EchoNav: Evidential BEV Completion from Historical Observations for Visual Navigation

Anonymous Authors

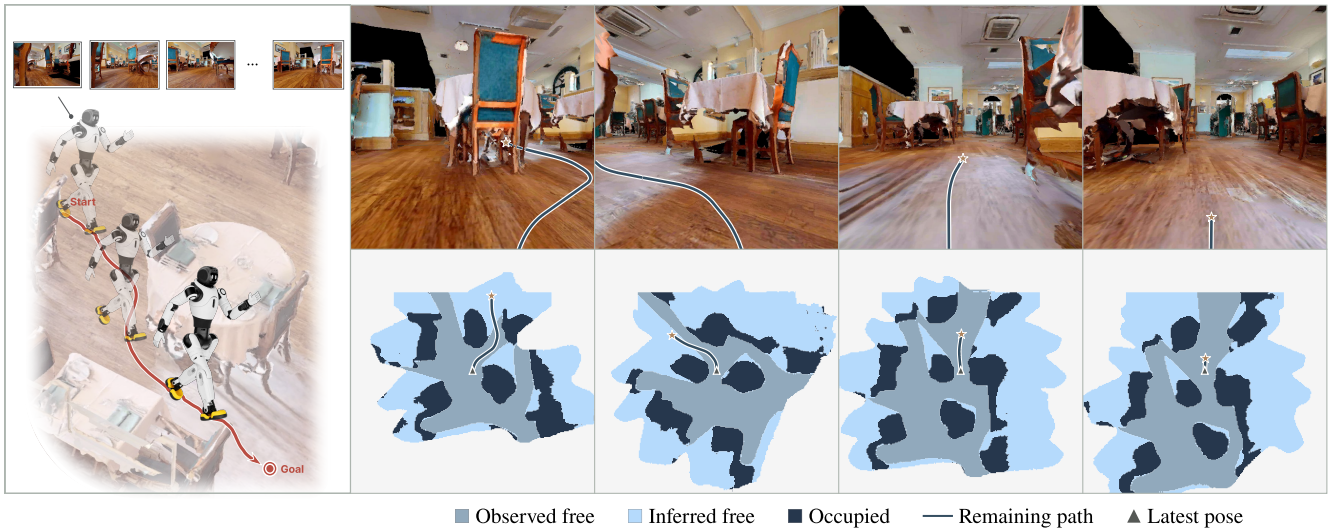


Fig. 1. **EchoNav: temporal BEV mapping and evidential occupancy completion for visual navigation.** The predictor integrates RGB history to retain previously observed structure and infer hidden obstacles and free space. Paths are planned on the predicted BEV maps and projected onto RGB views. Off-white denotes regions outside the input camera-footprint union.

***Abstract*— Visual navigation in cluttered environments requires robots to recover previously observed structure and infer obstacles and free space that remain occluded. We present a framework that combines temporal aggregation and occupancy completion in an evidential bird's-eye-view (BEV) representation predicted from short RGB sequences. A decoder adapts frozen VGGT-Omega features and combines historical residual proposals using cellwise weights and a learned no-update option. The BEV predictor produces a local map in the latest ego frame without depth or camera-pose inputs or explicit geometric warping. The evidential model treats observed free space separately and predicts occupancy with associated confidence across the remaining supported region. We train the predictor on simulated RGB sequences paired with complete and visibility-masked occupancy annotations. Our experiments evaluate navigation in simulation and on a wheeled robot, with additional demonstrations on quadruped and humanoid robots. In simulated sequential two-goal navigation, our system achieves 86.3% success, outperforming the strongest evaluated baseline by 32.2 percentage points. Across 60 wheeled-robot trials in 20 scenes, it achieves 73.3% success, compared with 33.3% for the strongest evaluated baseline.**

I. Introduction

Mobile robots navigating cluttered environments must make spatial decisions from incomplete visual observations (Fig. 1). Each camera view reveals only part of the surrounding layout. Relying only on the latest view can cause a robot to revisit unproductive routes or become trapped in local minima and fail to reach its goal. Observations retained within the RGB window provide evidence about nearby regions that are no longer visible in the current image. The more complete map gives the planner additional local context to choose more efficient paths. Yet some obstacles and free space remain hidden behind furniture and walls throughout this process. Predicting these hidden regions can inform planning, but the robot must account for the uncertainty of those predictions. The challenge is therefore to build a local BEV map that retains useful information across time, distinguishes observed space from inferred structure, and indicates how reliable its predictions are.

Learned occupancy anticipation has demonstrated that visual context can help predict unseen structure [1]. In autonomous driving, occupancy completion has shown strong results in scene understanding by predicting occupancy in both visible and occluded regions [2], [3]. For mobile robots, anticipating obstacles and free space beyond current observations can similarly provide a more complete map for planning, supporting exploration and navigation before those regions are directly observed [1], [4].

BEV occupancy maps place obstacles, free space, and candidate paths in a common top-down frame. This representation makes the basis of motion decisions transparent. The map describes the environment, while the planner and controller handle robot footprints, clearance margins, and motion constraints. This separation leaves the choice of planning algorithm open to the task and robot platform.

To address these challenges, we propose an RGB-based

framework for temporal BEV mapping and occupancy completion. It combines evidence from earlier observations with predictions of persistently hidden obstacles and free space, producing occupancy and reliability estimates for local planning. The BEV predictor requires neither depth nor camera-pose inputs and uses no explicit geometric warping.

