible hidden-room completions, lifts them into semantically labeled 3D point clouds, and passes the resulting sampled worlds to a planner in configuration space. The planner optimizes a joint success criterion over sampled worlds: a plan is successful in one sampled world only if it remains collision-free and terminates within reaching distance of the sampled target. This induces an Minimum Constraint Removal (MCR)-style objective that minimizes the number of sampled worlds in which either collision occurs or the target is not reached. Experiments on doorway-occluded indoor scenes derived from Matterport3D show that planning over multiple generated worlds substantially improves both safety and object-reaching performance over uniform and single-sample baselines, without task-specific fine-tuning. These results suggest that zero-shot generative world sampling can make pretrained foundation models available to embodied planners as uncertainty-aware, physically queryable world models for acting beyond the field of view.**

I. Introduction and Related Work

Robots operating in homes, offices, and shared indoor spaces rarely receive complete scene geometry before taking action. A doorway observation may expose the room type and some visible free space, while leaving hidden obstacles and the queried object entirely occluded. In this regime, planning from a single deterministic completion is brittle, yet planning without any prior over the hidden room is overly conservative or inefficient. The central question in this paper is whether pretrained generative vision models can be used as *zero-shot world samplers* that expose the missing geometry and task semantics needed for planning.

Classical planning under uncertainty addresses such settings by reasoning over belief space rather than a single deterministic reconstruction [1, 2, 3]. Related navigation and object-search systems use occupancy maps, semantic maps, commonsense spatial priors, or active perception to reduce uncertainty [4, 5, 6, 7, 8]. Sampling-based motion planning under environment uncertainty provides a particularly useful interface: if a sampler can produce realizations of the uncertain environment, paths can be evaluated by the set of worlds in which they fail, connecting uncertainty-aware planning to minimum-constraint-removal objectives [9, 10]. However, for doorway-occluded indoor scenes, the main practical difficulty is obtaining such a sampler. Handcrafted priors and coarse spatial relations can be useful, but they often lack the geometric detail required for collision checking and the semantic specificity required for object reaching.

Recent world-model and generative-model work seeks to connect advances in learned prediction and representation to the requirements of embodied control [11, 12, 13, 14, 15, 16]. In parallel, generative vision and vision-language models implicitly encode rich regularities of indoor scenes [17, 18, 19, 20, 21, 22]. Recent work suggests that pretrained generative models can be used as effective world-representation samplers, where repeated sampling exposes plausible scene structure under uncertainty [22, 16]. Complementary efforts show that even computationally cheap representations can support useful spatial abstraction for downstream reasoning [23]. This suggests a different role for world models in robot planning: instead of predicting only long-horizon image rollouts or returning one best reconstruction, a world model can sample plausible hidden spatial states that are directly queryable by a planner. In this paper, we formulate zero-shot hidden-room completion as *generative world sampling* for navigation and object reaching. The robot observes a partial RGB image and is given a queried object class, such as *chair* or *table*. A pretrained vision-language model first scores room–object plausibility and constructs a prompt. A generative image model then samples plausible hidden-room completions. Monocular depth estimation and semantic segmentation lift each completion into a labeled 3D point cloud. Each sample therefore defines a possible occupancy map and a possible target region.

Our planner follows a sampled-world formulation for planning under environment uncertainty [10]. A trajectory is judged jointly on two conditions in each sampled world: it must be collision-free along the path, and its terminal state must reach the sampled target. Intuitively, the selected path should fail in as few world samples as possible. This coupling is important: a path that is safe but terminates far from likely target regions is still poor, and a path that reaches a likely target while colliding in many plausible worlds is unusable.

As illustrated in Figure 1, we address this problem by converting a partial doorway observation and object query into sampled semantically labeled 3D world hypotheses, which are then consumed by an uncertainty-aware planner for safe navigation and object reaching.

