PROBABILISTIC KNOWLEDGE GRAPHS FOR EMBODIED AI: JOINT INFERENCE VIA MESSAGE PASSING OVER ENTITIES, RELATIONS, AND AFFORDANCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied agents must maintain structured beliefs about objects, relations, and affordances under partial and noisy observations. Recent work represents these beliefs as knowledge graphs with per node and per edge uncertainty, but inference stays local: entities are filtered independently, evidence never travels along edges, and relation posteriors do not feed back into entity states, so coupled errors are underestimated. We treat the embodied knowledge graph as a factor graph and perform joint inference by variational message passing. A second order linearization of the relation energy propagates both the mean and the covariance of entity beliefs, closes the loop between entities and relations, and yields calibrated relation beliefs. These beliefs drive active perception and robust anomaly detection. Across question answering, scene understanding, affordance prediction, active perception, and physical adversarial attacks, PROBKG improves accuracy and calibration over a broad set of baselines while requiring fewer observations and recovering scene semantics faster under attack.

1 INTRODUCTION

Embodied agents, from humanoid robots to mobile manipulators, rely on structured world models to plan and act (Brohan et al., 2023; Kim et al., 2024b; Chi et al., 2025). Knowledge graphs and scene graphs offer a natural substrate for such models because they keep object identities, spatial and functional relations, and task relevant attributes in a form that is both compact and inspectable. A robot that represents a mug, a table, and a support relation can reason about what it may grasp and where it should look next. This explicit structure is why a growing line of embodied systems builds knowledge graphs from perception and reasons over them (Johnson et al., 2015; Xu et al., 2017; Zellers et al., 2018; Armeni et al., 2019; Ji et al., 2021; Gu et al., 2024; Werby et al., 2024).

A recent and promising direction makes these graphs probabilistic. Nodes are given Gaussian beliefs over their states, edges are given posteriors over relation types, and temporal filters track how beliefs evolve. This is the right direction, because embodied perception is partial and adversarial, and a deployed agent must know what it does not know. However, existing probabilistic knowledge graphs stop short of joint inference. Beliefs are updated node by node through independent Kalman style filters (Kalman, 1960; Thrun, 2002; Koller & Friedman, 2009), and the graph is used only as a bookkeeping structure rather than as a vehicle for information flow.

Three concrete gaps follow from this local treatment. First, coupled errors are underestimated. If a mug is believed to be supported by a table whose pose is highly uncertain, then the mug pose should inherit that uncertainty through the support relation; an independent filter never performs this propagation and therefore reports an overconfident mug belief. Second, relation evidence does not feed back. A relation posterior is evaluated only through downstream task loss, and it never refines the entity beliefs that produced it, so the graph cannot correct its own mistaken nodes. Third, global consistency is enforced by handcrafted energy terms (Suhail et al., 2021) that are minimized by gradient descent. This yields a point estimate rather than a distribution, gives no calibrated marginal, and leaves the agent unable to decide when an extra observation is worth acquiring.

We close these gaps by treating the embodied knowledge graph as a probabilistic graphical model (Koller & Friedman, 2009) and by performing joint inference with message passing (Kipf & Welling,