The main contributions are:

1) **Spatially selective temporal BEV fusion.** We propose a framework that integrates RGB histories into an ego-centric occupancy BEV. Historical residual proposals from a shared decoder are weighted independently at each map location. A learned no-update option allows each cell to retain its initial features.
2) **Observation-aware evidential occupancy completion.** We develop an occupancy-learning formulation that models observed free space and inferred occupancy through separate components. It combines spatial support with occupancy uncertainty to estimate confidence for planning.
3) **Navigation integration and physical validation.** We integrate predicted occupancy and reliability with established local motion planners. Experiments show higher navigation success rates than the evaluated baselines in simulation and wheeled-robot trials. Additional quadruped and humanoid demonstrations illustrate deployment on other embodiments.

II. Related Work

A. Visual Navigation and Map-Based Planning

Learned visual navigation policies predict motion from visual observations and goal specifications. ViNT [5] learns from diverse robot trajectories, while NoMaD [6] uses a diffusion policy for goal-directed navigation and exploration. Model-based action reannotation [7] improves supervision for policy learning, and OmniVLA [8] supports goals specified through images, spatial coordinates, and language. X-NavDP [9] learns embodiment-specific behavior through robot-identity embeddings and reinforcement learning. Despite demonstrated cross-embodiment transfer, their behavior can remain sensitive to robot size, motion capabilities, and camera configurations represented in training [5].

Explicit traversability methods connect visual predictions to local planning. GeNIE [10] projects single-frame traversability into a BEV cost map and identifies the absence of a maintained local map as a source of local-minimum failures.

In autonomous driving, camera-based BEV methods provide spatial representations for downstream perception and planning. Lift-Splat-Shoot [11] and BEVFormer [12] combine RGB features from multiple cameras in a common BEV representation. These maps provide a spatial basis for navigation. More recent methods such as ViGT [13] predict continuous occupancy from surround-view RGB images without requiring camera calibration. The wider view helps planners identify obstacles and choose clear routes.

These methods motivate our use of a learned BEV representation. Our system supplies occupancy and reliability to motion planners. The robot footprint and traversal costs can be adjusted at planning time without changing the learned BEV predictor.

B. Occupancy Prediction and Anticipation

Occupancy anticipation addresses a further requirement for indoor navigation: inferring obstacles and free space hidden by furniture and walls. Map-based completion methods operate on partially observed top-down maps and predict occupancy in their unknown regions [4]. Geometry-guided semantic scene completion methods predict volumetric occupancy and semantic categories from RGB images using supplied camera geometry [14]. Navigation-oriented methods connect hidden-space prediction with exploration and planning, including occupancy anticipation in OccAnt [1] and confidence-aware semantic completion in SSCNav [15].

We focus on local occupied/free completion in regions that remain unobserved across the input views. Our evidential formulation models observed free space separately from inferred occupancy and accounts for spatial support when estimating prediction confidence.

C. Temporal Visual Geometry and Spatial Memory

Spatial memory retains information from past observations to support navigation [1], [16]. Dense mapping has long pursued this objective through active ranging, as in LiDAR-inertial systems such as FAST-LIO2 [17]. Cameras offer rich visual information without a dedicated range sensor, but recovering structure and motion from images depends on reliable correspondence and alignment. Weak texture, motion blur, and rapid viewpoint changes can challenge this estimation process [18]. Learned visual SLAM improves robustness by combining learned correspondence with geometric optimization, as in DROID-SLAM [18]; feed-forward models such as VGGT-Omega [19] further learn geometry across multiple images.

Using visual history for navigation requires combining observations from different viewpoints into a consistent spatial representation. Existing approaches use ego-motion or camera poses to align historical features and observations [12], [15]. Camera intrinsics and relative poses also provide geometric information for conditioning visual features [20]. These mechanisms depend on the supplied geometry to establish spatial relationships across views. Errors in this information can misalign historical features or maps, introducing false obstacles or distorting passage clearance and free-space connectivity.

Our work adopts the idea of learning spatial relationships across views to selectively integrate RGB history into a local BEV anchored to the latest ego frame, retaining evidence of previously observed structure. The BEV predictor decodes this history without depth or camera-pose inputs or explicit geometric warping.

III. Methodology

A. System Overview

Given an ordered RGB window $\mathcal{I}_t = (I_{t-T+1}, \ldots, I_t)$ of T observations from a moving camera, we seek a local