We evaluate this view on 173 doorway-occluded object-

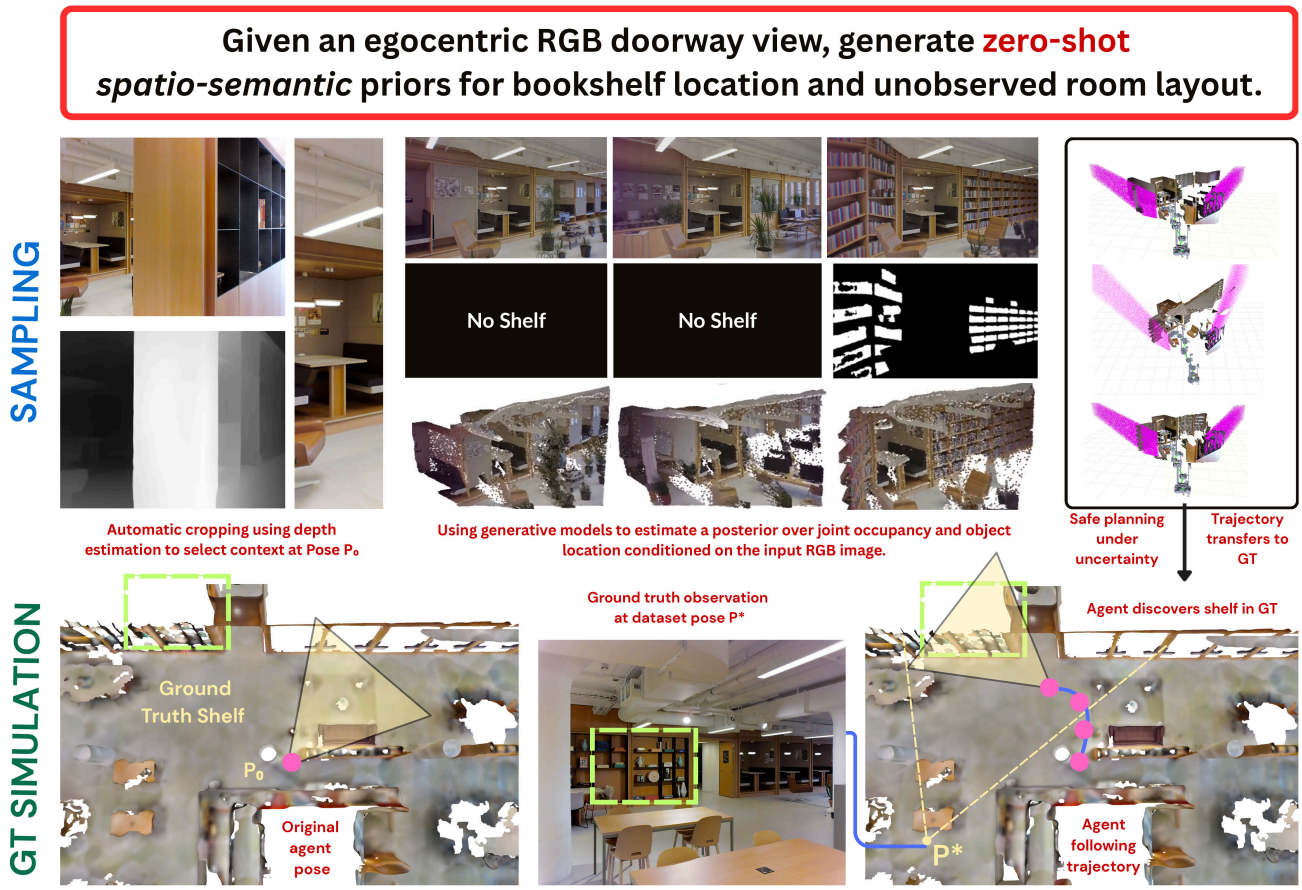

Fig. 1: Overview. Given a doorway RGB crop and queried object, the system uses VLM-guided prompting and outpainting to sample hidden-room images, lifts them into semantically labeled 3D point clouds with depth and segmentation, and plans over the resulting sampled worlds. The planner selects a path that minimizes failure across sampled worlds, where failure means either collision or unsuccessful object reaching.

reaching tasks derived from Matterport3D. The results show two consistent patterns. First, the generative pipeline produces informative world samples: compared with geometric heuristics and uniform sampling, it places target hypotheses substantially closer to the ground-truth target and yields meaningful occupancy structure. Second, planning over the *set* of sampled worlds is much stronger than planning with either no hidden-room prior or a single generated world. Relative to a uniform prior, the sampled-world planner delivers over three times higher target-reaching success and halves the collision rate. These findings suggest that zero-shot generative scene completion goes beyond predicting the appearance of unseen regions, instead producing a distribution over plausible world models that can be queried by an interpretable, composable planning layer.

II. Problem Formulation

Doorway-observation setting. Let Θ be an egocentric RGB observation captured from outside a doorway and let o be a queried object category. The hidden room is represented by a world variable $\xi \in \mathcal{W}$, where a world contains the observed evidence together with a completion of hidden occupancy, free space, and semantic labels. We model the scene uncertainty by a conditional distribution $Q_{\Theta,o}$ over the measurable scene space $(\mathcal{W}, \mathcal{A})$ and draw

$$\xi^1, \ldots, \xi^N \overset{\text{i.i.d.}}{\sim} Q_{\Theta,o}. \tag{1}$$

The samples form an empirical belief over the unobserved part of the room. Each world ξ^i contains a 3D occupied set and a target-evidence set $\mathcal{E}_o(\xi^i) \subseteq \mathbb{R}^3$ for the queried object, with $\mathcal{E}_o(\xi^i) = \emptyset$ when the object is absent or not realized in that completion.