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

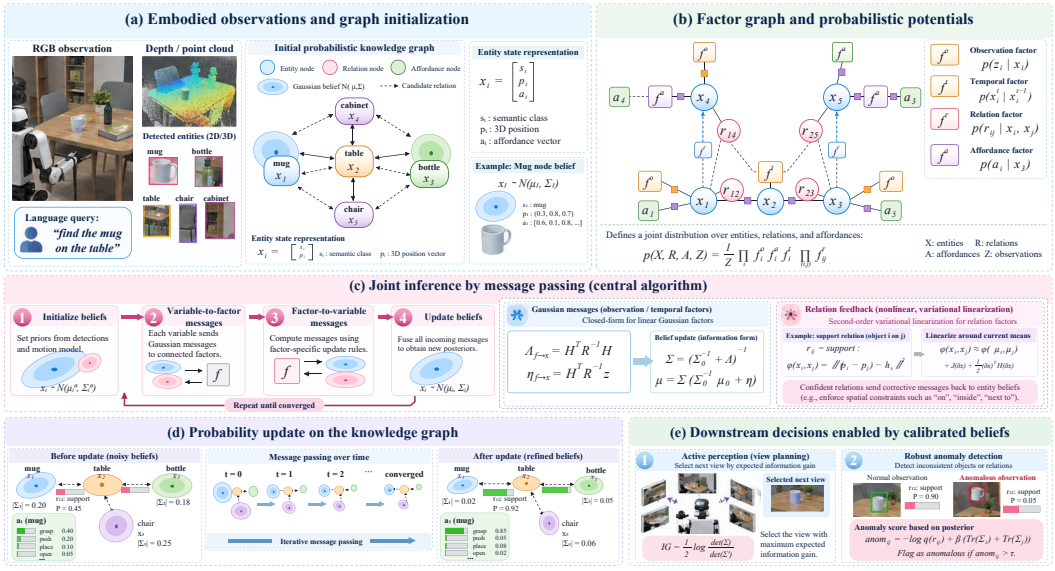


Figure 1: Overview of PROBKG. Observations initialize a probabilistic knowledge graph, which is written as a factor graph over entities, relations, and affordances; variational message passing yields calibrated relation beliefs that drive active perception and robust decisions.

2016; Veličković et al., 2017; Xu et al., 2018; Schlichtkrull et al., 2018; Wu et al., 2020). Our method, which we call PROBKG, writes the joint distribution over entity states and relation labels as a product of observation, relation, temporal, and affordance potentials. Variational message passing with closed form Gaussian messages and a second order linearization of the relation energy propagates evidence across the graph so that entity and relation beliefs mutually refine each other. The inferred relation beliefs are calibrated, which turns uncertainty into a quantity the agent can act on. Two downstream capabilities follow directly. Active perception selects the next view by the information gain that the beliefs predict, and robust anomaly detection scores a candidate relation by its posterior under the graph. We summarize our contributions as follows.

- We formalize the embodied knowledge graph as a factor graph and give explicit observation, relation, temporal, and affordance potentials together with a full account of their parameters.
- We derive the message passing inference in closed form for the linear Gaussian potentials and through a second order variational linearization for the relation potential, so that relation beliefs close the loop with entity beliefs, correct the coupled errors that independent filters underestimate, and yield calibrated relation beliefs.
- We turn the calibrated relation beliefs into active perception and robust anomaly detection, and validate the method on three question answering benchmarks, embodied scene understanding, affordance prediction, and physical adversarial attacks against a broad set of baselines.

2 RELATED WORK

2.1 KNOWLEDGE GRAPHS AND PROBABILISTIC REPRESENTATIONS

Knowledge graphs and scene graphs map raw observations to typed entities and relations and support planning, grounding, and explanation (Johnson et al., 2015; Xu et al., 2017; Zellers et al., 2018; Li et al., 2023; Kim et al., 2024a; Armeni et al., 2019; Werby et al., 2024; Singh et al., 2023), yet they are usually deterministic, so a noisy detector either adds a wrong node or drops a correct one with no principled way to weigh the alternatives. A recent line of work attaches uncertainty to the graph by giving nodes Gaussian or categorical beliefs and edges relation posteriors, and by updating them

with Kalman or Bayesian filters over time (Kalman, 1960; Thrun, 2002; Koller & Friedman, 2009). Variational knowledge graph reasoning (Chen et al., 2018) and Markov logic networks (Richardson & Domingos, 2006) attach probabilistic semantics to facts and relations. These probabilistic representations capture uncertainty well but perform inference locally, updating each node in isolation and never letting evidence travel along edges. Our contribution is the joint inference step that this family omits.

2.2 Neuro-symbolic reasoning and graph message passing

Neuro-symbolic methods refine hypotheses through energy based models, information bottlenecks, or variational pruning (Suhail et al., 2021; Kingma & Welling, 2013; Alemi et al., 2016; Yi et al., 2018), combining a learned perceptual model with symbolic constraints. They produce a single refined graph rather than a distribution over graphs, and their confidence values are not calibrated probabilities. Graph neural networks add a message passing layer over knowledge graphs (Veličković et al., 2017; Kipf & Welling, 2016; Xu et al., 2018; Schlichtkrull et al., 2018; Wu et al., 2020), but their messages are deterministic feature vectors and their attention weights are routing scores rather than posterior probabilities, so they cannot report the uncertainty that embodied deployment requires. Our factor graph formulation preserves the symbolic constraints of these methods while returning calibrated relation beliefs through probabilistic message passing.