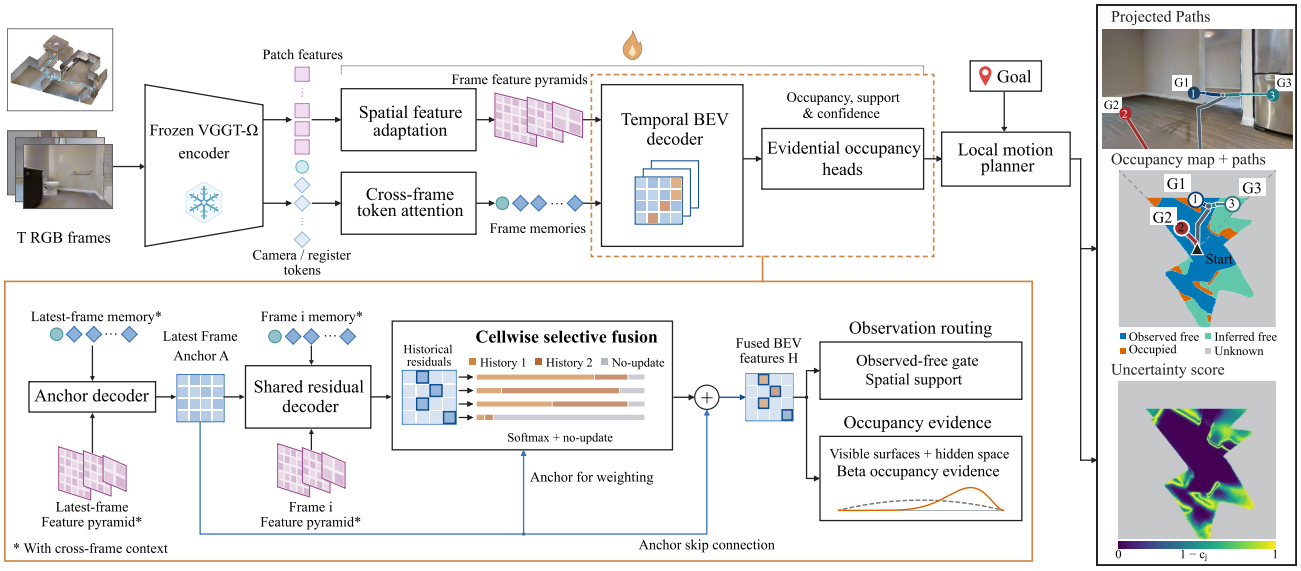


Fig. 2. **Framework overview.** A frozen VGGT-Omega encoder processes RGB history. Historical residuals refine a latest-frame BEV anchor through cellwise selective fusion with a learned no-update option. Observation-aware evidential heads predict occupancy, spatial support, and confidence for local planning. Right: predicted BEV with planned paths, path projections onto the latest RGB view, and uncertainty $1 - c_j$.

occupancy representation on a planar BEV grid $\mathcal{B}_t$ anchored to the latest ego frame. We learn the mapping

$$f_\theta(\mathcal{I}_t) = \{(p_j, g_j, s_j, c_j)\}_{j \in \mathcal{B}_t}, \tag{1}$$

where θ denotes the trainable parameters, and p_j, g_j, s_j, and c_j are the occupancy probability, observed-free probability, spatial-support probability, and prediction-confidence score at cell j, respectively. We use a 256×256 grid spanning 10×10 m, with metric extent fixed by the training targets.

Figure 2 summarizes the pipeline: a frozen VGGT-Omega encoder [19] extracts joint image features; a trainable decoder constructs a latest-frame BEV anchor and fuses historical residuals with cellwise weights (Sec. III-B); observation-aware heads produce occupancy, support, and confidence (Sec. III-C). Complete and visibility-masked targets supervise the predictor (Sec. III-D), whose outputs are supplied to a motion planner (Sec. III-E). The BEV predictor takes RGB images without depth or camera-pose inputs.

B. Selective Temporal BEV Decoding

1) Visual features and BEV anchor: To guide BEV queries in extracting image evidence across views, we use complementary spatial features and token context. The frozen VGGT-Omega encoder jointly processes the RGB window, with the latest image as the reference; we index this frame by $i = 0$ and historical frames by $i > 0$. Trainable adapters form multiscale image pyramids F_i^ℓ from patch features, providing spatially indexed content for sampling. Cross-frame attention over camera and register tokens produces frame-associated memories M_i, which condition the sampling queries with context from the window.

A learned readout of the frozen tokens produces positive frame weights ρ_i, normalized to mean one. These weights modulate each frame's influence when constructing the token memories: adding $\log \rho_i$ to attention logits for keys from frame i favors contributions from higher-weighted frames. The same weights provide a frame-level prior for the subsequent cellwise selection of historical residuals.

For cellwise temporal selection, historical proposals should refine the same starting representation. We therefore construct a common BEV anchor in the latest ego frame. Let Q denote the learned BEV queries with positional encodings, and let $\mathcal{R}_i(X) = \mathrm{Attn}(X, M_i)$ denote query attention to frame i's token memory. The anchor is

$$A = \mathcal{D}_0\big(Q + \mathcal{R}_0(Q), \{F_0^\ell\}_\ell\big)\,. \tag{2}$$

The anchor decoder samples the latest image pyramid using deformable cross-attention and learns image-to-BEV correspondence through BEV supervision without explicit camera projection. Subsequent historical branches propose residual refinements from this same anchor, allowing their contributions to be selected at each BEV location.

2) Historical residual proposals: A shared decoder generates historical residual proposals from the anchor A. Each proposal is conditioned on the corresponding frame's image pyramid and token context. The resulting residual is given by

$$C_i = \mathcal{R}_i(A), \qquad \Delta_i = \mathcal{D}_h\big(A, \{F_i^\ell\}_\ell, C_i\big) - A. \tag{3}$$

The token context C_i conditions the queries used to sample the historical image pyramid, producing a feature residual Δ_i at each anchor cell. Residual-output projections are zero-initialized to preserve the anchor at initialization.

3) Spatially selective fusion: Historical views provide different evidence at different map locations, motivating cellwise control over both their relative contributions and the strength of their updates. We include a zero-residual candidate so that fusion can also control the total strength of the historical update.

For $T > 1$, a learned scorer combines the anchor A_j, residual Δ_{ij}, and token context C_{ij} at each cell. Adding

$\log \rho_i$ and a learned frame-age bias gives the historical score e_{ij}. A no-update candidate has score $e_{\varnothing j} = f_\varnothing(A_j) + \log(T-1) + \log[\pi_0/(1-\pi_0)]$, where $f_\varnothing$ is learned. The candidate weights are

$$w_{ij} = \frac{\exp(e_{ij})}{\sum_{k\in\mathcal{J}} \exp(e_{kj})}, \qquad \mathcal{J} = \{\varnothing, 1, \ldots, T-1\}. \quad (4)$$

The learned frame weight ρ_i provides a spatially shared prior, while w_{ij} determines that frame's contribution separately at each BEV cell.

With uniform initial frame weights and zero learned scoring terms, the bias gives an initial no-update probability of $\pi_0 = 0.9$. The fused representation is

$$H_j = A_j + \sum_{i=1}^{T-1} w_{ij}\Delta_{ij}. \quad (5)$$

A higher no-update weight leaves less weight for historical residuals, reducing their combined correction to the anchor. The decoder maintains no recurrent BEV state and returns $H = A$ when $T = 1$. This design allows historical evidence to refine individual BEV cells while retaining the anchor as the default representation.