For a measurable workspace region r, the sampled target evidence in that region is

$$p^{\text{reg}}_{\text{det}}(\Theta, o; r) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\mathcal{E}_o(\xi^i) \cap r \neq \emptyset\right],$$
$$p_{\text{det}}(\Theta, o) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\mathcal{E}_o(\xi^i) \neq \emptyset\right]. \tag{2}$$

These quantities summarize semantic belief, but they are insufficient for action: the planner also needs to know whether

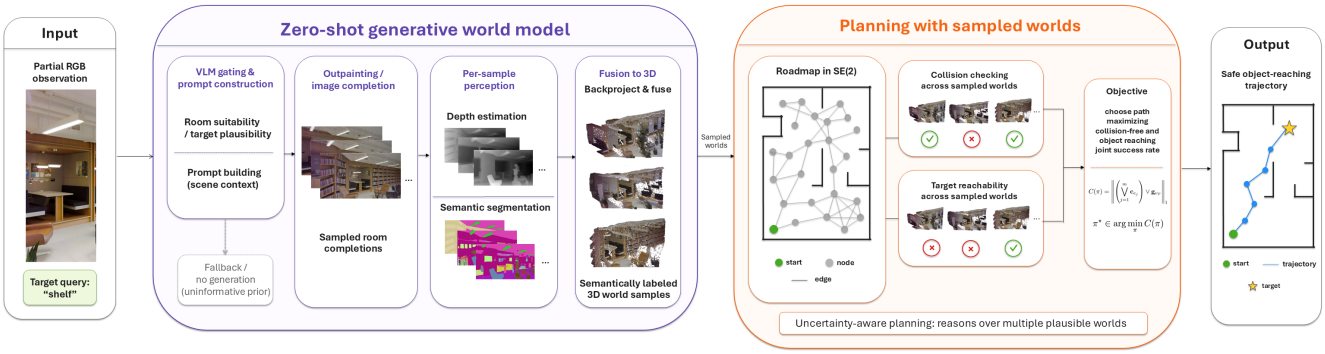

Fig. 2: **Pipeline.** Given a partial RGB observation through a doorway and a queried object category, the system first constructs a zero-shot generative world model of the unobserved room. A VLM-based gating and prompt-construction stage estimates whether the target is plausible in the hidden room and provides scene context for generation. Conditional outpainting then samples multiple plausible room completions. Each completed image is processed by depth estimation and semantic segmentation, and the resulting predictions are back-projected and fused into semantically labeled 3D point-cloud hypotheses. These hypotheses form sampled worlds containing both occupancy and target evidence. The planner consumes these sampled worlds directly: it constructs a roadmap in configuration space, checks collision and target reachability across the sampled worlds, and selects a trajectory that maximizes the joint success rate of remaining collision-free and reaching the object.

the robot can move safely to a configuration that reaches the target. The distinction matters because object likelihood and traversability are not independent in cluttered rooms. A likely target location behind a sofa may be unreachable, while a collision-free corridor may not lead to a useful viewpoint or grasping location. We therefore keep occupancy and target evidence coupled at the sample level throughout planning, rather than converting them into independent marginal heatmaps.

Object-reaching success. Let $\mathcal{X} \subseteq \mathrm{SE}(2)$ be the base configuration space of the mobile robot. In a sampled world ξ^i, let $\mathcal{F}(\xi^i) \subseteq \mathcal{X}$ denote the collision-free configurations and let $\mathcal{G}_o(\xi^i) \subseteq \mathcal{X}$ denote terminal configurations from which the robot can reach the sampled target. If no target evidence is present in ξ^i, then $\mathcal{G}_o(\xi^i) = \emptyset$.

A trajectory π is checked at a finite set of configurations $\mathcal{K}(\pi)$, including roadmap vertices and dense edge interpolants. If x_T is the terminal configuration, the sampled task-success probability is

$$p^N_{\mathrm{succ}}(\pi \mid \Theta, o) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[x_K \in G_o(\xi^i)\big] \prod_{x \in \mathcal{K}(\pi)} \mathbf{1}\big[x \in F(\xi^i)\big]\,. \tag{3}$$

The product couples all collision checks along the path within a single sampled world, and the terminal indicator couples target semantics with reachability. Thus a trajectory receives credit in a sampled world only when it satisfies both safety and goal-reaching requirements.