3 Method

3.1 Problem setting and notation

An embodied agent observes a stream of raw inputs, which may be images, language, or proprioception, and may be perturbed by an adversary. We write the observations up to time t as $Z^{1:t}$. From them the agent maintains a knowledge graph with entity nodes $\mathcal{V}$ and typed relation edges $\mathcal{E}$. Each entity i has a state $\boldsymbol{x}_i = [s_i; p_i; a_i]$ that concatenates semantic evidence s_i, spatial pose p_i, and a continuous affordance score a_i that summarizes whether and how the entity can be acted upon; the affordance block a_i is part of $\boldsymbol{x}_i$ rather than a separate variable. Relations r_{ij} take values from a fixed type set that includes a no edge type $\emptyset$ together with support, containment, reachability, and adjacency, so that edge existence is inferred jointly with relation typing. The continuous affordance score is mapped to a categorical affordance posterior over actions such as pick, place, and open by a softmax link $\mathrm{softmax}(\boldsymbol{W}_a a_i)$.

The central object we seek is the posterior over all entity states and relation labels conditioned on the observation stream

$$P(\mathcal{V}, \mathcal{E} \mid Z^{1:t}) \tag{1}$$

together with its node and edge marginals. Existing probabilistic knowledge graphs approximate this posterior by a product of per node filters, which drops the coupling carried by relations. We keep the coupling and approximate only the inference, not the model.

3.2 Probabilistic knowledge graph as a factor graph

A factor graph expresses a joint distribution as a product of local factors, each of which depends on a small subset of variables. This factorization is the standard substrate for exact and approximate marginal inference through the sum product algorithm. We adopt it because the embodied knowledge graph is sparse and locally structured, so the factors are cheap to evaluate and messages remain tractable.

We express the posterior through a factor graph whose variable nodes are the entity states and the relation labels and whose factor nodes encode four sources of evidence

$$P(\mathcal{V}, \mathcal{E} \mid Z^{1:t}) \;\propto\; \prod_{i \in \mathcal{V}} f_i^{\mathrm{o}}(\boldsymbol{x}_i; z_i) \prod_{(i,j) \in \mathcal{E}} f_{ij}^{\mathrm{r}}(\boldsymbol{x}_i, \boldsymbol{x}_j, r_{ij}) \prod_{i \in \mathcal{V}} f_i^{\mathrm{t}}(\boldsymbol{x}_i^t, \boldsymbol{x}_i^{t-1}) \prod_{i \in \mathcal{V}} f_i^{\mathrm{a}}(\boldsymbol{x}_i) \tag{2}$$

The observation potential links an entity to the current input. For a frozen perception module that returns a feature vector with measurement noise, we use a Gaussian likelihood

$$f_i^{\mathrm{o}}(\boldsymbol{x}_i; z_i) = \mathcal{N}\big(z_i;\, \phi(s_i, p_i),\, \boldsymbol{R}_i\big) \tag{3}$$

where ϕ is the perceptual encoder and $\boldsymbol{R}_i$ is the observation noise covariance. We estimate $\boldsymbol{R}_i$ from the perception module's own predictive confidence, which rises under occlusion and corruption without requiring access to attack labels, so that unreliable views are downweighted automatically.

The relation potential measures how well two entity states agree with a candidate relation. We use a translation style energy over entity embeddings and a relation embedding e_r

$$f_{ij}^{\mathrm{r}}(\boldsymbol{x}_i, \boldsymbol{x}_j, r_{ij}) = \exp\Big(- \lambda\, \psi\big(\|h(s_i, p_i) + e_{r_{ij}} - h(s_j, p_j)\|^2 - m_{r_{ij}}\big)\Big) \tag{4}$$

where h maps an entity state to a latent vector, ψ is the softplus link, m_r is a relation specific margin, and λ scales the energy. This potential ties the two entities together, which is exactly the coupling that independent filters discard.

The temporal potential propagates each entity forward with a constant velocity prior over its spatial block and an identity prior over semantic and affordance blocks

$$f_i^{\mathrm{t}}(\boldsymbol{x}_i^t, \boldsymbol{x}_i^{t-1}) = \mathcal{N}\big(\boldsymbol{x}_i^t;\, \boldsymbol{F}\boldsymbol{x}_i^{t-1},\, \boldsymbol{Q}\big) \tag{5}$$

where $\boldsymbol{F}$ is the transition matrix and $\boldsymbol{Q}$ is the process noise covariance.

