HYPOTHESIS-INDUCED ACTION FIELDS FOR OBJECT LOCALIZATION IN PARTIAL SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Object localization in partial scenes requires inferring the position of an unobserved object from incomplete spatial and semantic context. Commonsense relations provide contextual information about potential target locations, but incomplete observations may remain compatible with several spatially separated alternatives. Compressing these alternatives into point estimates can obscure the underlying ambiguity, and aggregating these estimates may yield a prediction with limited support from the observed context. We introduce Hypothesis-Induced Action Fields (HIAF), a method that couples the representation of spatial alternatives with the localization decision. Spatial hypotheses derived from observed objects jointly define an action field that estimates the probability of successful localization at different locations. Learning this field aligns hypothesis generation and scoring with the localization objective, allowing alternatives to inform the final prediction collectively. On Partial ScanNet, HIAF achieves the highest mean success rates and lowest mean successful localization errors among the evaluated methods across three evaluation protocols. At 1 m, query-level success reaches 30.06%, exceeding the strongest evaluated baseline by 4.86 percentage points. At the stricter 0.5 m tolerance, HIAF achieves 13.19% query-level success, compared with 9.54% for the strongest baseline at that tolerance.

1 INTRODUCTION

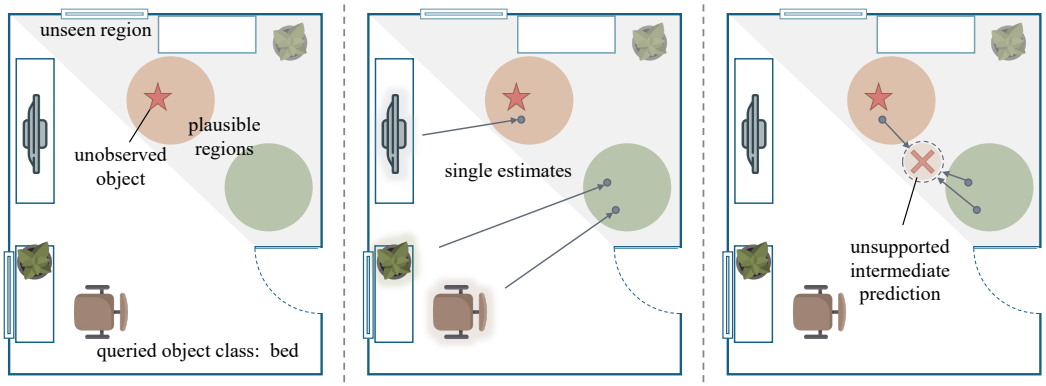
Object localization in partial scenes concerns inferring the location of an unobserved object from incomplete observations. In embodied perception and assistive robotics, agents often need to decide where to search before a target becomes visible (Martinez-Martin & Del Pobil, 2017; Chaplot et al., 2020; Batra et al., 2020). Objects already in view provide spatial and semantic clues, but occlusion leaves much of the surrounding context unresolved. Several separated locations may remain plausible, and more than one instance of the queried class may be hidden in the scene.

Methods for this task range from MLP with Commonsense (MLPC) and Graphormer-based localization to the spatial commonsense graphs used by SCG-OL and D-SCG-OL (Ying et al., 2021; Giuliari et al., 2022; 2023). SCG-OL predicts pairwise distances before recovering a position, whereas D-SCG-OL learns directed relative positions end to end. Although these methods use different scene representations, they train spatial estimates with error-based objectives. A richer representation can refine those estimates, but the task also raises a question about what to estimate when only one location can be returned. Minimizing coordinate error need not favor the location most likely to satisfy the evaluation criterion. Even in the single-instance case, squared-error coordinate regression targets the conditional mean, which can fall between separated regions supported by the observed context. The resulting prediction may lie in an unsupported intermediate location (Figure 1).

The established evaluation criterion for this task allows a spatial tolerance. A prediction succeeds if it lies within a specified distance of any hidden instance of the queried class. We use this existing criterion to define the estimation target, namely the probability that a prediction at each location succeeds. This probability provides a common target for single- and multi-instance scenes, since any hidden instance is an admissible match. It is also the expected reward of choosing that location, so maximizing it yields a Bayes action for the task.

We introduce Hypothesis-Induced Action Fields (HIAF) to estimate this success probability as a continuous field over the ground plane. A target-conditioned relational encoder supplies context for

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.



(a) Ambiguity under partial observation (b) Object-conditioned point estimates (c) Averaging between plausible regions

Figure 1: Why the estimation target matters. (a) In a partially observed scene, the observed context can support several separated regions for a hidden instance of the queried class. (b) A single location estimate per observed object compresses the spatial alternatives supported by that object. (c) Averaging separated estimates can place the prediction between supported regions, outside the success radius of a hidden instance.

Gaussian hypotheses anchored at observed objects. These hypotheses jointly define the action field and provide candidate locations for the final prediction. The full hidden target set labels where a prediction would succeed, providing binary supervision over the ground plane. At inference, HIAF evaluates the field at the hypothesis centers and returns the location with the highest estimated success probability.

We evaluate HIAF on Partial ScanNet using three protocols that differ in query selection and outcome aggregation. The main contributions of this work are summarized as follows:

- We formulate partial-scene localization in terms of success probabilities within the evaluation tolerance, providing a common estimation target for single- and multi-instance queries.
- We propose HIAF, which combines anchor-conditioned Gaussian hypotheses into a continuous action field. Each hypothesis supplies a candidate location, whose success score incorporates contributions from all hypotheses.
- We derive a full-plane binary cross-entropy objective using the full hidden target set. An analytic background integral and sampling with overlap correction enable tractable end-to-end training.
- Experiments on Partial ScanNet show that, among the evaluated methods, HIAF achieves the highest mean success rate and lowest mean error among successful predictions across three protocols. The success-rate gains extend to both single- and multi-instance queries.

2 Related Work

Object localization in partial scenes. MLP with Commonsense (MLPC) and SCG-OL estimate distances between an unobserved target and observed objects, then recover the target position through a separate localization stage (Giuliari et al., 2022). MLPC uses a graph convolutional network (Kipf & Welling, 2017) to aggregate commonsense information before predicting pairwise distances with an MLP. SCG-OL also incorporates the spatial relations among observed objects. D-SCG-OL replaces distance prediction with directed relative position regression, combining object-conditioned estimates within an end-to-end model (Giuliari et al., 2023). Graphormer (Ying et al., 2021) extends the Transformer architecture to graph representation learning through structural encodings. For object localization, Giuliari et al. (2023) use Graphormer to update node features and regress the target position from the target node representation. These methods learn spatial distances or coordinates; HIAF instead learns the probability of success within the task's evaluation radius. Related formulations include relation prediction for unobserved-object localization in RP-SG (Ying et al., 2024) and map-based target localization in CSG-TL (Ge et al., 2024), which combines room context with LLM-derived commonsense.

The use of spatial context connects localization to scene synthesis and object placement. Scene synthesis models learn priors over plausible object arrangements (Wang et al., 2019; Dhamo et al., 2021; Paschalidou et al., 2021). SceneGraphNet predicts which object class fits a query location (Zhou et al., 2019), while semantic placement identifies plausible image regions for a queried object (Ramrakhya et al., 2024). These tasks concern constructing or modifying a scene; object localization in partial scenes seeks the position of an object already present but unobserved.

Scene graphs and commonsense reasoning. Scene graphs represent scene entities and their spatial and semantic relationships (Armeni et al., 2019; Rosinol et al., 2021). For example, Wald et al. (2020) learn 3D scene graphs from reconstructed point clouds, while SceneGraphFusion constructs and updates scene graphs incrementally from RGB-D sequences (Wu et al., 2021). Open3DSG supports open-vocabulary object queries and predicts relationships beyond a predefined label set (Koch et al., 2024). External knowledge offers another source of relational information: Feng et al. (2023) combine ConceptNet (Speer et al., 2017) knowledge with visual context for 3D scene graph prediction. Our model reuses D-SCG-OL's directed spatial commonsense graph (D-SCG) encoder with 2D ground-plane positions, object classes, and ConceptNet embeddings. The contribution concerns the prediction target and decision rule built on this representation.

Multiple-hypothesis prediction. Producing several candidates for the same input allows a model to represent output uncertainty (Rupprecht et al., 2017). Examples in trajectory forecasting generate or score multiple possible future paths conditioned on the observed history (Gupta et al., 2018; Chai et al., 2020; Phan-Minh et al., 2020; Zhao et al., 2021). Multiple-choice learning bases its learning objective on the lowest task loss among the predictions for each training example (Guzman-Rivera et al., 2012; Lee et al., 2016). HIAF learns its hypotheses through their combined contribution to a continuous success-probability field, supervised by the full hidden target set.