Roadmap objective. We implement Eq. (3) on a PRM-style roadmap $G = (V, E)$ [24]. For an edge $e \in E$, define a sampled-world failure vector $\mathbf{c}_e \in \{0,1\}^N$ with component

$$c^i_e = \mathbf{1}\big[\exists x \in \mathcal{K}(e) : x \notin \mathcal{F}(\xi^i)\big]\,. \tag{4}$$

For a terminal vertex $v \in V$, define a goal-failure vector $\mathbf{g}_v \in \{0,1\}^N$ with component

$$g^i_v = \mathbf{1}\big[v \notin \mathcal{G}_o(\xi^i)\big]\,. \tag{5}$$

For a roadmap path $\pi = (e_1, \ldots, e_m, v_T)$, the planner minimizes

$$C(\pi) = \left\| \left(\bigvee_{j=1}^{m} \mathbf{c}_{e_j} \right) \vee \mathbf{g}_{v_T} \right\|_1$$

$$\pi^\star \in \arg\min_{\pi} C(\pi), \tag{6}$$

where $\vee$ is the elementwise logical union. The objective counts the sampled worlds in which the trajectory is invalid due to either path collision or terminal target unreachability. Equivalently, minimizing $C(\pi)$ maximizes the empirical task-success probability in Eq. (3). This objective is related to minimum-constraint-removal planning [9]: sampled worlds play the role of constraints, and a preferred path violates fewer of them.

As shown in Figure 2, the proposed system separates world sampling from planning while keeping the interface between them explicit and queryable. The generative component is not used to produce a single deterministic reconstruction or a direct action prediction. Instead, it produces a set of semantically labeled 3D world hypotheses, each of which can be queried by a geometric planner for occupancy, collision, and target reachability. This representation is particularly suitable for partially observed rooms, where multiple completions may be visually and semantically plausible from the doorway observation. The planning component then aggregates these hypotheses through the sampled-world objective, preferring trajectories that remain valid and task-successful across many possible hidden-room completions. This makes the uncertainty represented by the generative model operational: uncertainty

is not only visualized as alternative completions, but directly affects which path and terminal goal configuration are selected.

III. Zero-Shot Generative World Sampling

The world sampler maps a doorway crop and target query to up to $N = 100$ completed worlds without retraining. The pipeline includes four stages.

Semantic prompting and gating. A vision-language model scores whether the queried object is plausible behind the doorway and generates the prompt used for hidden-room completion [21]. This step injects language-level semantic structure into the world samples before any planning occurs.

Image-space completion. Conditioned on the doorway crop, completion side, and query, an outpainting model samples multiple RGB completions of the hidden room while preserving the visible evidence [20, 25]. In our implementation, the visible band is protected using the observed depth range, and each accepted query is expanded by 500 pixels on both sides before completion. These samples represent diverse but observation-consistent hypotheses of room layout and object placement.

3D lifting. Each sampled RGB completion is converted into a 3D world using monocular metric depth estimation and semantic segmentation [26, 27, 28]. We use an ADE20K-trained segmenter mapped to the task vocabulary, filter small or low-confidence regions, align predicted depth to the observed crop to reduce scale ambiguity using RANSAC [29], remove statistical outliers, and back-project to semantically labeled point clouds. Using known camera intrinsics preserves the image rays needed for consistent lifting across samples.

Planning representation. Each lifted sample is split into an occupancy point cloud and a target point cloud for the queried object. We fuse the observed ground-truth visible geometry into every sample so that all planners agree on the observed region and differ only in how they model the hidden room. This produces a sampled set of world hypotheses that can be queried for collision and target reachability.

This representation is intentionally explicit rather than latent. A planner can collision-check a configuration against a sampled occupancy point cloud, test whether a terminal state reaches a sampled target point cloud, and aggregate these outcomes across sampled worlds. In this sense, the generative model serves as a world model through a *sampling interface*: it does not need to output actions directly, but it must produce a distribution over physically queryable worlds. Excluding amortized preprocessing, world generation runs at approximately 1 ± 0.25 s per sample, which makes offline belief construction practical for batched planning experiments even though online deployment would require further optimization.

IV. Planning with Sampled Worlds

We instantiate the planner for a Stretch mobile robot navigating with its base in configuration space [30]. For each task, we build a PRM-style roadmap [24] with 3000 sampled configurations inside the room-specific search volume, producing approximately 5×10^4 edges on average. Collision checking is performed along interpolated edge states using FCL [31] through OMPL [32], with step sizes of 0.05 m in translation and 0.10 rad in rotation. A terminal state is considered successful in sampled world ξ^i if it lies within 0.25 m of that sample's target hypothesis.

We compare four planning priors. *Uniform* uses the visible room geometry and a single target sampled uniformly in the hidden search region. *1-sample* uses the visible geometry plus one generated occupancy–target hypothesis. *Ours* uses all 100 generated hypotheses and minimizes the world-failure objective in Eq. (6). *GT oracle* uses ground-truth hidden-room geometry and target location as a non-deployable upper bound. All resulting trajectories are replayed against the same ground-truth scene and target.