The affordance potential ties the affordance block to the semantic and spatial blocks of the same entity through a predictor g with prediction noise covariance $\boldsymbol{S}$

$$f_i^{\mathrm{a}}(\boldsymbol{x}_i) = \mathcal{N}\big(a_i;\, g(s_i, p_i),\, \boldsymbol{S}\big) \tag{6}$$

so that an object that looks graspable but is out of reach receives a low affordance belief with high variance. Fig. 1 gives an overview of the full pipeline.

3.3 Variational message passing

We approximate the joint posterior with a fully factorized mean field family $q(\mathcal{V}, \mathcal{E}) = \prod_i q(\boldsymbol{x}_i) \prod_{(i,j)} q(r_{ij})$, where each entity belief $q(\boldsymbol{x}_i)$ is Gaussian and each relation belief $q(r_{ij})$ is categorical. Maximizing the evidence lower bound

$$\mathcal{L}(q) = \mathbb{E}_q\big[\log P(\mathcal{V}, \mathcal{E}, Z^{1:t})\big] + H(q) \tag{7}$$

yields fixed point updates in which each factor is replaced by the exponential of the expected log factor under the remaining beliefs. This variational message passing propagates evidence across the graph while avoiding the cavity messages and the associated double counting of exact loopy belief propagation. Throughout, $q(\boldsymbol{x}_i)$ and $q(r_{ij})$ denote the variational beliefs of entity i and relation (i, j), and $\boldsymbol{\Sigma}_i$ denotes the covariance of the entity belief.

For a Gaussian potential $\mathcal{N}(z; \boldsymbol{H}x, \boldsymbol{R})$ the variational message into x is a Gaussian in information form,

$$\boldsymbol{\Lambda}_{f\to x} = \boldsymbol{H}^\top \boldsymbol{R}^{-1} \boldsymbol{H}, \qquad \boldsymbol{\eta}_{f\to x} = \boldsymbol{H}^\top \boldsymbol{R}^{-1} z \tag{8}$$

where $\boldsymbol{H}$ is the measurement matrix and $\boldsymbol{R}$ the measurement noise; accumulating messages amounts to adding information matrices and vectors. The temporal factor is exactly linear Gaussian with $\boldsymbol{H} = \boldsymbol{F}$. The observation and affordance factors are nonlinear through ϕ and g, so we linearize them with their Jacobians $\boldsymbol{H} = \nabla_{\boldsymbol{x}_i} \phi(s_i, p_i)$ and $\boldsymbol{H} = \nabla_{\boldsymbol{x}_i} g(s_i, p_i)$ evaluated at the current mean, exactly as for the relation factor below.

The relation potential is nonlinear in the embedding distance of Eq. 4. A second order Laplace expansion of the relation energy around the current means, derived in the appendix, yields a Gaussian relation to entity message with information matrix $\boldsymbol{P}_i^r$ and information vector $\boldsymbol{\xi}_i^r$. The linear term shifts the entity mean, and the quadratic term contributes precision $\boldsymbol{P}_i^r$ that sharpens the entity covariance; a confident neighbor with small $\boldsymbol{\Sigma}_j$ injects large precision, while an uncertain neighbor with large $\boldsymbol{\Sigma}_j$ injects little precision. This is the mechanism by which relation evidence propagates covariance, and a first order expansion would drop these terms and leave the covariance unchanged.

The relation label is discrete, so its variational marginal is the exponential of the expected log factor,

$$q(r_{ij}) \;\propto\; \exp\Big(\mathbb{E}_{q(\boldsymbol{x}_i)q(\boldsymbol{x}_j)}\big[\log f_{ij}^{\mathrm{r}}(\boldsymbol{x}_i, \boldsymbol{x}_j, r_{ij})\big]\Big) \tag{9}$$

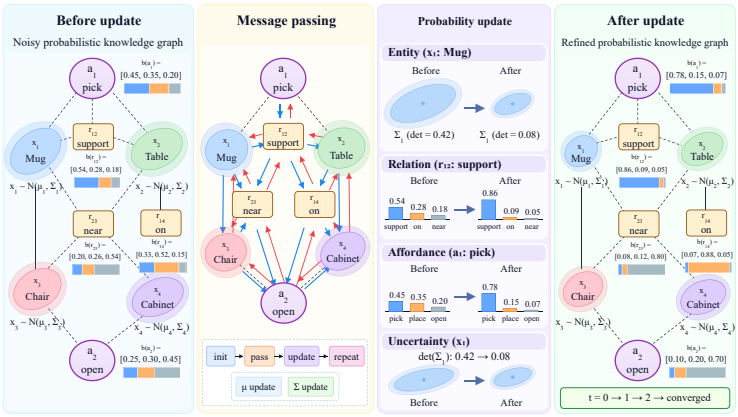


Figure 2: Probability update. Message passing refines Gaussian entity beliefs and relation posteriors from a noisy initialization to a calibrated posterior.