Probabilistic prediction and decision-making. Mixture density networks model uncertainty in continuous outputs through conditional probability distributions (Bishop, 1994). For outputs containing multiple objects, set prediction accommodates their unordered structure. DETR uses bipartite matching to train such predictions (Carion et al., 2020), while Hess et al. (2022) model distributions over object sets using negative log-likelihood. When a task requires a single location, converting uncertainty into a prediction requires a decision rule aligned with the evaluation loss (Gneiting, 2011). For our task, the relevant criterion is whether the predicted location lies within the evaluation radius of any hidden instance of the queried class. HIAF learns success probabilities at each location from the full hidden target set, forming an action field that guides the final localization decision.

3 Problem formulation

Given a partial scene and a queried object class $c \in \mathcal{C}$, the task is to predict a single ground-plane location $\hat{x} \in \mathbb{R}^2$ for a hidden instance of that class. We represent the N observed objects by $\mathcal{O} = \{(p_i, c_i)\}_{i=1}^{N}$, with ground-plane positions $p_i \in \mathbb{R}^2$ and class labels $c_i \in \mathcal{C}$, where $\mathcal{C}$ denotes the set of object classes. The input context X combines these observations with ConceptNet embeddings and semantic relations (Speer et al., 2017), following the baseline input representation (Giuliari et al., 2022; 2023). Let $Y = \{y_m\}_{m=1}^{M} \subset \mathbb{R}^2$ contain the positions of class-c instances outside the observed-object set. We call these instances hidden and assume $N \geq 1$ and $M \geq 1$.

For any candidate location $x \in \mathbb{R}^2$, let $D_r(x)$ denote the closed disk of radius $r > 0$ centered at x. A prediction at x succeeds if this disk contains at least one point in Y, giving the binary success indicator

$$S_r(x, Y) = \mathbb{I}[\, Y \cap D_r(x) \neq \emptyset \,]. \tag{1}$$

For each candidate location, the expected utility conditioned on (X, c) is its success probability:

$$E_r(x \mid X, c) = \mathbb{E}[S_r(x, Y) \mid X, c] = \mathbb{P}(Y \cap D_r(x) \neq \emptyset \mid X, c)\,. \tag{2}$$

Any location maximizing E_r over $\mathbb{R}^2$ is a Bayes action under this criterion. Localization success rate (LSR@r) is the fraction of evaluation queries whose final prediction satisfies $S_r(\hat{x}, Y) = 1$.

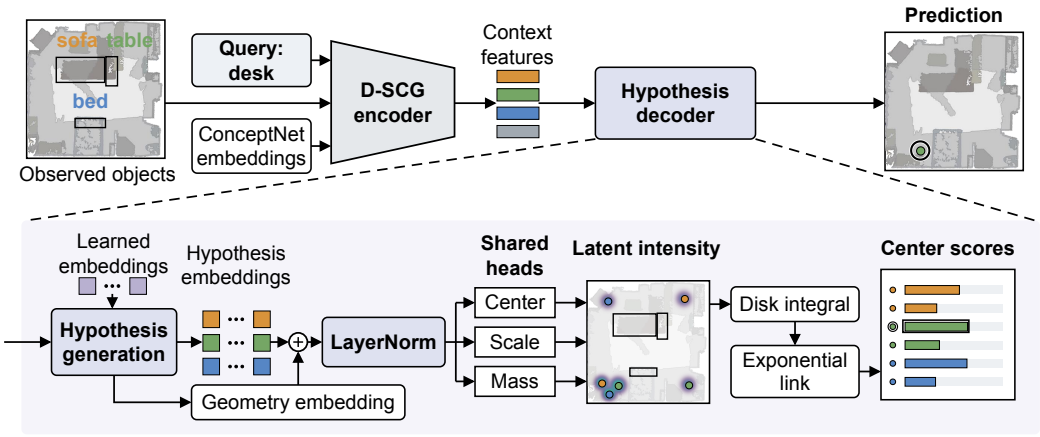


Figure 2: HIAF prediction pipeline. The D-SCG encoder produces context features from observed-object classes and 2D positions, the query class, and ConceptNet embeddings. The decoder uses learned embeddings to generate multiple hypotheses anchored at each observed object. Geometry embeddings are added to the hypothesis representations before layer normalization. Shared heads predict the center, scale, and mass of each hypothesis, together defining the latent intensity. Disk integration followed by an exponential link produces a success score at each candidate center. The highest-scoring center is returned as the prediction.

4 METHOD

HIAF estimates the conditional success probability in Eq. (2) through an action field $E_\theta(x \mid X, c)$, where θ denotes the model parameters. We fix the training and scoring radius to $r = 1$ m and suppress this fixed argument in the field notation. As shown in Figure 2, a target-conditioned D-SCG encoder extracts relational context for anchor-conditioned spatial hypotheses. The decoder predicts Gaussian contributions whose combined disk mass defines the action field. Continuous event supervision trains the encoder and decoder end to end, and the field selects one final point.

4.1 CONTEXT ENCODING

We adopt the graph construction and message-passing architecture of D-SCG-OL (Giuliari et al., 2023) as the D-SCG encoder. The graph contains observed-object nodes, a query node for class c, and concept nodes from ConceptNet. Spatial edges describe directed relative positions between objects, while semantic edges connect objects and the query to concepts through `AtLocation` and `UsedFor` relations. The query's position is treated as unknown. Starting from ConceptNet embeddings, each message-passing layer updates node representations by aggregating information from their graph neighbors. Edge features enter both the attention scores and the messages, incorporating spatial displacements and semantic relation types. Applying ReLU to scaled query–key scores gives sparse, nonnegative weights for this aggregation.

Let h_i and h_{q} denote the encoded representations of observed object i and the query node, respectively. Each is formed by concatenating the node's input embedding with its final message-passing representation, as in D-SCG-OL.

4.2 HYPOTHESIS GENERATION

Our decoder generates K target-location hypotheses per observed object, using its position p_i as a spatial anchor. An affine projection followed by GELU and layer normalization maps $[h_i; h_{\mathrm{q}}]$ to a query-conditioned object representation $b_i \in \mathbb{R}^d$:

$$b_i = \mathrm{LayerNorm}(\mathrm{GELU}(W_b[h_i; h_{\mathrm{q}}] + a_b)), \tag{3}$$

where W_b and a_b are learned projection parameters. We add a learned embedding $e_k \in \mathbb{R}^d$ to distinguish the K hypotheses associated with this object. For each resulting representation $z_{i,k}$, a

two-layer MLP with a GELU activation predicts an offset $\Delta p_{i,k} \in \mathbb{R}^2$ from the anchor:

$$z_{i,k} = b_i + e_k, \qquad \Delta p_{i,k} = \mathrm{MLP}_{\mathrm{offset}}(z_{i,k}), \qquad \mu^{(0)}_{i,k} = p_i + \Delta p_{i,k}. \tag{4}$$

The resulting $H = NK$ preliminary centers provide the positions used for geometry conditioning in the next stage. This MLP is shared across all objects and hypotheses. The embeddings e_k are shared across objects and queried classes, with no predefined spatial meaning or hidden-instance assignment. We use $K = 2$ in the main method, independently of the hidden-instance count.

4.3 Geometry-conditioned prediction

We now index the H hypotheses by j, replacing the pair (i, k). Each preliminary center is expressed relative to the mean observed-object position $\bar{p} = N^{-1} \sum_{i=1}^{N} p_i$. This gives all hypotheses a common spatial reference independent of the global coordinate origin. A geometry MLP embeds $\mu^{(0)}_j - \bar{p}$ and adds its output to the hypothesis representation. Final layer normalization gives

$$\widetilde{z}_j = \mathrm{LayerNorm}\Big(z_j + \mathrm{MLP}_{\mathrm{geom}}(\mu^{(0)}_j - \bar{p})\Big). \tag{5}$$

Three prediction heads map $\widetilde{z}_j$ to a center correction $\Delta\mu_j = \mathrm{MLP}_{\mathrm{refine}}(\widetilde{z}_j)$ and scalar outputs s_j and a_j. The resulting Gaussian parameters are

$$\mu_j = \mu^{(0)}_j + \Delta\mu_j, \qquad \sigma_j = \mathrm{softplus}(s_j), \qquad \alpha_j = \mathrm{softplus}(a_j) + \epsilon_\alpha, \tag{6}$$

where $\epsilon_\alpha > 0$ is a small constant. Each head has its own network parameters, shared across hypotheses. The hypotheses share relational context and receive supervision through their combined contribution to the action field.

4.4 Action field construction

Each refined hypothesis contributes an isotropic Gaussian with center μ_j, scale σ_j, and mass α_j. Summing these contributions defines a nonnegative latent intensity:

$$\lambda_\theta(u \mid X, c) = \sum_{j=1}^{H} \alpha_j \, \mathcal{N}(u; \mu_j, \sigma_j^2 I), \tag{7}$$

where u is a ground-plane location and I is the 2×2 identity matrix. The intensity has total mass $\sum_{j=1}^{H} \alpha_j$ and is not constrained to integrate to one. We omit conditioning on (X, c) below.