This formulation highlights why multiple sampled worlds matter. A single generated completion may look plausible yet still induce a brittle plan because it commits to one occupancy layout and one target placement. In contrast, sampled-world planning prefers paths that remain valid for many hypotheses and terminate near regions where the target is repeatedly hypothesized to be reachable. The optimization therefore balances geometric caution with semantic intent. Because collision tests for different edges and sampled worlds are independent, the planner is also amenable to parallelization; this matters in our setting because repeated geometric validation, rather than graph search, dominates total runtime.

V. Experiments

Evaluation tasks. We evaluate on doorway-occluded indoor scenes derived from Matterport3D [33]. The full dataset contains 100 doorway contexts crossed with 10 object categories; for planning we use the 173 verified queries with ground-truth target instances and reliable region-based navigation support. The target may be partially visible, fully occluded, or absent in the visible crop. For each generated world, we fuse the observed geometry from the visible region into the sampled completion, crop the floor plane to avoid treating traversable floor as an obstacle, and voxelize the occupancy point cloud for faster collision checking. The search volume is task-specific and derived from the doorway-view setup rather than from a fixed global field of view.

World-sampling metrics. We report semantic alignment of room–object priors (D_{sem}), object realization rate p_{det}, target localization via mean and best object-placement recall at 1 m (mOPR and bOPR), geometry quality via symmetric Chamfer distance [34], and placement uncertainty via an energy score [35, 36]. These metrics measure whether the sampled worlds capture the hidden-room structure needed by a planner. We emphasize these generative metrics because our goal is not to claim a new image completion model. Instead, they quantify whether the sampled worlds preserve the geometric and semantic cues required for planning.

Generative baselines. We compare the sampled world model against several baselines. The *uniform object prior* assigns equal probability to every object category in every room and is evaluated only by D_{sem}. The *no generation* baseline emits no object or geometry. The *uniform random search-volume*

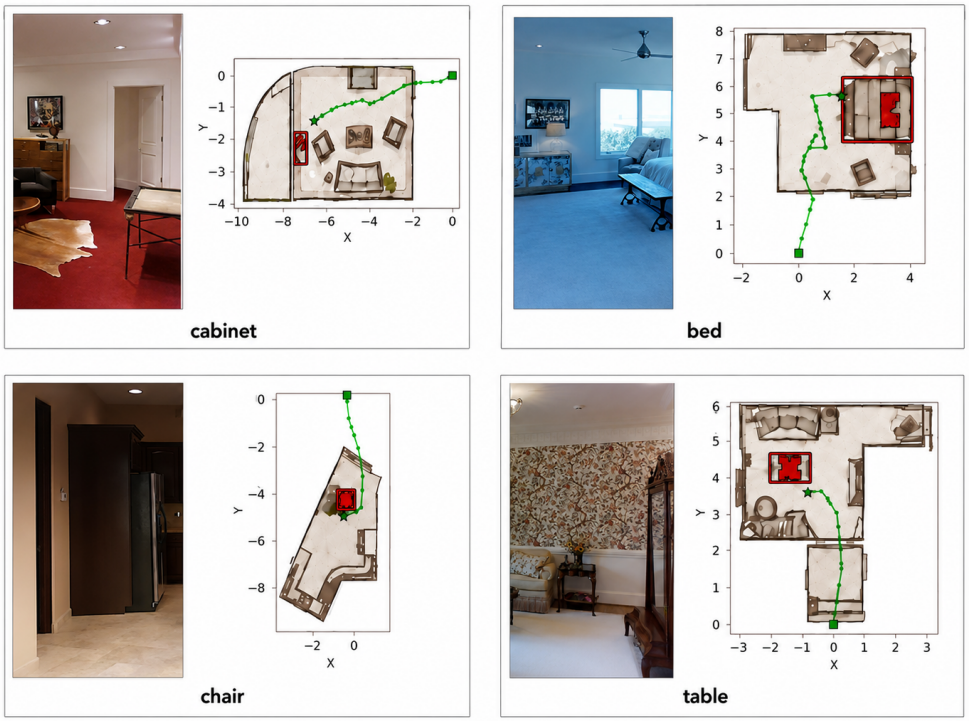

Fig. 3: **Qualitative trajectories.** Representative examples drawn from the subset of tasks in which the planned trajectory successfully reaches the ground-truth target. For each example, the left image shows the input crop and the right image shows the planned trajectory evaluated in the ground-truth scene. Light-brown points show ground-truth occupancy, red regions show the queried target object, green curves show the trajectories, squares indicate start states, and stars indicate terminal states. The examples cover different queried objects, room layouts, and visibility conditions, including partially visible and out-of-crop targets. During planning, the trajectories are optimized using the 100 generated occupancy–target hypotheses. The ground-truth occupancy and target are shown here only for visual evaluation.

baseline samples target locations uniformly in the valid hidden-room volume without RGB or object cues, using $K = 100$ samples averaged over seeds. We also include two deterministic one-hypothesis controls that place a target-sized box at either the search-volume centroid or the doorway-side frontier [37]. The *GT oracle* is a non-deployable ceiling using Matterport3D ground-truth target boxes and scene geometry.