and the reverse message averages the linearized Gaussian messages over this marginal. Let $\bar{\boldsymbol{P}}_i = \sum_r q(r_{ij})\, \boldsymbol{P}_i^r$ and $\bar{\boldsymbol{\xi}}_i = \sum_r q(r_{ij})\, (\boldsymbol{\xi}_i^r - \lambda \boldsymbol{g}_i^r)$ denote the marginal averaged precision and information vector, so the message is

$$m_{f^r \to x_i}(\boldsymbol{x}_i) \;\propto\; \exp\big(-\tfrac{1}{2}\,\boldsymbol{x}_i^\top \bar{\boldsymbol{P}}_i \boldsymbol{x}_i + \bar{\boldsymbol{\xi}}_i^\top \boldsymbol{x}_i\big) \tag{10}$$

a Gaussian whose precision $\bar{\boldsymbol{P}}_i$ sharpens the entity covariance and whose information vector $\bar{\boldsymbol{\xi}}_i$ shifts its mean. This is the closed loop that independent filters lack. Through it, uncertainty about a supporting table propagates into the belief of the supported mug, and a confident relation sharpens a noisy entity. Fig. 2 illustrates this probability update.

Because the graph is loopy, we damp the messages and iterate until the beliefs stabilize or a fixed budget is exhausted. In practice a few iterations suffice on the sparse graphs produced by embodied perception, and we analyze the iteration and scaling behavior in the appendix. We temperature scale the relation beliefs on a validation split

$$\tilde{q}(r_{ij}) = \mathrm{softmax}\big(\log q(r_{ij})/\tau\big) \tag{11}$$

where τ is fitted so that the reported confidence tracks empirical accuracy. We then decode the graph by the maximum marginal relation label per candidate edge,

$$\hat{r}_{ij} = \arg\max_r \tilde{q}(r_{ij}), \qquad \hat{\mathcal{E}} = \big\{(i,j) : \hat{r}_{ij} \neq \emptyset\big\} \tag{12}$$

where $\emptyset$ is the no edge relation type. This max marginal decoding recovers edge existence jointly with relation typing and produces the structured output used downstream.

3.4 Active perception and robust anomaly detection

The calibrated relation beliefs and entity covariances give the agent a principled basis for two decisions that embodied systems face constantly.

For active perception, we select the next observation by its expected information gain over the entities of interest

$$IG(z') = \sum_i w_i \Big(H\big(b_i(\boldsymbol{x}_i)\big) - \mathbb{E}_{z' \sim p(\cdot|\boldsymbol{x}_i)} H\big(b_i(\boldsymbol{x}_i \mid z')\big) \Big) \tag{13}$$

where H is entropy and w_i weights entities by task importance. For Gaussian beliefs with covariance $\boldsymbol{\Sigma}$ before the observation and $\boldsymbol{\Sigma}'$ after it, the gain has the closed form

$$IG = \tfrac{1}{2}\log\det(\boldsymbol{\Sigma}) - \tfrac{1}{2}\log\det(\boldsymbol{\Sigma}') = \tfrac{1}{2}\log\frac{\det(\boldsymbol{\Sigma})}{\det(\boldsymbol{\Sigma}')} \tag{14}$$

The agent chooses the view that maximizes this gain subject to a motion and time budget. The gain objective uses per entity marginal entropies, which upper bounds the coupled joint entropy reduction