C. Observation-Aware Evidential Occupancy

From the fused representation H, the routing branch predicts observed-free probability g_j and spatial-support probability s_j. The former estimates whether a cell is free and visible in any input frame. Spatial support covers the union of the input camera footprints wherever complete labels are valid, including occluded cells. The evidence branch models visible occupied surfaces and hidden free or occupied cells. This separation distinguishes observed free space from inferred free space, even when both yield a low occupancy probability.

Two logits parameterize a Beta distribution [21] over occupancy probability, with $\alpha_j = 1 + \mathrm{softplus}(z_j^o)$, $\beta_j = 1 + \mathrm{softplus}(z_j^f)$, and mean $\mu_j = \alpha_j/(\alpha_j + \beta_j)$. Conditional on the observed-free component, occupancy is modeled as zero. Mixing this component with the Beta distribution using weights g_j and $1 - g_j$ gives

$$p_j = (1 - g_j)\mu_j. \quad (6)$$

With Beta variance $v_j = \mu_j(1-\mu_j)/(\alpha_j + \beta_j + 1)$, the mixture variance and composite confidence are

$$V_j = (1 - g_j)v_j + g_j(1 - g_j)\mu_j^2, \quad (7)$$

$$c_j = s_j\,|2p_j - 1|\,\mathrm{clip}(1 - 4V_j, 0, 1). \quad (8)$$

The mixture variance captures spread within the evidence distribution and disagreement between components. High confidence requires strong spatial support, a clear occupied/free preference, and low mixture variance. These factors reduce confidence where occupancy predictions are ambiguous or variable, while s_j suppresses confidence in cells with low predicted support even when one class is strongly favored. The resulting confidence is a constructed score rather than a calibrated collision probability.

D. Supervision and Optimization

Complete and visibility-masked targets are expressed in the latest ego frame. We denote spatial support by U, observed free space by $O \subseteq U$, and the remaining region by $G = U \setminus O$. These regions supervise spatial support s_j, observed-free routing g_j, and occupancy evidence, respectively.

For each cell $j \in G$, we minimize the Bernoulli negative log-likelihood averaged under the predicted Beta distribution, with $y_j = 1$ denoting an occupied cell:

$$\ell_j^E = \psi(\alpha_j + \beta_j) - y_j\psi(\alpha_j) - (1 - y_j)\psi(\beta_j), \quad (9)$$

where ψ is the digamma function. The occupancy objective is $\mathcal{L}_{\mathrm{occ}} = \langle \ell^E + \lambda_K \ell^K \rangle_w$. The KL regularizer suppresses evidence for the incorrect class. It compares the predicted Beta distribution to the uniform $\mathrm{Beta}(1,1)$ prior after resetting the correct-class parameter to one. To balance regions of unequal area, we average the occupancy loss separately over hidden free cells, visible occupied surfaces, and hidden occupied cells in each sample. Their means are then combined using relative weights normalized over the regions present.

Class-balanced BCE defines $\mathcal{L}_{\mathrm{gate}}$ using positives in O and negatives in G, and $\mathcal{L}_{\mathrm{support}}$ using positives in U and negatives outside U. The resulting map objective jointly supervises observed-free routing, occupancy evidence, and spatial coverage:

$$\mathcal{L}_{\mathrm{BEV}} = \mathcal{L}_{\mathrm{gate}} + \mathcal{L}_{\mathrm{occ}} + \lambda_B \mathcal{L}_{\mathrm{support}}. \quad (10)$$

We use $\lambda_B = 0.5$ and a maximum regularization weight $\lambda_K = 0.005$. Hidden-occupied supervision and wrong-class evidence regularization are introduced gradually during training.

E. Inference and Planning

The planner uses the predicted BEV and a metric point goal, both expressed in the latest ego frame. Supported occupied cells are inflated according to the robot footprint and safety margin to form hard collision constraints. Traversal costs combine occupancy p_j, obstacle clearance, and distribution reliability $\kappa_j = s_j\,\mathrm{clip}(1 - 4V_j, 0, 1)$. Within supported traversable regions, lower reliability incurs a larger penalty. Unsupported cells retain their unknown labels and incur finite traversal penalties, subject to the collision constraints.

When the goal lies outside the map or the predicted reachable region, the planner selects a local subgoal within that region. Only a short path segment is executed, with an additional limit on travel through unknown space. The next RGB observation updates the BEV before replanning.

IV. Experiments

In this section, we evaluate the proposed representation from map prediction to navigation. We present our model training procedure in Sec. IV-A and the comparison with baselines in Sec. IV-B. Section IV-C presents the temporal fusion ablation. Sections IV-D and IV-E report simulation navigation and physical-robot evaluation, respectively.