To score a location x, we integrate the intensity over its tolerance disk. For hypothesis j, the Gaussian disk probability is

$$q_j(x) = \int_{D_r(x)} \mathcal{N}(u; \mu_j, \sigma_j^2 I)\, du.$$

We sum the component masses within the disk and apply an exponential link to estimate localization success:

$$\Lambda_\theta(x) = \int_{D_r(x)} \lambda_\theta(u)\, du = \sum_{j=1}^{H} \alpha_j q_j(x), \qquad E_\theta(x) = 1 - \exp\big(-\Lambda_\theta(x)\big). \tag{8}$$

At inference, the decoder supplies the finite candidate set $\mathcal{M}_\theta = \{\mu_j\}_{j=1}^{H}$. The prediction rule selects the highest-scoring generated center:

$$\hat{x} \in \arg\max_{x \in \mathcal{M}_\theta} E_\theta(x). \tag{9}$$

Each candidate's score incorporates contributions from all hypotheses. Appendix A details the architecture, parameterization, numerical integration, and candidate-selection procedure.

4.5 TRAINING OBJECTIVE

The hidden target set Y defines the success region $A_Y = \bigcup_{y \in Y} D_r(y)$, where $S_r(x, Y) = 1$. We supervise the action field with binary cross-entropy integrated over the ground plane and normalized by the disk area:

$$\mathcal{L}_{\text{succ}} = \frac{1}{\pi r^2} \int_{\mathbb{R}^2} \big[-S_r(x, Y) \log E_\theta(x) - (1 - S_r(x, Y)) \log(1 - E_\theta(x)) \big] \, dx. \tag{10}$$

At each location, the expected binary cross-entropy is minimized by the true conditional success probability (Gneiting & Raftery, 2007).

The exponential link gives $-\log(1 - E_\theta(x)) = \Lambda_\theta(x)$ for a negative label. Gaussian normalization yields $\int_{\mathbb{R}^2} \Lambda_\theta(x) \, dx = \pi r^2 \sum_{j=1}^{H} \alpha_j$. Integrating this negative-label contribution over the plane and correcting it on A_Y gives

$$\mathcal{L}_{\text{succ}} = \sum_{j=1}^{H} \alpha_j + \frac{1}{\pi r^2} \int_{A_Y} \big[-\log E_\theta(x) - \Lambda_\theta(x) \big] \, dx. \tag{11}$$

Thus, the full-plane negative-label contribution is analytic, and only the correction over the success region requires sampling.

For each sample, we select a hidden instance uniformly and draw a location uniformly by area from its radius-r disk. Let $n_Y(x)$ count the target disks covering x. For $L = 256$ independent samples x_ℓ, the loss estimator is

$$\widehat{\mathcal{L}}_{\text{succ}} = \sum_{j=1}^{H} \alpha_j + \frac{1}{L} \sum_{\ell=1}^{L} \frac{M}{n_Y(x_\ell)} \big[-\log E_\theta(x_\ell) - \Lambda_\theta(x_\ell) \big]. \tag{12}$$

The weight $M/n_Y(x_\ell)$ corrects the higher sampling density in regions covered by multiple disks.

We average this loss across queries in each batch. Gradients reach the encoder and decoder through the continuous field, without differentiating through the final candidate argmax. Appendix A provides the loss derivations and details of the sampling estimator.

5 EXPERIMENTS

We evaluate localization success and conditional error on Partial ScanNet, then isolate the effects of the decoder and training objective. Additional comparisons examine the evaluation tolerance, hypothesis count, and number of hidden target instances.

5.1 EXPERIMENTAL SETUP

Data and comparisons. We evaluate on Partial ScanNet, reconstructed from ScanNet using the SCG-OL and D-SCG-OL benchmark resources (Dai et al., 2017; Giuliari et al., 2022; 2023). The test set contains $16{,}925$ queries across $4{,}385$ partial scenes. We compare with MLPC, SCG-OL, Graphormer, and D-SCG-OL using the same query splits and object-level inputs. Appendices B and C describe the splits, baselines, and training settings.

Protocols. Exhaustive Class-Query Evaluation (ECQE) is the primary protocol and evaluates every applicable class in each partial scene. Random Single-Class-Query Evaluation (RSCQE) samples one class per scene and averages ten fixed repetitions within each model run. Relaxed Scene-Level Success Evaluation (RSSE) counts a scene as successful when any class query succeeds.

Metrics. Following prior work (Giuliari et al., 2022; 2023), we use localization success rate at 1 m (LSR@1m) as the primary metric. We also report mean successful localization error (mSLE), computed over each method's own 1 m successes. Under RSSE, we report scene success rate (SSR@1m), and mSLE averages successful scenes' minimum query distances (Appendix B). We additionally report success rates at the stricter 0.5 m tolerance.

Implementation details. HIAF uses $K = 2$ hypotheses per observed object and $r = 1$ m for training and scoring. We report means and standard deviations across three random seeds. Appendices A and B provide the implementation details, training settings, and aggregation rules.

5.2 LOCALIZATION PERFORMANCE

HIAF achieves the highest mean success rate and lowest mSLE across the three protocols at 1 m (Table 1). Its ECQE LSR@1m reaches 30.06%, compared with 25.21% for D-SCG-OL, an improvement of 4.86 percentage points. The advantage persists under RSCQE's uniform scene sampling and RSSE's scene-level success aggregation.

Table 1: Localization at 1 m on Partial ScanNet. Success is query-level LSR under ECQE/RSCQE and scene-level SSR under RSSE (%). mSLE (m) is conditional on each method's own 1 m success. Mean ± standard deviation over three seeds; bold marks the best mean.

Method	ECQE	RSCQE	RSSE
	Success@1m (%) ↑		
MLPC	19.08 ± 0.59	20.11 ± 0.79	46.75 ± 1.56
SCG-OL	24.79 ± 0.49	24.48 ± 1.05	50.98 ± 2.69
Graphormer	24.42 ± 0.92	25.37 ± 1.34	52.68 ± 1.40
D-SCG-OL	25.21 ± 0.55	26.29 ± 0.66	53.66 ± 1.13
HIAF	$\mathbf{30.06 \pm 0.72}$	$\mathbf{32.19 \pm 0.82}$	$\mathbf{60.77 \pm 1.07}$
	mSLE (m) ↓		
MLPC	0.618 ± 0.003	0.630 ± 0.007	0.563 ± 0.003
SCG-OL	0.581 ± 0.007	0.587 ± 0.008	0.523 ± 0.010
Graphormer	0.602 ± 0.009	0.599 ± 0.003	0.538 ± 0.004
D-SCG-OL	0.585 ± 0.008	0.574 ± 0.007	0.514 ± 0.002
HIAF	$\mathbf{0.546 \pm 0.005}$	$\mathbf{0.528 \pm 0.009}$	$\mathbf{0.467 \pm 0.003}$

The success-rate advantage persists at the stricter 0.5 m tolerance (Table 2). With the predictions unchanged, HIAF retains the highest mean success rate under all three protocols, with 13.19% under ECQE and 15.11% under RSCQE.

Table 2: Success rates (%) at the stricter 0.5 m tolerance, using the same predictions as in Table 1. Mean ± standard deviation over three seeds; bold marks the best mean.

Method	ECQE LSR ↑	RSCQE LSR ↑	RSSE SSR ↑
MLPC	6.05 ± 0.31	5.76 ± 0.34	19.12 ± 0.88
SCG-OL	9.54 ± 0.20	9.06 ± 0.41	24.16 ± 0.82
Graphormer	8.68 ± 0.45	9.11 ± 0.50	24.57 ± 0.57
D-SCG-OL	9.44 ± 0.34	10.23 ± 0.33	26.27 ± 0.11
HIAF	$\mathbf{13.19 \pm 0.37}$	$\mathbf{15.11 \pm 1.01}$	$\mathbf{34.78 \pm 1.09}$

5.3 ABLATION STUDIES

We examine the effects of hypothesis interaction and the training objective through two controlled variants (Table 3). Both variants retain $K = 2$, the relational encoder, and the action-field scoring rule. Appendix C specifies the matched settings.

Success supervision versus set likelihood. The objective ablation replaces full-plane event supervision with a Poisson point process (PPP) set negative log-likelihood, while retaining the decoder without hypothesis interaction. Under these settings, event supervision improves mean success at 1 m by 2.05, 2.20, and 2.74 percentage points under ECQE, RSCQE, and RSSE, respectively. These

gains extend to 0.5 m, while mean errors among successful predictions remain close across the two objectives.

Hypothesis interaction. Adding two self-attention blocks between the geometry embedding and final normalization gives similar mean success at 1 m. HIAF without these blocks has lower mSLE and higher success at 0.5 m under all three protocols. The comparison supports the simpler decoder at the primary and stricter tolerances. Additional comparisons appear in Appendix D.