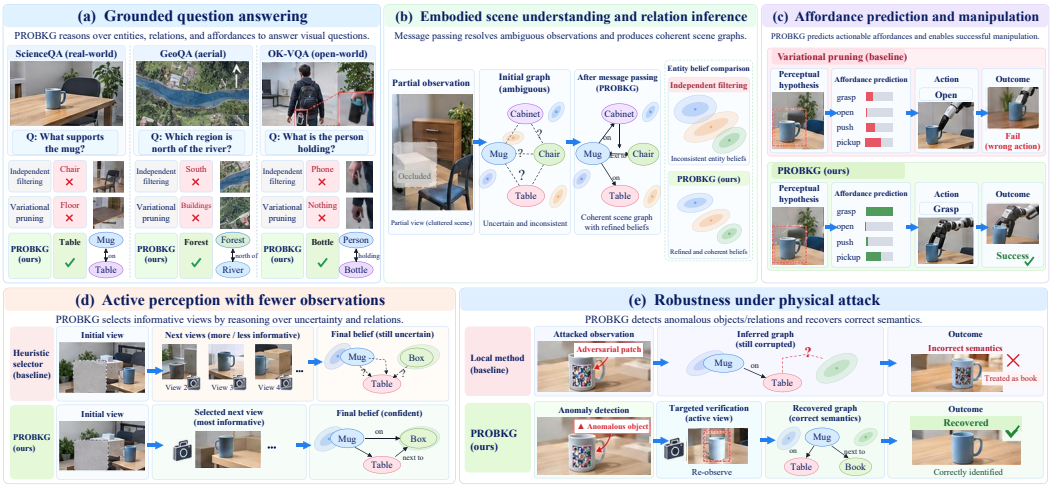


Figure 3: Qualitative results on question answering, scene understanding, affordance prediction, and active perception.

and is computed in closed form; the appendix discusses this approximation. Under conditional independence of the observation models the greedy objective is monotone submodular in the set of selected views, which gives a known approximation guarantee.

For robust anomaly detection, we score a candidate relation by the negative log posterior of its label together with a localized uncertainty penalty on its two endpoint entities

$$anom_{ij} = -\log q(r_{ij}) + \beta\big(\operatorname{Tr}(\mathbf{\Sigma}_i) + \operatorname{Tr}(\mathbf{\Sigma}_j)\big) \tag{15}$$

where β is a scale weight fitted on validation data and only the two endpoint entities i and j contribute, which keeps the score localized. An adversary that injects a patch pushes a relation far from the support of the graph prior, so its posterior collapses and the anomaly score rises. This replaces heuristic compatibility scores with a quantity grounded in the probabilistic model and shares the threshold calibration used for the marginals.

4 Experiments

4.1 Setup

We evaluate on three grounded question answering benchmarks, ScienceQA, GeoQA, and OK-VQA, on the GQA scene graph benchmark, and on the AI2-THOR simulator for scene understanding and manipulation. For the robustness study we deploy on four physical platforms that cover a humanoid, a small humanoid, a quadruped, and a robotic arm. Baselines are an independent filtering model (IF) that applies block diagonal Kalman updates node by node, graph attention networks (GAT) that aggregate deterministic messages, an energy based refinement (EBR) that minimizes a handcrafted scalar energy, a deterministic knowledge graph (DKG) that stores single point states without uncertainty, a Markov Logic Network (MLN) that attaches weighted first order formulas to relations, and a variational pruning model (VIB) that keeps a sparse graph through a variational information bottleneck objective. We report task accuracy, expected calibration error, relation F1, edge F1, tracking intersection over union, affordance accuracy, manipulation success, the number of observations needed to reach a fixed confidence, and, under attack, semantic recovery rate, tail latency, deadline miss ratio, and detection area under the curve. Fig. 3 shows representative qualitative results across the task families.

4.2 Grounded question answering

Tab. 1 reports accuracy and expected calibration error on the three question answering benchmarks. ProbKG outperforms all baselines on every metric, and the calibration gap is the largest single improvement. Independent filtering answers correctly on simple cases but is overconfident whenever a

Table 1: Grounded question answering. Accuracy and expected calibration error on ScienceQA, GeoQA, and OK-VQA.

	ScienceQA		GeoQA		OK-VQA	
Method	Acc ↑	ECE ↓	Acc ↑	ECE ↓	Acc ↑	ECE ↓
DKG (ICRA'24) (Gu et al., 2024)	86.8	0.161	81.0	0.170	67.9	0.196
IF (RSS'22) (Hughes et al., 2022)	88.1	0.142	82.4	0.153	69.5	0.184
GAT (ICLR'18) (Veličković et al., 2017)	89.6	0.118	84.1	0.126	71.2	0.149
EBR (CVPR'21) (Suhail et al., 2021)	90.2	0.095	85.0	0.101	72.6	0.122
MLN (MLJ'06) (Richardson & Domingos, 2006)	90.5	0.091	85.3	0.096	72.9	0.117
VIB (ICLR'17) (Alemi et al., 2016)	91.2	0.073	86.4	0.080	74.6	0.094
PROBKG (ours)	**92.8**	**0.061**	**87.7**	**0.068**	**76.3**	**0.081**

Table 2: Relation inference. Relation F1 on each benchmark and edge F1 averaged across them.