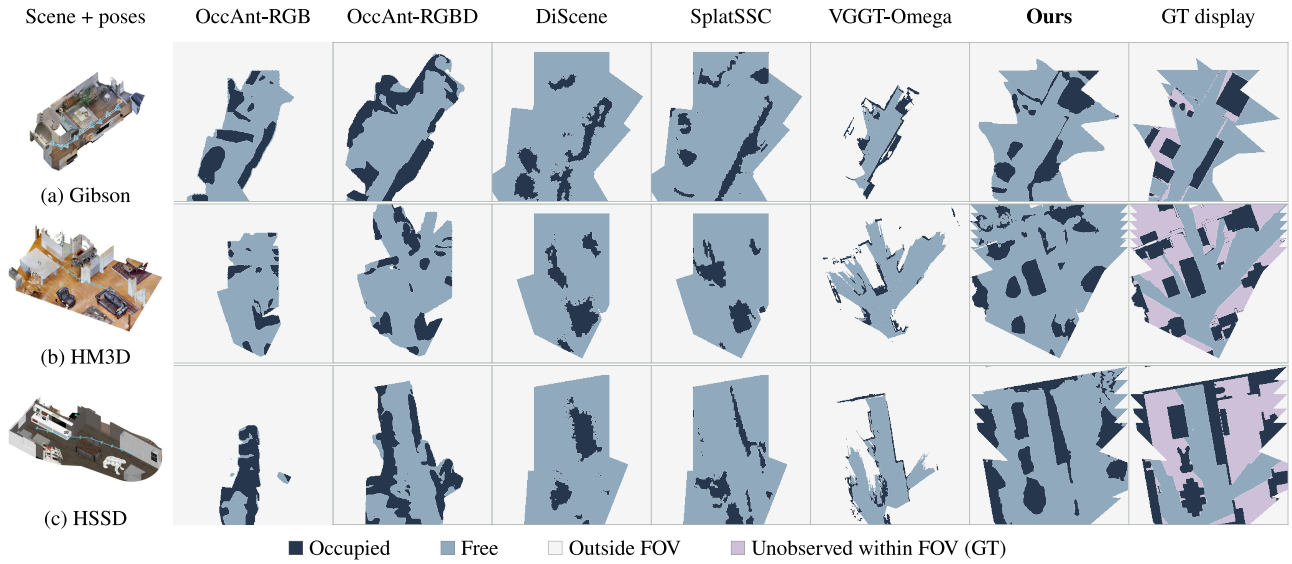


Fig. 3. **Representative BEV completion results.** Maps from image history are shown on a common grid for Gibson, HM3D, and HSSD scenes. Purple in GT highlights regions within the input camera-footprint union that remain unobserved across all views and require completion. Off-white denotes regions without an occupancy label.

A. Data Collection and Training

1) Data collection: We collect RGB sequences along sampled navigable paths in Habitat environments from HM3D [22], HSSD [23], and Matterport3D [24], varying camera height, field of view, and motion speed. The indoor corpus contains approximately 535k training sessions from 1k scenes and 26k validation sessions from 54 disjoint scenes. For each session, we generate complete and visibility-masked BEV targets in the latest ego frame, following the target definitions in Sec. III-D.

2) Training configuration: We freeze the pretrained VGGT-Omega encoder and optimize the BEV adapters, temporal decoder, and prediction heads. The decoder accepts variable-length RGB sequences. During training, BoQ-based visual similarity guides frame selection to reduce visual redundancy, with sequence lengths centered on ten frames. Predictions use a 256×256 BEV grid spanning 10×10 m. Hidden-occupied supervision and wrong-class evidence regularization are introduced gradually during training.

Gibson [25] and ProcTHOR [26] are unseen datasets used only for evaluation and are entirely excluded from our BEV training data.

B. Occupancy Prediction under Partial Observability

1) Evaluation settings: We evaluate 1,200 matched routes from 101 scenes in HM3D, HSSD, and Gibson, with 400 routes per dataset. Scenes used for our BEV training are excluded. Single-image prediction uses the final image of each ten-frame route on a forward-facing 3.25×3.25 m grid with 5 cm cells. History mapping uses all ten images on a 256×256 grid spanning 10×10 m. Both settings share the same final ego frame, and evaluation covers valid labels within the corresponding camera-footprint union.

We compare against DiScene [27], SplatSSC [14], VGGT-Omega [19], OccAnt-RGB, and OccAnt-RGBD [1]. DiScene and SplatSSC receive camera calibration; their volumetric predictions are projected into BEV and accumulated across views using ground-truth poses. For OccAnt, we bypass learned pose estimation and use ground-truth poses for inter-frame accumulation, removing odometry errors from the comparison. OccAnt-RGBD additionally receives sensor depth. VGGT-Omega maps are projected from predicted geometry and cameras using camera-height calibration. Our predictor receives only RGB images, without camera intrinsics or inter-frame extrinsics.

2) Metrics and filtering: Following Sec. III-D, O denotes observed free space and $G = U \setminus O$ the remaining supported region. We report free-space IoU in O and occupied/free mIoU in G and U. Unknown predictions count as errors in O and are excluded from G. Thus, mIoU_G measures accuracy conditional on a known prediction, while mIoU_U combines the regional counts under the same policy. We also report Cov_G, the fraction of G assigned an occupied or free label.

For our model, dense prediction labels every evaluated cell as occupied when $p_j \geq 0.5$ and free otherwise. Filtered prediction applies the same labeling rule only to cells satisfying $g_j \geq 0.5$ or $c_j \geq 0.4$; the remaining cells are marked unknown. OccAnt assigns occupied/free labels only to cells with explored probability > 0.6; the remaining cells are marked unknown.

3) Occupancy completion: Region G includes visible obstacle surfaces and regions that remain occluded across the input views, where free space and obstacles must be inferred. Figure 3 shows that VGGT-Omega leaves parts of the surrounding layout unknown, while our model predicts both obstacles and free space in these regions. For single-image prediction, our model achieves the highest reported mIoU_G among RGB-based methods, exceeding the strongest RGB baseline, VGGT-Omega, by 4.03 percentage points with dense prediction and 8.53 points after filtering. Dense