Table 3: Decoder and objective ablations at $K = 2$. Each variant changes one component of HIAF. Success denotes LSR under ECQE/RSCQE and SSR under RSSE (%); mSLE (m) is conditional on 1 m success. Mean $\pm$ standard deviation over three seeds; bold marks the best mean.

Protocol	Variant	Success@1m ↑	Success@0.5m ↑	mSLE ↓
ECQE	+ interaction	29.96 ± 0.10	12.11 ± 0.32	0.572 ± 0.005
	PPP objective	28.01 ± 0.74	12.34 ± 0.77	0.547 ± 0.010
	HIAF	$\mathbf{30.06 \pm 0.72}$	$\mathbf{13.19 \pm 0.37}$	$\mathbf{0.546 \pm 0.005}$
RSCQE	+ interaction	31.95 ± 0.68	13.50 ± 0.80	0.557 ± 0.009
	PPP objective	29.99 ± 1.39	13.75 ± 0.89	0.531 ± 0.003
	HIAF	$\mathbf{32.19 \pm 0.82}$	$\mathbf{15.11 \pm 1.01}$	$\mathbf{0.528 \pm 0.009}$
RSSE	+ interaction	60.45 ± 1.03	32.51 ± 1.24	0.487 ± 0.011
	PPP objective	58.02 ± 2.46	32.48 ± 1.83	0.471 ± 0.011
	HIAF	$\mathbf{60.77 \pm 1.07}$	$\mathbf{34.78 \pm 1.09}$	$\mathbf{0.467 \pm 0.003}$

Number of hypotheses. Mean ECQE LSR@1m ranges from 29.87% to 30.06% across $K \in \{1, 2, 4\}$ without hypothesis interaction (Table 4). The main setting, $K = 2$, has the highest mean LSR@1m and lowest mSLE under both ECQE and RSCQE. At $K = 1$, the field already combines N hypotheses, one per observed object. The comparison therefore varies per-anchor capacity while retaining the aggregation of hypotheses across observed objects.

Table 4: Hypothesis-count sensitivity of HIAF without hypothesis interaction. K is the number of hypotheses per observed object. LSR is in percent; mSLE is in meters and uses each model's own 1 m successes. Mean $\pm$ standard deviation over three seeds; bold marks the best mean.

	ECQE		RSCQE	
K	LSR@1m ↑	mSLE ↓	LSR@1m ↑	mSLE ↓
1	29.87 ± 0.70	0.558 ± 0.001	31.89 ± 1.35	0.537 ± 0.008
2	$\mathbf{30.06 \pm 0.72}$	$\mathbf{0.546 \pm 0.005}$	$\mathbf{32.19 \pm 0.82}$	$\mathbf{0.528 \pm 0.009}$
4	29.95 ± 0.62	0.553 ± 0.005	31.63 ± 0.95	0.534 ± 0.011

5.4 Single- and multi-instance queries

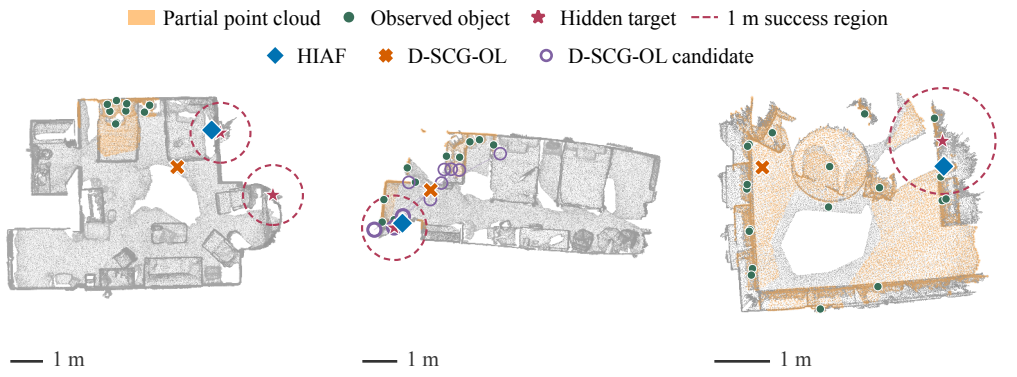
HIAF achieves the highest mean LSR@1m for both single- and multi-instance queries under ECQE and RSCQE (Table 5). Under ECQE, it improves over the strongest baseline in each stratum by 5.04 percentage points for single-instance queries and 3.68 points for multi-instance queries. HIAF also has the lowest mean mSLE in both strata under both protocols (Appendix D), with errors conditional on each method's own successful-query subset.

5.5 Qualitative analysis

We visualize three selected cases to compare the predictions of HIAF and D-SCG-OL (Figure 3). HIAF successfully localizes a hidden target in each case, whereas D-SCG-OL fails. In the sink example, three of the ten D-SCG-OL anchor predictions lie within the target's success region, but the mean of all ten falls outside it. Appendix D shows the corresponding HIAF hypotheses and action fields.

Table 5: LSR@1m (%) by hidden-instance count. ECQE has 11,017 single-instance and 5,908 multi-instance queries; RSCQE averages each stratum over ten fixed repetitions per model run. Mean ± standard deviation over three seeds; bold marks the best mean.

	ECQE		RSCQE	
Method	Single	Multiple	Single	Multiple
MLPC	14.69 ± 0.66	27.26 ± 0.60	15.25 ± 1.17	29.70 ± 0.21
SCG-OL	21.92 ± 0.54	30.15 ± 0.43	21.09 ± 0.47	31.17 ± 2.33
Graphormer	22.21 ± 1.06	28.55 ± 0.78	22.36 ± 1.36	31.31 ± 1.64
D-SCG-OL	23.00 ± 0.33	29.31 ± 1.17	23.96 ± 0.73	30.88 ± 0.75
HIAF	**28.04 ± 0.44**	**33.83 ± 1.73**	**30.53 ± 0.74**	**35.48 ± 2.40**



HIAF 0.31 m | D-SCG-OL 1.84 m (a) Window

HIAF 0.29 m | D-SCG-OL 1.62 m (b) Sink

HIAF 0.48 m | D-SCG-OL 3.44 m (c) Books

Figure 3: Qualitative comparison of HIAF and D-SCG-OL in three selected cases. Gray and orange layers show full-scene and partial point-cloud projections for visualization only. Green circles and red stars mark observed objects and hidden targets; dashed circles indicate 1 m success regions. Blue diamonds and orange crosses mark HIAF and D-SCG-OL predictions, with nearest-target distances reported below each scene. In (b), open circles show the ten D-SCG-OL anchor predictions, connected to their mean.

6 Conclusion

This paper introduced HIAF for partial-scene object localization, using a continuous action field to estimate the probability of localizing a hidden target within a fixed radius. Relational context supplies anchor-conditioned Gaussian hypotheses, and an analytic background term makes full-plane event supervision tractable. Supervision uses the full hidden target set to identify successful prediction locations, providing a common learning target for single- and multi-instance queries. At inference, the field scores the generated centers to select a single prediction. Each candidate's score incorporates contributions from all hypotheses, allowing spatial alternatives to inform the final decision collectively. On Partial ScanNet, HIAF achieves higher mean success rates and lower mean successful localization error than the evaluated baselines across three protocols at 1 m. The success-rate gains extend to 0.5 m and to both single- and multi-instance queries at 1 m. Full-plane event supervision also yields higher mean success rates than PPP set likelihood when the decoder and scoring rule are held fixed. The method offers a decision-focused alternative to coordinate-error training for partial-scene localization.

Limitations. We evaluate HIAF on Partial ScanNet using object-level positions and class labels. The current setting assumes at least one hidden instance of the queried class. Performance with noisy perception inputs, absent targets, and other scene distributions remains to be evaluated. Training and scoring use a fixed radius of 1 m; the effects of training for other radii have not been examined.

AI USE STATEMENT

We used generative AI tools to assist with code implementation and to provide feedback on the methodology, experimental comparisons, and interpretation of results. They also assisted with the exposition and checking of mathematical derivations, manuscript drafting and language editing, literature and citation checks, simulated pre-submission review, plotting code, figure and table preparation, and LaTeX formatting. Reported numerical results were checked against supplied experiment records, and qualitative comparison figures were constructed from saved model predictions and scene data.

ETHICS STATEMENT

This work studies object localization using an existing ScanNet-derived benchmark, with object-level positions and semantic labels as inputs. Indoor scene reconstructions can reveal information about private spaces; their use and redistribution must respect the source datasets' access conditions and the privacy of the environments represented.

REPRODUCIBILITY STATEMENT

Section 4 specifies the model, training objective, and prediction rule. Appendix A provides the architecture, loss derivations, numerical integration and sampling procedures, optimization settings, and model-selection criterion. Appendix B describes benchmark construction, query eligibility, data splits, evaluation protocols, and aggregation across runs, and identifies the artifacts needed to reconstruct the benchmark and evaluation. Appendix C documents the baseline adaptations and controlled ablation settings. Additional quantitative comparisons and visualizations are provided in Appendix D.