Planning metrics. We evaluate each planned trajectory by replaying it in the corresponding ground-truth scene. We report ground-truth target-reaching success $p_{\mathrm{tgt}}^{\mathrm{GT}}$, which measures whether the terminal robot configuration reaches the true target object within the specified success threshold ($1\mathrm{m}$). We report collision rate $p_{\mathrm{coll}}^{\mathrm{GT}}$, which measures the fraction of trajectories that collide with the ground-truth occupancy during replay. The terminal target distance d_{tgt} measures how close the final robot configuration is to the ground-truth target. The trajectory distance d_{traj} compares the planned trajectory with the ground-truth-oracle trajectory using Fréchet distance [38], indicating how closely the selected path follows a trajectory planned with full scene information. We also report the trajectory-length-weighted collision-free fraction χ_{GT}, which measures the proportion of the trajectory that remains collision-free in the ground-truth scene, and the collision-free prefix fraction C_f, which measures how much of the trajectory can be executed before the first collision occurs. Higher values are better for $p_{\mathrm{tgt}}^{\mathrm{GT}}$, χ_{GT}, and C_f, while lower values are better for $p_{\mathrm{coll}}^{\mathrm{GT}}$, d_{tgt}, and d_{traj}. We further report runtime to expose the computational cost of sampled-world planning.

TABLE I: World-sampling results on doorway-occluded scenes. Lower is better for D_{sem}, Chamfer (CD), and energy score.

Method	D_{sem}	p_{det}	mOPR	bOPR	CD	Energy
Uniform object prior	0.642	–	–	–	–	–
No generation	–	0.000	0.000	0.000	–	–
Uniform random	–	–	0.001	0.039	–	3.42
Centroid	–	–	0.006	0.006	–	5.04
Frontier	–	–	0.003	0.003	–	10.07
Ours	**0.144**	**0.903**	**0.901**	**0.925**	1.81	**2.26**
GT oracle	0.000	1.000	1.000	1.000	0.00	0.00

VI. Results and Discussion

Table I shows that the zero-shot sampler is informative enough to support planning. It realizes the queried object in 90.3% of generated samples and achieves 90.1% mean OPR@1m from a single draw, far above the uniform random,

TABLE II: Planning evaluation over 173 object-reaching tasks. Bold indicates the best deployable method.

Method	$p_{\mathrm{tgt}}^{\mathrm{GT}}$	$p_{\mathrm{coll}}^{\mathrm{GT}}$	d_{tgt}	d_{traj}	χ_{GT}	C_f
Uniform	0.069	0.815	5.050	6.175	0.316	0.291
1-sample	0.041	0.971	6.034	6.695	0.049	0.043
Ours	**0.231**	**0.428**	**2.693**	**5.928**	**0.884**	**0.849**
GT oracle	1.000	0.075	1.098	0.000	0.925	0.925

TABLE III: Planning runtime in milliseconds over the same 173 tasks. Roadmap construction dominates total time.

Method	Total mean	Total median	Roadmap mean	A* mean
Uniform	1567.0	1518.7	1498.9	0.73
1-sample	1556.4	1490.8	1473.6	3.08
Ours	79333.0	78023.1	79239.6	10.54
GT oracle	1589.7	1517.6	1508.6	1.63

centroid, and frontier baselines. The improvement is not merely semantic: the lifted 3D worlds also reduce placement energy relative to non-generative baselines, indicating that target hypotheses concentrate near the correct regions while preserving useful diversity.

The planning results in Table II are the main finding. Planning over all 100 sampled worlds improves target-reaching success by more than $3\times$ over the uniform prior (0.231 vs. 0.069), cuts collision rate nearly in half (0.428 vs. 0.815), and greatly improves both collision-free path fraction and collision-free prefix fraction. The selected paths also end much closer to the ground-truth object and remain closer to the ground-truth oracle trajectory. The single-sample baseline is notably worse than even the uniform prior, confirming that a single completion is an unstable planning substrate. A plausible but inaccurate sample can pull the planner toward both incorrect geometry and incorrect target placement. By contrast, aggregating over many samples suppresses idiosyncratic errors and emphasizes consistent structure across hypotheses.