TABLE I

OCCUPANCY PREDICTION RESULTS

Method	mIoU_U	IoU_O	mIoU_G	Cov_G
Single image: 3.25×3.25 m				
DiScene	43.21	97.28	32.02	100.00
SplatSSC	40.58	94.97	31.11	100.00
VGGT-Omega	42.62	53.12	44.26	13.06
OccAnt-RGB	32.40	44.88	33.88	39.14
OccAnt-RGBD	63.96	81.57	**55.16**	56.32
Ours (dense)	57.32	**97.54**	48.29	100.00
Ours (filtered)	**64.72**	96.98	52.79	50.84
Ten-image history: 10×10 m				
DiScene	39.74	80.36	32.68	64.76
SplatSSC	37.71	75.97	32.12	64.76
VGGT-Omega	40.61	55.25	38.22	11.86
OccAnt-RGB	31.10	47.15	31.27	22.26
OccAnt-RGBD	55.82	80.14	46.72	42.51
Ours (dense)	55.27	**95.67**	47.91	100.00
Ours (filtered)	**68.32**	89.58	**61.17**	35.03

prediction covers all of G, compared with 13.06% coverage for VGGT-Omega. Our filtered predictions also achieve mIoU_G within 2.37 percentage points of depth-assisted OccAnt-RGBD, with coverage 5.48 percentage points lower. In history mapping, our dense predictions outperform all baselines, exceeding OccAnt-RGBD by 1.19 points while providing full coverage. Filtering increases this margin over the strongest baseline, OccAnt-RGBD, to 14.45 percentage points in mIoU_G at 35.03% coverage of G. These results support dense occupancy prediction beyond observed free space and a selective alternative with higher conditional accuracy at reduced coverage.

4) Temporal BEV mapping: Temporal mapping errors can produce fragmented occupied regions and displaced or duplicated boundaries, such as those seen in several baseline maps in Fig. 3. Such errors can falsely block passages or omit obstacles, leading to detours or increased collision risk. Our model predicts a unified BEV from RGB history in one forward pass, while the baselines accumulate predictions using predicted or ground-truth camera extrinsics. Our filtered predictions achieve the highest reported mIoU_U in both settings. Their advantage over the strongest baseline increases from 0.76 percentage points for single-image prediction to 12.50 points for history mapping. Dense prediction also shows a larger advantage over the strongest RGB baseline in the history setting. Its mIoU_U decreases by only 2.05 percentage points, compared with 8.14 points for OccAnt-RGBD. These results indicate that our model effectively retains and integrates information from past RGB observations to construct a more complete map.

C. Temporal Fusion Ablation

We examine temporal fusion on 190 examples, comprising 11 HM3D and 179 ProcTHOR scenes excluded from training. All variants use the same checkpoint. Anchor-only prediction omits historical residual updates while retaining historical context from joint encoding. Equal history weighting preserves the learned no-update weight at each cell and distributes the remaining weight equally among historical residuals.

All variants produce dense occupancy predictions (Table II). Recovery denotes cells visible in earlier observations but not in the latest view; Hidden denotes cells unobserved across all input views.

Full selective fusion raises overall mIoU from 47.59% for anchor-only prediction to 58.43%. Compared with equal history weighting, it improves mIoU by 1.90 percentage points overall, 2.35 in Recovery, and 2.25 in Hidden. These results demonstrate the contribution of learned historical weighting to temporal recovery and occupancy completion within the trained model.

TABLE II

TEMPORAL FUSION ABLATION.

Fusion variant (mIoU, %)	Overall	Recovery	Hidden
No history fusion	47.59	45.79	37.78
Equal-weight fusion	56.53	46.60	48.01
Full selective fusion	**58.43**	**48.95**	**50.26**

D. Simulation Navigation

1) Comparisons and metrics: We evaluate simulated wheeled-robot navigation against LoGoNav [7], OmniVLA [8], X-NavDP [9], and OccAnt-RGBD [1] in both the sequential and hidden-goal benchmarks. Ours and OccAnt use Dijkstra with a shared path executor. X-NavDP and OccAnt-RGBD receive sensor depth; the remaining methods use RGB only. OccAnt accumulates map predictions using ground-truth odometry. Goal acceptance requires a collision-free connection from the robot to the goal.

For N episodes, let $\zeta_n \in \{0, 1\}$, P_n, and L_n denote success, executed distance, and precomputed ground-truth reference path length, respectively. We compute L_n by summing edge lengths along the precomputed reference path in the ground-truth navigability graph. Goal tolerance and continuous motion allow successful trajectories to be shorter than this discrete reference, yielding SCPR below one. We report success rate (SR), SPL [28], and success-conditioned path ratio (SCPR):

$$\text{SR} = \frac{1}{N}\sum_n \zeta_n, \qquad \text{SCPR} = \frac{\sum_n \zeta_n P_n / L_n}{\sum_n \zeta_n},$$
$$\text{SPL} = \frac{1}{N}\sum_n \zeta_n \frac{L_n}{\max(L_n, P_n)}. \tag{11}$$

Failures contribute zero to SPL. SCPR is averaged over successful episodes; methods with no successful episodes are excluded from the SCPR comparison.

2) Sequential two-goal navigation: The sequential two-goal navigation benchmark comprises 1,000 shared routes across 308 scenes excluded from training, with 250 routes each from Gibson, HSSD, HM3D, and ProcTHOR. The second destination is revealed only after the first is reached (Fig. 5). The current waypoint comes from a precomputed ground-truth route and is expressed relative to the robot at planning time. Ground-truth occupancy maps and future waypoints are not provided. Success requires reaching both destinations in order without collision.

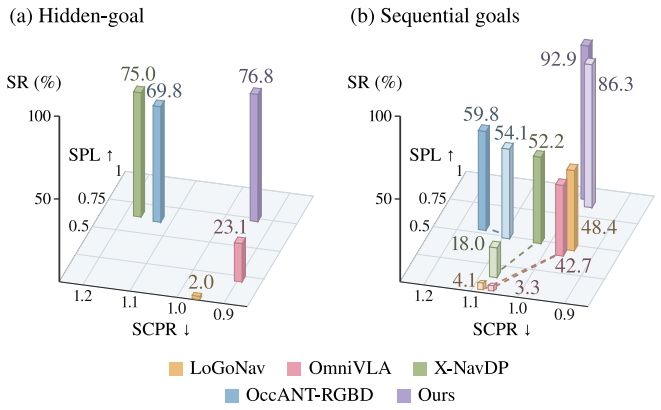


Fig. 4. Simulation navigation results for (a) hidden-goal and (b) sequential two-goal tasks. Bar height indicates SR; ground-plane coordinates indicate SPL and SCPR. In (b), darker and lighter bars represent reaching the first goal and both goals, respectively.