REFERENCES

Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5664–5673, 2019.

Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020.

Christopher M Bishop. Mixture density networks. 1994.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura (eds.), *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 86–99. PMLR, 30 Oct–01 Nov 2020.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in neural information processing systems*, 33:4247–4258, 2020.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

Helisa Dhamo, Fabian Manhardt, Nassir Navab, and Federico Tombari. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16332–16341. IEEE, 2021.

Mingtao Feng, Haoran Hou, Liang Zhang, Zijie Wu, Yulan Guo, and Ajmal Mian. 3d spatial multimodal knowledge accumulation for scene graph prediction in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9182–9191, 2023.

Wenqi Ge, Chao Tang, and Hong Zhang. Commonsense scene graph-based target localization for object search. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13318–13325. IEEE, 2024.

Francesco Giuliari, Geri Skenderi, Marco Cristani, Yiming Wang, and Alessio Del Bue. Spatial commonsense graph for object localisation in partial scenes. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19496–19505. IEEE, 2022.

Francesco Giuliari, Geri Skenderi, Marco Cristani, Alessio Del Bue, and Yiming Wang. Leveraging commonsense for object localisation in partial scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12038–12049, 2023.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2255–2264. IEEE, 2018.

Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. *Advances in neural information processing systems*, 25, 2012.

Georg Hess, Christoffer Petersson, and Lennart Svensson. Object detection as probabilistic set prediction. In *European Conference on Computer Vision*, pp. 550–566. Springer, 2022.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14183–14193. IEEE, 2024.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. *Advances in Neural Information Processing Systems*, 29, 2016.

Ester Martinez-Martin and Angel P Del Pobil. Object detection and recognition for assistive robots: Experimentation and implementation. *IEEE Robotics & Automation Magazine*, 24(3):123–138, 2017.

Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. *Advances in neural information processing systems*, 34:12013–12026, 2021.

Tung Phan-Minh, Elena Corina Grigore, Freddy A Boulton, Oscar Beijbom, and Eric M Wolff. Covernet: Multimodal behavior prediction using trajectory sets. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14062–14071. IEEE, 2020.

Ram Ramrakhya, Aniruddha Kembhavi, Dhruv Batra, Zsolt Kira, Kuo-Hao Zeng, and Luca Weihs. Seeing the unseen: Visual common sense for semantic placement. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16273–16283. IEEE, 2024.

Antoni Rosinol, Andrew Violette, Marcus Abate, Nathan Hughes, Yun Chang, Jingnan Shi, Arjun Gupta, and Luca Carlone. Kimera: From slam to spatial perception with 3d dynamic scene graphs. *The International Journal of Robotics Research*, 40(12-14):1510–1546, 2021.

Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 3611–3620. IEEE, 2017.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3960–3969. IEEE, 2020.

Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X Chang, and Daniel Ritchie. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019.

Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7511–7521. IEEE, 2021.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888, 2021.

Zhongmou Ying, Xianfeng Yuan, Baojiang Yang, Yong Song, Qingyang Xu, Fengyu Zhou, and Weihua Sheng. Rp-sg: Relation prediction in 3d scene graphs for unobserved objects localization. *IEEE Robotics and Automation Letters*, 9(2):1412–1419, 2024. doi: 10.1109/LRA.2023.3342666.

Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Ben Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. In *Conference on robot learning*, pp. 895–904. PMLR, 2021.

Yang Zhou, Zachary While, and Evangelos Kalogerakis. Scenegraphnet: Neural message passing for 3d indoor scene augmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7383–7391. IEEE, 2019.

APPENDIX CONTENTS

- **Appendix A**: Implementation details, numerical procedures, and loss derivations.
- **Appendix B**: Dataset construction, evaluation protocols, and reproducibility.
- **Appendix C**: Baseline adaptations and ablation settings.
- **Appendix D**: Complementary evaluations and candidate-level visualizations.

A IMPLEMENTATION DETAILS

A.1 ENCODER INPUTS

The graph has observed-object nodes, a query-class node, and concept nodes, each initialized with a 300-dimensional ConceptNet embedding. Observed-object and query nodes form a directed complete graph without self-loops. Each edge has six attributes: three relation-type indicators, one unknown-position indicator, and a two-dimensional displacement. Spatial edges between observed objects encode the displacement from the source to the destination. Proximity edges involving the query node carry zero displacement and have the unknown-position indicator set to one. Semantic edges connect objects and the query to concepts in both directions through `AtLocation` and `UsedFor` relations. Their displacement attributes are zero. The unknown-position indicator is zero on semantic edges. The encoder has four message-passing layers with output widths 256, 512, 512, and 512, each with four attention heads. Its sparse attention applies ReLU to scaled query–key scores before aggregating edge-conditioned values, without a neighbor softmax. The blocks include residual connections, normalization, gated residual fusion, and dropout, following the D-SCG encoder. The final 512-dimensional message-passing representation is concatenated with the original 300-dimensional embedding, giving 812-dimensional features h_i and h_{q}.

A.2 HYPOTHESIS DECODER

Within each projection or prediction head, parameters are reused across objects and hypotheses; different modules have separate parameters. The query-conditioned object representation b_i is obtained from $[h_i; h_{\mathrm{q}}]$ by an affine projection from 1,624 to $d = 256$ dimensions, followed by a GELU activation and layer normalization. The learned hypothesis embeddings e_k are initialized from a zero-mean Gaussian with standard deviation 0.02. The offset MLP $\mathrm{MLP}_{\mathrm{offset}}$ maps $z_{i,k}$ to $\Delta p_{i,k}$. It has two affine layers, with dimensions $d \to d \to 2$, an intermediate GELU activation, and dropout. The geometry embedding of $\mu_j^{(0)} - \bar{p}$ uses a separate MLP with dimensions $2 \to d \to d$ and an intermediate GELU activation. Its output is added to the hypothesis representation.

The geometry-conditioned representations pass directly through final layer normalization and the three prediction heads in Eq. (6). The refinement MLP $\mathrm{MLP}_{\mathrm{refine}}$ has the same layer dimensions as the offset MLP, with separate parameters. Scale and mass use affine heads from d dimensions to one. The mass offset is $\epsilon_\alpha = 10^{-4}$, and scales use unbounded softplus.

A.3 GAUSSIAN DISK INTEGRATION

For $u \sim \mathcal{N}(\mu_j, \sigma_j^2 I)$, the squared distance $\|u - x\|_2^2/\sigma_j^2$ follows a noncentral chi-square distribution with two degrees of freedom. The disk mass is therefore

$$q_j(x) = F_{\chi'^2_2(\nu_j(x))}\left(\frac{r^2}{\sigma_j^2}\right), \qquad \nu_j(x) = \frac{\|x - \mu_j\|_2^2}{\sigma_j^2}, \tag{13}$$

where $\nu_j(x)$ is the noncentrality parameter. Writing $\rho = \|x - \mu_j\|_2$ and $\sigma = \sigma_j$, angular integration gives the equivalent radial integral

$$q(\rho, \sigma) = \int_0^r \frac{t}{\sigma^2} \exp\left[-\frac{(t-\rho)^2}{2\sigma^2}\right] I_{0e}\left(\frac{t\rho}{\sigma^2}\right) dt, \tag{14}$$

where $I_{\nu e}(z) = e^{-z} I_\nu(z)$ for $z \geq 0$ is the scaled modified Bessel function. We evaluate this integral by adaptive Gauss–Legendre quadrature, using scaled Bessel functions and log-domain arithmetic. Training and inference use the same computation.

Quadrature starts with 16 nodes and doubles the order up to 256. For orders n and $2n$, the convergence check is

$$|\log q_{2n} - \log q_n| \le \varepsilon_q + 16\epsilon_{64} \max(1, |\log q_{2n}|), \qquad \varepsilon_q = 10^{-10}, \tag{15}$$

where ϵ_{64} is double-precision machine epsilon. For interval-expansion attempt $a \in \{0, 1, 2, 3\}$, set $B_a = -\log \varepsilon_q + 16 + 24a$. The radial interval is

$$[\ell_a, u_a] = \begin{cases} [\max(0, \rho - \sigma\sqrt{2B_a}),\ \min(r, \rho + \sigma\sqrt{2B_a})], & \rho \le r, \\ [\max(0, r - w_a),\ r], & \rho > r, \end{cases} \qquad w_a = \frac{2\sigma B_a}{\sqrt{z^2 + 2B_a} + z}, \tag{16}$$

where $z = (\rho - r)/\sigma$ in the second case. On each omitted interval, the Gaussian exponential is largest at the retained endpoint. Using $I_{0e}(z) \le 1$ for $z \ge 0$ and $\int_0^r t/\sigma^2\, dt = r^2/(2\sigma^2)$ bounds the omitted mass by

$$b_a = \frac{r^2}{2\sigma^2} \left(\mathbb{I}[\ell_a > 0] e^{-(\rho - \ell_a)^2/(2\sigma^2)} + \mathbb{I}[u_a < r] e^{-(u_a - \rho)^2/(2\sigma^2)} \right).$$

We accept the interval when $b_a \le \varepsilon_q q_a$, evaluated in the log domain, where q_a is the converged quadrature estimate. Otherwise, we enlarge the interval at the next attempt. Disk probabilities and the continuous loss are computed in double precision. We accumulate component contributions in the log domain and evaluate the exponential link with a stable $\mathrm{expm1}$ expression, using its small-intensity limit when needed. The routine raises an error if the convergence, omitted-mass, or finite-precision checks fail.