Runtime analysis in Table III shows that the additional cost of our method is concentrated in roadmap construction and edge validation rather than graph search. This is expected: the single-world baselines only need to evaluate each roadmap edge against one occupancy representation and one target hypothesis, whereas our sampled-world planner evaluates the same roadmap against all 100 occupancy–target hypotheses. Consequently, the single-world planners require less time for both roadmap construction and search.

Despite this larger sampled-world representation, the search remains fast. Once the roadmap and edge weights are available, A^* search requires only 10.54 ms on average for our 100-sample planner. The dominant cost instead comes from collision checking during roadmap construction and edge validation, which takes approximately 1.5 s per occupancy–target hypothesis in our implementation. This overhead arises because edge validity is evaluated by dense interpolation in continuous configuration space, with each interpolated configuration checked using FCL [31] through the OMPL [32] planning stack. Since collision checks across roadmap edges and sampled worlds are independent, this stage is naturally parallelizable. Incorporating vectorized or parallel collision-checking backends such as VAMP [39] could therefore substantially reduce the total runtime while preserving the same sampled-world planning objective.

Figure 3 provides a qualitative view of how the sampled-world planner behaves in object-reaching success cases. These examples are drawn from the subset of tasks in which the planned trajectory reaches the ground-truth target. They are selected to cover different queried objects, room layouts, and visibility conditions. For each task, the planner computes a trajectory using all the generated occupancy–target hypotheses, and the resulting trajectory is then overlaid on the ground-truth room and target occupancy for visual inspection.

The examples show that successful object-reaching plans often approach the target from nearby free space instead of terminating inside the target occupancy itself. This is important for object reaching: the terminal state should place the robot close enough to interact with or observe the object while still remaining collision-free. The trajectories also adapt to the structure of the room, producing longer entry paths when the target is deeper inside the scene. In partially visible cases, the generated samples provide relatively constrained target evidence, leading to direct and stable plans. In out-of-crop cases, the planner relies more strongly on semantic hypotheses across samples. When these hypotheses are consistent with the true layout, the roadmap search is guided toward useful terminal configurations. These visual results complement the quantitative results by showing how sampled generative worlds can induce task-directed paths in geometrically diverse scenes.

VII. Conclusion

We presented a planning-centric formulation of zero-shot generative scene completion for robots facing partially observed rooms. The key idea is to treat pretrained generative models as a conditional world sampler that maps doorway observations and object queries to explicit 3D world hypotheses. A sampled-world planner then chooses paths that minimize joint failure across those hypotheses, where failure means either geometric collision or unsuccessful object reaching. Across doorway-occluded indoor tasks, this combination yields safer and more goal-directed behavior than planning from either a uniform hidden-room prior or a single generated world. More broadly, the results suggest that sampled 3D generative priors can serve as practical robot world models for acting beyond the visible scene. The current system is offline and limited by collision checking cost across many generated worlds. Future work should integrate parallel collision checking, incremental replanning, active perception, and closed-loop execution so that sampled world models can be updated as the robot enters the room. More broadly, the work points toward a coevolution of generative world models and robotic reasoning: pretrained models can supply rich zero-shot hypotheses, while planners provide the structure needed to turn those hypotheses into safe, goal-directed action.

References

[1] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, no. 1–2, pp. 99–134, 1998.

[2] D. Silver and J. Veness, "Monte-Carlo Planning in Large POMDPs," in *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[3] H. Kurniawati, "Partially Observable Markov Decision Processes and Robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 253–277, 2022.

[4] J. van den Berg, P. Abbeel, and K. Goldberg, "LQG-MP: Optimized Path Planning for Robots with Motion Uncertainty and Imperfect State Information," *The International Journal of Robotics Research*, vol. 30, no. 7, pp. 895–913, 2011.

[5] Z. Zeng, A. Röfer, and O. C. Jenkins, "Semantic Linking Maps for Active Visual Object Search (Extended Abstract)," in *International Joint Conference on Artificial Intelligence*, 2021, pp. 4864–4868.

[6] W. Yang, X. Wang, A. Farhadi, A. Gupta, and R. Mottaghi, "Visual Semantic Navigation Using Scene Priors," in *International Conference on Learning Representations*, 2019.

[7] J.-P. Töberg, A.-C. Ngonga Ngomo, M. Beetz, and P. Cimiano, "Commonsense Knowledge in Cognitive Robotics: A Systematic Literature Review," *Frontiers in Robotics and AI*, vol. 11, p. 1328934, 2024.

[8] L. Zhao, W. McClinton, A. Curtis, N. Kumar, T. Silver, L. P. Kaelbling, and L. L. S. Wong, "Seeing is Believing: Belief-Space Planning with Foundation Models as Uncertainty Estimators," *arXiv preprint arXiv:2504.03245*, 2025.