For the first goal, Ours achieves an SR of 92.9% and an SPL of 0.922, exceeding the strongest remaining baseline, OccAnt-RGBD, by 33.1 percentage points in SR and 0.370 in SPL. Ours and OccAnt account for the robot's footprint and clearance. Both show stronger performance in guiding the robot along efficient, collision-free paths.

After the goal changes, Ours maintains 86.3% SR and 0.860 SPL. Its SR decreases by only 6.6 percentage points, compared with decreases of 5.7 to 44.3 points for the baselines. Figure 5 shows two challenging navigation cases where only our method reaches both goals. OccAnt fails due to a missed obstacle boundary in Case 1 and overestimated obstacle extent in Case 2. Our maps more accurately represent the obstacles and surrounding free space, allowing the planner to find routes to both goals.

Our method achieves a first-goal SCPR of 0.983, comparable to LoGoNav's 0.951. It completes nearly twice as many episodes as LoGoNav. It also succeeds in challenging scenarios with similar path efficiency. For the complete trajectory, our SCPR of 0.967 is the lowest among all evaluated methods. Only our method and OccAnt-RGBD show lower SCPR over the complete trajectory, while the other baselines show increases. This pattern provides the evidence that retained historical observations can help guide more efficient path planning as navigation continues.

3) Navigation toward initially hidden destinations: To assess navigation beyond initially observed free space, we construct 1,000 tasks in 503 existing ProcTHOR houses. Each task specifies a reachable goal 1–3 m from the start, occluded by an object visible in the initial image. The relative goal is supplied once, and navigation proceeds with subsequent observations without intermediate waypoints.

Our method achieves the highest SR and SPL among all evaluated methods, with a substantial advantage over the other RGB-based approaches (Fig. 4). It also exceeds the depth-assisted baselines on both metrics, demonstrating effective navigation toward initially hidden destinations using RGB observations alone.

Our method achieves substantially lower SCPR than X-NavDP and OccAnt-RGBD, the two strongest baselines in SR and SPL. OmniVLA achieves a lower SCPR than ours, but succeeds in far fewer episodes. Our method combines high navigation success with efficient paths toward initially hidden goals.

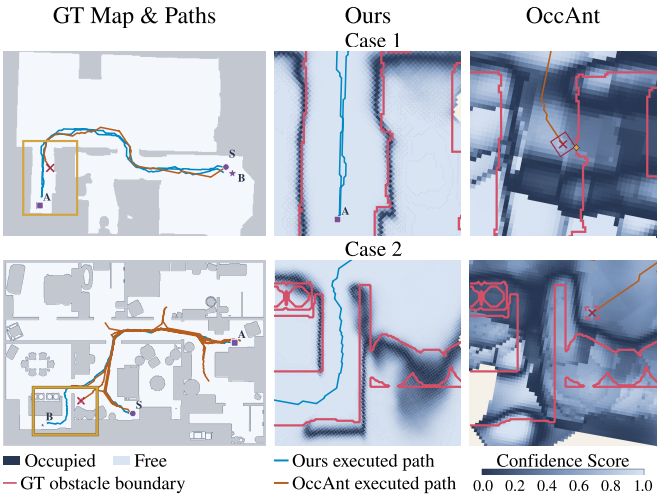


Fig. 5. Sequential two-goal navigation comparing Ours and OccAnt. Yellow boxes mark the enlarged regions; S denotes the start and A/B the goals.

E. Physical-Robot Evaluation

We deploy our framework on a Mecanum-wheeled robot with an Intel RealSense D435 camera. Our BEV predictor processes temporal RGB observations, and Dijkstra search plans paths on the predicted occupancy map. We evaluate Ours, X-NavDP, and OmniVLA in 20 scenes, with three matched trials per scene. Success requires reaching within 30 cm of the goal. A trial fails if the robot collides with an obstacle or the planner cannot find a feasible path. Figure 6 shows representative scenes from our physical-robot experiments. The frosted-glass doorway in (a) restricts visibility into the adjacent space, while (b) requires planning through a doorway toward a goal outside the room. Very low illumination in (c) reduces the visual detail available for perception.

Our method achieves a success rate of 73.3%, compared with 33.3% for X-NavDP and 26.7% for OmniVLA. It completes more than twice as many trials as the strongest evaluated baseline, X-NavDP. These results demonstrate the practical value of integrating temporal BEV mapping and occupancy completion with conventional path planning, enabling successful physical navigation using a spatial representation predicted from RGB observations. We further demonstrate navigation on humanoid and quadruped robots. These qualitative demonstrations extend the framework's deployment to platforms with different locomotion capabilities.