To obtain the derivatives, center the disk at the origin and write $f(u) = \mathcal{N}(u; (\rho, 0), \sigma^2 I)$. The identities

$$\partial_\rho f = -\partial_{u_1} f, \qquad \partial_\sigma f = -\frac{1}{\sigma} \nabla_u \cdot \big[(u - (\rho, 0)) f \big]$$

convert the derivatives to boundary integrals by the divergence theorem. Evaluating them on $u = r(\cos\phi, \sin\phi)$ gives, with $z = r\rho/\sigma^2$,

$$\frac{\partial q}{\partial \rho} = -\frac{r}{\sigma^2} \exp\left[-\frac{(\rho - r)^2}{2\sigma^2} \right] I_{1e}(z), \tag{17}$$

$$\frac{\partial q}{\partial \sigma} = \frac{\exp[-(\rho - r)^2/(2\sigma^2)]}{\sigma^3} \big[r\rho I_{1e}(z) - r^2 I_{0e}(z) \big]. \tag{18}$$

The custom backward evaluates these derivative-to-probability ratios in log-domain arithmetic to obtain the gradients of $\log q$. At $\rho = 0$, the center gradient is zero; the stable vector expression uses $I_{1e}(z)/z \to 1/2$ as $z \to 0$. The backward implementation supports first derivatives. The centered limit $q(0, \sigma) = 1 - \exp[-r^2/(2\sigma^2)]$ provides an analytic check on the numerical calculation.

A.4 Sampling and candidate selection

To sample uniformly by disk area, draw a target index J uniformly from $\{1, \ldots, M\}$ and independent $U, V \sim \mathrm{Uniform}[0, 1)$:

$$x = y_J + r\sqrt{U}\, (\cos(2\pi V), \sin(2\pi V)). \tag{19}$$

Selecting a target uniformly and then sampling uniformly by area from its disk gives the density

$$g_Y(x) = \frac{n_Y(x)}{M\pi r^2}, \qquad n_Y(x) = \sum_{y \in Y} \mathbb{I}[\, x \in D_r(y)\,]. \tag{20}$$

Thus the importance weight for the area-normalized integral in Eq. (11) is $1/(\pi r^2 g_Y(x)) = M/n_Y(x)$ on A_Y. For any integrable f on A_Y,

$$\mathbb{E}_{g_Y} \left[\frac{M}{n_Y(x)} f(x) \right] = \int_{A_Y} \frac{M}{n_Y(x)} f(x) g_Y(x)\, dx = \frac{1}{\pi r^2} \int_{A_Y} f(x)\, dx.$$

Substituting $f(x) = -\log E_\theta(x) - \Lambda_\theta(x)$ gives the correction term in Eq. (12). With exact disk masses, this estimator is unbiased for Eq. (11). Each update draws $L = 256$ independent samples per query.

Although the continuous BCE loss is nonnegative, its Monte Carlo estimator can be negative; we use these estimates without clipping.

At inference, we select among the H final hypothesis centers according to Eq. (9). Each score includes contributions from all hypotheses. We keep the scoring radius at its training value and use the same prediction for every reported metric and evaluation threshold.

Candidate probabilities are evaluated in double precision and cast to single precision for selection. Ties after this cast are resolved lexicographically by the first, then the second center coordinate.

A.5 Model and optimization settings

Table 6 summarizes the HIAF architecture, field, and optimization settings. Query losses are averaged within each batch. Ground-plane rotation augmentation draws an angle uniformly from $[-\pi, \pi)$ for each training query and rotates observed and target coordinates about the observed-object mean, with consistent edge displacements. The cosine learning-rate schedule advances once per epoch, without warmup. Validation is evaluated after each epoch. Training runs for 100 epochs without early stopping. We evaluate the checkpoint with the highest validation ECQE LSR@1m. Experiments use an NVIDIA A100-SXM4-80GB GPU.

Table 6: HIAF architecture, field, and optimization settings.

Setting	Value
ConceptNet embedding dimension	300
Edge-feature dimension	6
Message-passing layers / attention heads	4 / 4
Encoder output widths	$256, 512, 512, 512$
Hypotheses per observed object K	2
Hypothesis representation dimension d	256
Object–query projection	$1{,}624 \to 256$
Learned-slot initialization	$\mathcal{N}(0, 0.02^2)$
Offset / refinement MLP	$256 \to 256 \to 2$
Geometry MLP	$2 \to 256 \to 256$
MLP activation	GELU
Scale / mass head	$256 \to 1$
Dropout	0.2
Training and scoring radius r	1 m
Scale / mass parameterization	Softplus / softplus $+ 10^{-4}$
Loss	Full-plane binary cross-entropy
Target-disk samples per query L	256
Quadrature initial / maximum order	16 / 256
Quadrature tolerance ε_q	10^{-10}
Optimizer / weight decay	AdamW / 10^{-4}
Initial / minimum learning rate	10^{-4} / 10^{-6}
Learning-rate schedule	Cosine decay
Batch size / training epochs	16 / 100
Gradient clipping norm	1
Ground-plane rotation augmentation	Enabled

A.6 Loss identities and statistical interpretation

By Tonelli's theorem, integrating over action centers first contributes the disk area πr^2. Gaussian normalization then gives

$$\int_{\mathbb{R}^2} \Lambda_\theta(x)\, dx = \pi r^2 \sum_{j=1}^{H} \alpha_j \int_{\mathbb{R}^2} \mathcal{N}(u; \mu_j, \sigma_j^2 I)\, du = \pi r^2 \sum_{j=1}^{H} \alpha_j. \tag{21}$$

Thus the total-mass term in Eq. (11) is fixed by the continuous objective, rather than an independently weighted regularizer. For finite Y, finite positive masses, and positive finite scales, the success region is bounded and the field is continuous and strictly positive there, so the per-query loss is finite.

For fixed (X, c), let $R(E)$ be the expected loss over Y for a prediction function E. Under finite expected losses,

$$R(E) - R(E_r) = \frac{1}{\pi r^2} \int_{\mathbb{R}^2} D_{\mathrm{KL}}(\mathrm{Bern}(E_r(x \mid X, c)) \parallel \mathrm{Bern}(E(x))) \, dx \geq 0. \tag{22}$$

This follows by taking the conditional expectation of binary cross-entropy at each location (Gneiting & Raftery, 2007). It identifies the unrestricted population optimum almost everywhere; it does not guarantee that the finite Gaussian family contains that optimum or that optimization finds it. Nor does integrated risk alone guarantee the correct ranking at every generated center. The exponential link defines a family of event probabilities without assuming that real hidden objects follow a Poisson point process. Individual components need not correspond to distinct instances or modes.

A.7 Computational scope

With $H = NK$ hypotheses, scoring all centers requires H^2 disk integrals. Loss evaluation adds LH disk integrals and LM distance checks for union coverage. If Q denotes the largest quadrature order reached, the quadrature work scales as $O((LH + H^2)Q)$, up to the adaptive refinements. The training forward pass also computes candidate scores, so its cost includes the H^2 term. These operation counts are not measurements of latency or memory use.

B Dataset and evaluation protocols

B.1 Partial scenes and query construction

Partial ScanNet uses partial observations of ScanNet indoor reconstructions (Dai et al., 2017), following the benchmark-generation resources of SCG-OL and D-SCG-OL (Giuliari et al., 2022; 2023). The benchmark pipeline divides RGB-D sequences into subsequences, reconstructs their partial point clouds using camera calibration and poses, and transfers annotations from the complete reconstruction. For localization, a partial scene supplies observed object classes and two-dimensional ground-plane positions. Positions are ground-plane projections of axis-aligned bounding-box centers. Observed centers use partial-scene instance points; hidden target centers use complete-scene instance points. The model operates on the observed-object classes and positions and ConceptNet information, not on raw RGB-D frames, full meshes, or room boundaries.

Each partial scene used in our experiments contains at least three observed object instances. The fixed query lists contain 33 object classes, excluding wall, floor, ceiling, and person. Hidden instances are those outside the observed-object set.

Let $\mathcal{C}_s$ contain the classes with at least one hidden instance in partial scene s. Each pair (s, c) with $c \in \mathcal{C}_s$ defines one query, with hidden target set $Y_{s,c}$ and one prediction $\hat{x}_{s,c}$. All hidden instances of the queried class are admissible matches. Queries with empty hidden target sets are outside the task definition. Hidden coordinates supply supervision and evaluation labels only.