[9] K. Hauser, "The Minimum Constraint Removal Problem with Three Robotics Applications," *The International Journal of Robotics Research*, vol. 33, no. 1, pp. 5–17, 2014.

[10] H. Lu, H. Kurniawati, and R. Shome, "Sampling-Based Motion Planning for Optimal Probability of Collision under Environment Uncertainty," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024, pp. 3138–3145.

[11] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to Control: Learning Behaviors by Latent Imagination," in *International Conference on Learning Representations*, 2020.

[12] N. A. Hansen, H. Su, and X. Wang, "Temporal Difference Learning for Model Predictive Control," in *International Conference on Machine Learning*, 2022, pp. 8387–8406.

[13] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "OccWorld: Learning a 3D Occupancy World Model for Autonomous Driving," in *European Conference on Computer Vision*, 2024.

[14] J. Berg, C. Zhu, Y. Bao, I. Durugkar, and A. Gupta, "Semantic World Models," *arXiv preprint arXiv:2510.19818*, 2025.

[15] C. Zhu, R. Yu, S. Feng, B. Burchfiel, P. Shah, and A. Gupta, "Unified World Models: Coupling Video and Action Diffusion for Pretraining on Large Robotic Datasets," in *Robotics: Science and Systems*, 2025.

[16] J. Zhang, M. Jiang, N. Dai, T. Lu, A. Uzunoglu, S. Zhang, Y. Wei, J. Wang, V. M. Patel, P. P. Liang, D. Khashabi, C. Peng, R. Chellappa, T. Shu, A. Yuille, Y. Du, and J. Chen, "World-In-World: World models in a closed-loop world," in *International Conference on Learning Representations*, 2026.

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis With Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp. 10 684–10 695.

[18] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. G. Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 36 479–36 494.

[19] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis," in *International Conference on Learning Representations*, 2024.

[20] Black Forest Labs, "FLUX.1 Fill [dev]: Image Inpainting and Outpainting Model," 2024, model release.

[21] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-VL: Technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[22] S. S. Bhattacharjee, D. Campbell, and R. Shome, "Believing is Seeing: Unobserved object detection using generative models," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, June 2025, pp. 19 366–19 377.

[23] ——, "FlatLands: Generative floormap completion from a single egocentric view," *arXiv preprint arXiv:2603.16016*, 2026.

[24] L. E. Kavraki, P. Švestka, J.-C. Latombe, and M. H. Overmars, "Probabilistic Roadmaps for Path Planning in High-Dimensional Configuration Spaces," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 4, pp. 566–580, 1996.

[25] Black Forest Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne, T. Dockhorn, J. English, Z. English, P. Esser, S. Kulal, K. Lacey, Y. Levi, C. Li, D. Lorenz, J. Müller, D. Podell, R. Rombach, H. Saini, A. Sauer, and L. Smith, "FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space," *arXiv preprint arXiv:2506.15742*, 2025.

[26] A. Bochkovskiy, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun, "Depth Pro: Sharp Monocular Metric Depth in Less Than a Second," in *International Conference on Learning Representations*, 2025.

[27] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene Parsing through ADE20K Dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5122–5130.

[28] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 12 077–12 090.

[29] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[30] C. C. Kemp, A. Edsinger, H. M. Clever, and B. Matulevich, "The Design of Stretch: A Compact, Lightweight Mobile Manipulator for Indoor Human Environments," in *IEEE International Conference on Robotics and Automation*, 2022, pp. 3150–3157.

[31] J. Pan, S. Chitta, and D. Manocha, "FCL: A General Purpose Library for Collision and Proximity Queries," in *IEEE International Conference on Robotics and Automation*, 2012, pp. 3859–3866.

[32] I. A. Şucan, M. Moll, and L. E. Kavraki, "The Open Motion Planning Library," *IEEE Robotics & Automation Magazine*, vol. 19, no. 4, pp. 72–82, 2012.

[33] A. X. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D data in indoor environments," in *International Conference on 3D Vision*, 2017, pp. 667–676.

[34] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 605–613.

[35] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[36] G. J. Székely and M. L. Rizzo, "Energy Statistics: A Class of Statistics Based on Distances," *Journal of Statistical Planning and Inference*, vol. 143, no. 8, pp. 1249–1272, 2013.

[37] B. Yamauchi, "A Frontier-Based Approach for Autonomous Exploration," in *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, 1997, pp. 146–151.

[38] H. Alt and M. Godau, "Computing the Fréchet Distance between Two Polygonal Curves," *International Journal of Computational Geometry & Applications*, vol. 5, no. 1–2, pp. 75–91, 1995.

[39] W. Thomason, Z. Kingston, and L. E. Kavraki, "Motions in microseconds via vectorized sampling-based planning," in *IEEE International Conference on Robotics and Automation*, 2024, pp. 8749–8756.