V. Conclusion

We presented an RGB-based local navigation framework combining selective temporal BEV fusion with observation-aware evidential occupancy completion. The predictor integrates historical observations and infers occluded obstacles

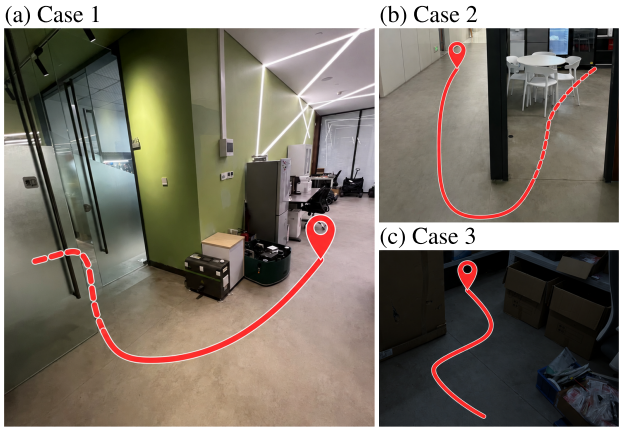


Fig. 6. Physical-robot navigation examples in three indoor scenes, with routes overlaid in red.

and free space without depth or camera-pose inputs. It distinguishes observed free space from inferred occupancy and provides confidence for local planning. The integrated system achieves the highest success rate and SPL in both simulated navigation benchmarks and the highest success rate in matched wheeled-robot trials. Additional qualitative demonstrations extend deployment to humanoid and quadruped robots. These results establish BEV maps that integrate RGB history and complete occluded regions as an effective interface between perception and local planning under partial observability.

REFERENCES

[1] S. K. Ramakrishnan, Z. Al-Halah, and K. Grauman, "Occupancy anticipation for efficient exploration and navigation," in *European conference on computer vision*. Springer, 2020, pp. 400–418.

[2] A.-Q. Cao and R. De Charette, "Monoscene: Monocular 3d semantic scene completion," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 3981–3991.

[3] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 9087–9098.

[4] V. D. Sharma, J. Chen, and P. Tokekar, "Proxmap: Proximal occupancy map prediction for efficient indoor robot navigation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7135–7140.

[5] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine, "ViNT: A foundation model for visual navigation," in *7th Annual Conference on Robot Learning*, 2023.

[6] A. Sridhar, D. Shah, C. Glossop, and S. Levine, "Nomad: Goal masked diffusion policies for navigation and exploration," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 63–70.

[7] N. Hirose, L. Ignatova, K. Stachowicz, C. Glossop, S. Levine, and D. Shah, "Learning to drive anywhere with model-based reannotation," *IEEE Robotics and Automation Letters*, vol. 11, no. 2, pp. 1242–1249, 2025.

[8] N. Hirose, C. Glossop, D. Shah, and S. Levine, "Omnivla: An omni-modal vision-language-action model for robot navigation," in *2026 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2026, pp. 12 651–12 658.

[9] T. Yang, Y. Zeng, W. Cai, Y. Yang, J. Peng, H. Cheng, J. Pang, and T. Wang, "X-NavDP: Generalizing navigation diffusion policy to novel behavior and embodiments with group Q-score reweighted matching," in *Conference on Robot Learning (CoRL)*, 2026.

[10] J. Wang, D. Liu, J. Chen, J. Da, N. Qian, M. M. Tram, and H. Soh, "Genie: A generalizable navigation system for in-the-wild environments," *IEEE Robotics and Automation Letters*, 2025.

[11] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *European conference on computer vision*. Springer, 2020, pp. 194–210.

[12] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 3, pp. 2020–2036, 2024.

[13] A. Shirokov, M. Kuznetsov, D. Stepochkin, E. Evdokimov, D. Glazkov, N. Patakin, A. Konushin, and D. Senushkin, "Visual implicit geometry transformer for autonomous driving," in *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, D. Calvanese, Ed. International Joint Conferences on Artificial Intelligence Organization, Aug. 2026, pp. 1604–1612, main Track.

[14] R. Qian, H. Cao, T. Deng, S. Yuan, and L. Xie, "Splatssc: Decoupled depth-guided gaussian splatting for semantic scene completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 10, 2026, pp. 8520–8528.

[15] Y. Liang, B. Chen, and S. Song, "Sscnav: Confidence-aware semantic scene completion for visual semantic navigation," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 13 194–13 200.

[16] J. Wang, J. Chen, D. Liu, A. Ghotavadekar, J. Da, L. Kästner, and H. Soh, "Change-robust online topological memory for long-term relocalization and semantic navigation," in *Advances in neural information processing systems*, 2026.

[17] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, "Fast-lio2: Fast direct lidar-inertial odometry," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2053–2073, 2022.

[18] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," in *Advances in neural information processing systems*, vol. 34, 2021, pp. 16 558–16 569.

[19] J. Wang, M. Chen, S. Zhang, N. Karaev, J. Schönberger, P. Labatut, P. Bojanowski, D. Novotny, A. Vedaldi, and C. Rupprecht, "VGGT-Ω," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[20] L. Xu, X. Wang, Z. Huang, H. Li, Y. Yang, N. Wang, and S. Liu, "Geometry-grounded unified 3d perception for autonomous driving," arXiv preprint arXiv:2608.13147, 2026.

[21] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Advances in neural information processing systems*, vol. 31, 2018.

[22] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. Chang, M. Savva, Y. Zhao, and D. Batra, "Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung, Eds., vol. 1, 2021.

[23] M. Khanna, Y. Mao, H. Jiang, S. Haresh, B. Shacklett, D. Batra, A. Clegg, E. Undersander, A. X. Chang, and M. Savva, "Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 16 384–16 393.

[24] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," in *2017 International Conference on 3D Vision (3DV)*, 2017, pp. 667–676.

[25] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson env: Real-world perception for embodied agents," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 9068–9079.

[26] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, K. Ehsani, J. Salvador, W. Han, E. Kolve, A. Kembhavi, and R. Mottaghi, "Procthor: Large-scale embodied ai using procedural generation," in *Advances in neural information processing systems*, vol. 35, 2022, pp. 5982–5994.

[27] X. Li, Y. Zheng, P. Li, Y. Chen, Y.-Q. Zhang, and W. Ding, "Enhancing indoor occupancy prediction via sparse query-based multi-level consistent knowledge distillation," *IEEE Robotics and Automation Letters*, 2025.

[28] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, *et al.*, "On evaluation of embodied navigation agents," *arXiv preprint arXiv:1807.06757*, 2018.