ScanNet's official training scans are divided 9:1 between our training and validation sets; official validation scans form the test set. Source-scan identifiers are disjoint across the three splits, so partial scenes from one scan cannot straddle them. Table 7 gives counts from the query manifests. The test set contains 11,017 queries with $|Y| = 1$ and 5,908 with $|Y| \geq 2$.

Table 7: Dataset counts at three distinct levels of aggregation.

Split	Source scans	Partial scenes	Scene-class queries
Training	1,044	14,244	54,851
Validation	116	1,472	6,170
Test	299	4,385	16,925

All methods use the same fixed training, validation, and test query lists. The scan and scene-class query manifests specify the data splits and evaluation population.

B.2 Query and scene metrics

For each query, define the nearest-target distance

$$d_{s,c} = \min_{y \in Y_{s,c}} \|\hat{x}_{s,c} - y\|_2. \tag{23}$$

For a set of evaluated queries $\mathcal{Q}$ and evaluation threshold t, query-level success is

$$\mathrm{LSR@}t = \frac{1}{|\mathcal{Q}|} \sum_{(s,c)\in\mathcal{Q}} \mathbb{I}[d_{s,c} \le t]. \tag{24}$$

At the primary threshold $t = 1\,\mathrm{m}$, let $\mathcal{Q}_{\mathrm{succ}} = \{(s,c) \in \mathcal{Q} : d_{s,c} \le 1\,\mathrm{m}\}$. Then

$$\mathrm{mSLE} = \frac{1}{|\mathcal{Q}_{\mathrm{succ}}|} \sum_{(s,c)\in\mathcal{Q}_{\mathrm{succ}}} d_{s,c}. \tag{25}$$

mSLE is defined for a nonempty set of successful queries. Success rates are displayed as percentages, and errors are measured in meters.

ECQE. The primary protocol sets $\mathcal{Q} = \{(s,c) : s \in \mathcal{S},\ c \in \mathcal{C}_s\}$, where $\mathcal{S}$ is the set of test partial scenes. Each query receives equal weight. Consequently, partial scenes with more applicable classes contribute more queries.

RSCQE. For each of ten repetitions, one class is sampled uniformly from $\mathcal{C}_s$ for every partial scene s. The resulting set has one query per scene. We compute each reported success rate and the 1 m-conditional mSLE separately for each repetition, then average within a model run. In particular, the conditional mSLE ratios are averaged, rather than pooling successful queries across repetitions. All methods use the same saved class mappings. This gives each partial scene equal weight in success rate and is distinct from ECQE's uniform weighting over queries.

RSSE. Define the best class-query distance in partial scene s as $d_s^{\min} = \min_{c\in\mathcal{C}_s} d_{s,c}$ and let $\mathcal{S}_{\mathrm{succ}} = \{s : d_s^{\min} \le 1\,\mathrm{m}\}$. The scene-level success at tolerance t and the conditional error at 1 m are

$$\mathrm{SSR@}t = \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \mathbb{I}[d_s^{\min} \le t], \qquad \mathrm{mSLE}_{\mathrm{scene}} = \frac{1}{|\mathcal{S}_{\mathrm{succ}}|} \sum_{s\in\mathcal{S}_{\mathrm{succ}}} d_s^{\min}. \tag{26}$$

RSSE measures whether at least one queried class is successfully localized in a partial scene. Results are aggregated by scene and reported without hidden-instance-count stratification.

Evaluation thresholds. The radius comparison varies t in Eq. (24) while keeping every $\hat{x}_{s,c}$ fixed. HIAF's training and scoring radius stays at $r = 1\,\mathrm{m}$. The distinction between t and r separates evaluation tolerance from the field used to select a point.

B.3 Aggregation and uncertainty

For three model seeds, we report the mean and standard deviation with a divisor of three. For seed-level metric values m_1, m_2, m_3,

$$\overline{m} = \frac{1}{3}\sum_{b=1}^{3} m_b, \qquad s = \sqrt{\frac{1}{3}\sum_{b=1}^{3}(m_b - \overline{m})^2}.$$

This convention applies to baseline, HIAF, ablation, and sensitivity tables. For RSCQE, each model seed first contributes its mean over ten fixed class-sampling repetitions. These repetitions are not additional model seeds. ECQE evaluates $16{,}925$ queries; RSSE applies query-success OR within each of $4{,}385$ partial samples and reports SSR without instance-count stratification. Standard deviations summarize variation across the three model runs. All percentage-point differences are computed from unrounded means.

B.4 Reproducibility materials

The HIAF implementation and dataset construction scripts are available at `https://anonymous.4open.science/r/HIAF`. The scan and query manifests define the splits and evaluation queries; the graph-generation configuration and class-eligibility rules specify the benchmark construction. Evaluation uses the ten fixed RSCQE class mappings and the predictions from each selected checkpoint.

C Baseline adaptations and ablation settings

C.1 Comparison on scene-class queries

Every method receives the observed scene and queried class, returns one ground-plane point, and is evaluated against all hidden instances of that class. The comparison uses the same training, validation, and test query sets. The final predictions are reused across ECQE, RSCQE, and RSSE, so differences between protocols arise from their evaluation units rather than a change in model input or prediction.

MLPC predicts pairwise distances from object-class and commonsense features; SCG-OL additionally uses spatial relations among observed objects (Giuliari et al., 2022). Both recover a target position in a separate localization stage. Graphormer (Ying et al., 2021) is used in the localization adaptation of Giuliari et al. (2023), with position regression from the updated query-node representation. D-SCG-OL predicts directed relative positions conditioned on observed objects and combines them into one output (Giuliari et al., 2023). HIAF reuses its encoder family, replacing the localization head and training objective. The matched objective control in Table 3 compares event supervision with set likelihood within the same Gaussian decoder.

Baseline supervision and hyperparameters. For multi-instance queries, each baseline selects the hidden instance nearest to its current localization output. D-SCG-OL and Graphormer minimize squared Euclidean position error to this instance. MLPC and SCG-OL minimize the mean squared error between predicted anchor-to-target distances and distances to the selected instance, averaged over observed objects. Each scene-class query contributes one loss, and query losses are averaged within each batch. We retain baseline hyperparameters explicitly reported in prior work (Giuliari et al., 2022; 2023) and tune those left unspecified. Checkpoints are selected by validation LSR@1m.

C.2 Hypothesis-interaction and objective ablations

HIAF + hypothesis interaction augments the main decoder with two self-attention blocks between the geometry embedding and final layer normalization. Each block has four attention heads and a feed-forward network of width $256 \to 1{,}024 \to 256$, with GELU, pre-normalization, residual connections, and dropout of 0.2. These permutation-equivariant operations follow the set-attention principle (Lee et al., 2019). The target-conditioned relational encoder, learned slots, provisional centers, geometry embedding, final normalization, prediction heads, Gaussian field, and continuous objective are retained in both versions. Here, hypothesis interaction refers to the added self-attention blocks; both versions score candidates using contributions from all hypotheses.

PPP objective. A PPP set negative log-likelihood uses the Gaussian intensity at the hidden target positions, up to constants independent of the parameters:

$$\mathcal{L}_{\text{set}} = \sum_{j=1}^{H} \alpha_j - \sum_{y \in Y} \log \lambda_\theta(y). \tag{27}$$

HIAF instead supervises the binary success event over action locations with Eq. (10). The shared total-mass term does not make these objectives equivalent. The PPP objective variant in Table 3 uses the main decoder without hypothesis interaction. It retains $K = 2$, the encoder, geometry embedding, final norm, heads, positive parameterization, and candidate-selection rule, with the same training and validation settings. Only the supervision objective changes. All three PPP runs complete the 100-epoch budget. For seeds 42, 100, and 123, validation selects epochs 61, 54, and 52, respectively. Both objectives use the hidden target set only for supervision; scoring remains at $r = 1\,\text{m}$ with one

prediction for all evaluation radii. The PPP loss does not require the event objective's Monte Carlo integral.

The PPP objective can be unbounded below as a positive-mass component concentrates at a training target and its scale tends to zero. The reported localization results do not establish whether such concentration occurs during training.

Hypothesis-count sensitivity. Table 4 compares $K \in \{1, 2, 4\}$ in the main decoder without hypothesis interaction. The encoder, geometry embedding, prediction heads, continuous event objective, optimization, and 100-epoch training budget follow Table 6. Each model is selected by validation ECQE LSR@1m. ECQE uses all 16,925 test queries; RSCQE uses the same ten fixed query samplings as the main comparison, averaged within each model run before aggregation across the three runs.

D Complementary evaluations

D.1 Tolerance sensitivity

Table 8 evaluates the same predictions at 0.5, 1.0, and 1.5 m. HIAF exceeds the four baselines in mean ECQE success at each reported radius. At 1.5 m, it reaches 43.32%, compared with 40.57% for D-SCG-OL, the strongest baseline at this radius.

Table 8: ECQE LSR (%) across evaluation radii, using the same point selected with $r = 1$ m. Mean $\pm$ standard deviation over three seeds; bold marks the best mean.

Method	0.5 m	1.0 m	1.5 m
MLPC	6.05 $\pm$ 0.31	19.08 $\pm$ 0.59	33.25 $\pm$ 0.73
SCG-OL	9.54 $\pm$ 0.20	24.79 $\pm$ 0.49	39.36 $\pm$ 0.47
Graphormer	8.68 $\pm$ 0.45	24.42 $\pm$ 0.92	39.98 $\pm$ 0.44
D-SCG-OL	9.44 $\pm$ 0.34	25.21 $\pm$ 0.55	40.57 $\pm$ 0.77
HIAF	**13.19 $\pm$ 0.37**	**30.06 $\pm$ 0.72**	**43.32 $\pm$ 0.50**

D.2 Conditional localization error by instance count

Table 9 complements the success rates in Table 5. HIAF has the lowest mSLE in both strata under ECQE and RSCQE, with each method evaluated on its own successful-query subset.

Table 9: mSLE (m) by hidden-instance count, conditional on each method's own 1 m successes. Mean $\pm$ standard deviation over three seeds; bold marks the lowest mean.

	ECQE		RSCQE	
Method	Single	Multiple	Single	Multiple
MLPC	0.642 $\pm$ 0.012	0.594 $\pm$ 0.005	0.660 $\pm$ 0.018	0.600 $\pm$ 0.008
SCG-OL	0.571 $\pm$ 0.008	0.595 $\pm$ 0.009	0.577 $\pm$ 0.010	0.601 $\pm$ 0.006
Graphormer	0.598 $\pm$ 0.010	0.608 $\pm$ 0.015	0.593 $\pm$ 0.008	0.607 $\pm$ 0.004
D-SCG-OL	0.590 $\pm$ 0.012	0.579 $\pm$ 0.007	0.574 $\pm$ 0.006	0.575 $\pm$ 0.013
HIAF	**0.547 $\pm$ 0.008**	**0.544 $\pm$ 0.011**	**0.526 $\pm$ 0.005**	**0.532 $\pm$ 0.017**

D.3 Additional hypothesis-interaction ablations

Table 10 extends the ablation in Table 3 across evaluation radii. Relative to the interaction variant, HIAF's 0.5 m ECQE success is higher by 1.08 percentage points (13.19% versus 12.11%). Its conditional mSLE is lower under all three protocols (Table 3). The interaction variant has higher ECQE success at 1.5 m. Changes across individual runs vary in direction; these comparisons are descriptive, without significance testing.

Table 10: ECQE LSR (%) with and without hypothesis interaction, using the same predictions across radii. Mean ± standard deviation over three seeds; bold marks the higher mean.

Variant	0.5 m	1.0 m	1.5 m
HIAF	**13.19 ± 0.37**	**30.06 ± 0.72**	43.32 ± 0.50
+ interaction	12.11 ± 0.32	29.96 ± 0.10	**44.04 ± 0.33**

Table 11 reports the same ablation by hidden-instance count. HIAF has slightly higher mean LSR@1m in both strata under ECQE and RSCQE. It also has lower conditional mSLE in both strata under both protocols. Each mSLE is computed on that variant's own successful subset.

Table 11: Interaction ablation by hidden-instance count at $K = 2$. mSLE uses each variant's own 1 m successes. Mean ± standard deviation over three seeds; bold marks the better mean.

	ECQE		RSCQE	
Variant	Single	Multiple	Single	Multiple
		LSR@1m (%) ↑		
HIAF	**28.04 ± 0.44**	**33.83 ± 1.73**	**30.53 ± 0.74**	**35.48 ± 2.40**
+ interaction	27.90 ± 0.41	33.81 ± 0.93	30.25 ± 1.34	35.30 ± 1.20
		mSLE (m) ↓		
HIAF	**0.547 ± 0.008**	**0.544 ± 0.011**	**0.526 ± 0.005**	**0.532 ± 0.017**
+ interaction	0.575 ± 0.007	0.566 ± 0.007	0.552 ± 0.011	0.564 ± 0.008

D.4 CANDIDATE-LEVEL QUALITATIVE COMPARISON

Figure 4 shows HIAF's final hypothesis centers and action fields for the three examples in Figure 3. Each candidate is scored using the contributions of all Gaussian hypotheses, and the highest-scoring center supplies the final point. Overlapping centers are shown at their actual positions.

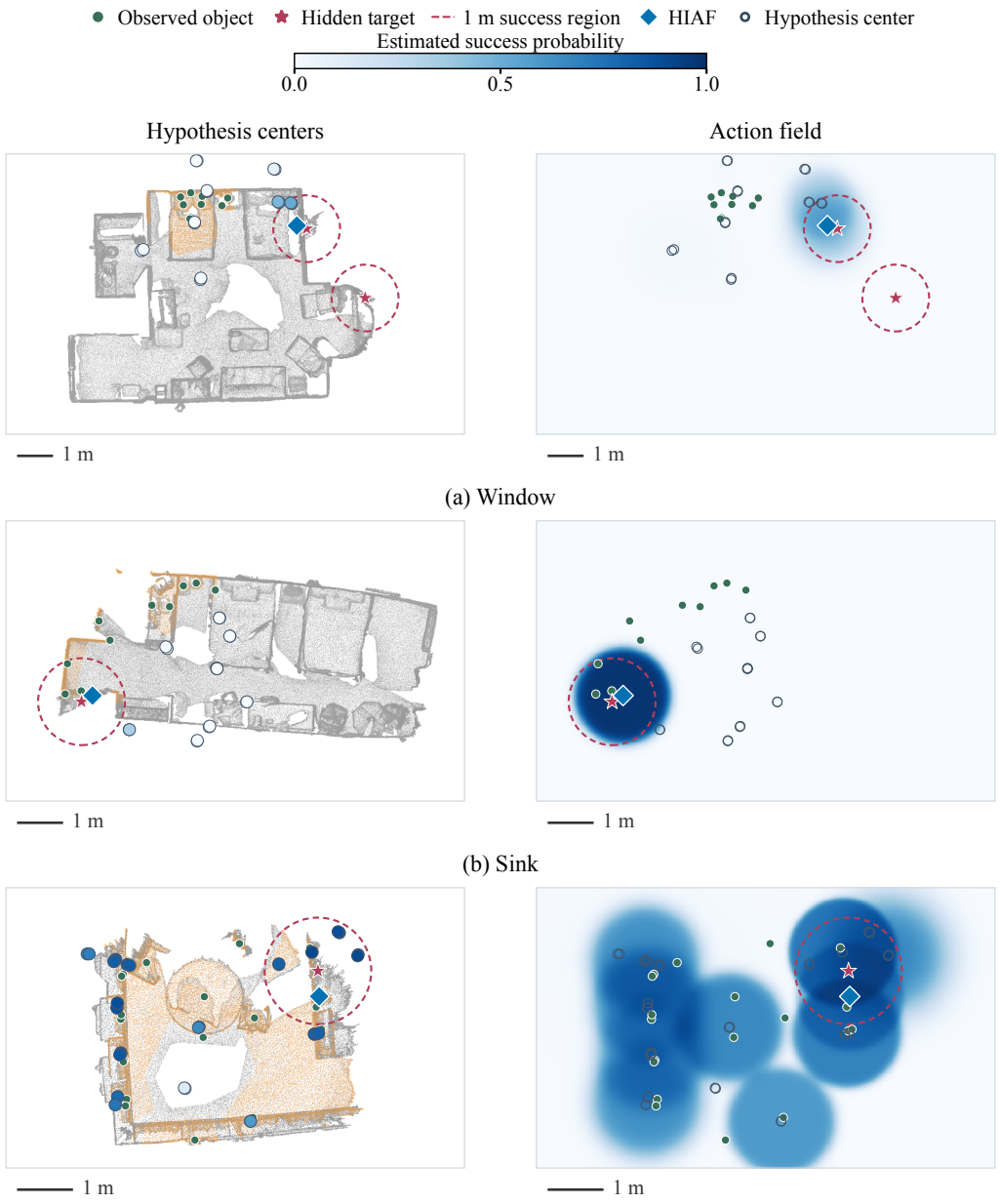


Figure 4: Hypothesis centers and action fields for the selected cases in Figure 3. Left: all final centers colored by their hit probabilities $E_\theta(\mu_j)$, overlaid on gray full-scene and orange partial point-cloud projections for visualization only. Right: the continuous action field $E_\theta(x)$ with candidate centers outlined. Blue diamonds mark the selected outputs; red stars and dashed circles show hidden targets and their 1 m success regions. Green circles mark observed objects. Both columns in each row share the same orientation, coordinate limits, and spatial scale; all panels share the probability scale.