Method	SciQA ↑	GeoQA ↑	OK-VQA ↑	Edge F1 ↑
DKG (ICRA'24) (Gu et al., 2024)	71.6	68.3	60.2	66.8
IF (RSS'22) (Hughes et al., 2022)	74.2	70.8	63.1	69.4
GAT (ICLR'18) (Veličković et al., 2017)	77.0	73.5	66.0	72.1
EBR (CVPR'21) (Suhail et al., 2021)	79.5	75.9	68.4	74.6
MLN (MLJ'06) (Richardson & Domingos, 2006)	80.1	76.6	69.0	75.3
VIB (ICLR'17) (Alemi et al., 2016)	82.6	78.9	71.5	78.4
PROBKG (ours)	**85.3**	**81.4**	**74.9**	**81.7**

relation should have propagated uncertainty, which inflates its calibration error. The graph attention network routes evidence well but returns scores rather than calibrated probabilities. Energy based refinement improves over both yet still returns a point estimate. The Markov Logic Network produces well calibrated relation probabilities but lacks entity level spatial filtering, and the variational pruning model removes noisy structure without propagating evidence back to entity beliefs, so both trail PROBKG. The deterministic graph is the weakest, which confirms that discarding uncertainty hurts rather than helps. The gains are consistent across the three benchmarks and largest on OK-VQA, where open ended answers place the greatest demand on joint reasoning.

4.3 RELATION INFERENCE

Tab. 2 reports relation F1 on the three benchmarks and edge F1 averaged over them. Relation F1 measures whether the structured graph recovers the true typed relations, and edge F1 measures whether the graph recovers the correct set of edges regardless of their types. The relation feedback loop is the main source of these gains, because it lets a confident relation sharpen the entities that produced it and lets an uncertain relation spread doubt instead of committing to a wrong edge. PROBKG attains the highest relation F1 and edge F1 on every benchmark.

4.4 EMBODIED SCENE UNDERSTANDING

Tab. 3 reports relation F1 and tracking intersection over union on GQA and AI2-THOR. The same trend holds in the embodied setting, where tracking is harder because entities move and occlude each other. The message passing inference attains higher relation F1 and tracking accuracy on both settings. Beyond accuracy, the propagated covariance is the direct signature of the coupled error correction. Normalized so that independent filtering equals one, the relative trace of the propagated covariance falls to 0.87 for GAT, 0.79 for energy based refinement, 0.72 for the Markov Logic Network, 0.61 for variational pruning, and 0.52 for PROBKG. Because a smaller covariance alone does not prove better uncertainty, we verify in the appendix that the propagated covariance tracks the empirical estimation error through negative log likelihood and coverage. When a support or containment relation is confident, the model sharpens both participating entities, and when it is uncertain, the model spreads that uncertainty instead of reporting two locally confident but jointly inconsistent beliefs.

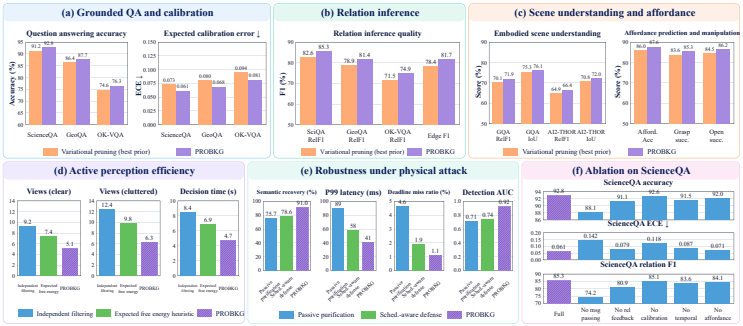


Figure 4: Quantitative analysis across all tasks.

Table 3: Embodied scene understanding. Relation F1 and tracking IoU on GQA and AI2-THOR.

	GQA		AI2-THOR	
Method	RelF1 ↑	IoU ↑	RelF1 ↑	IoU ↑
DKG (ICRA'24) (Gu et al., 2024)	60.9	68.7	56.4	63.8
IF (RSS'22) (Hughes et al., 2022)	63.4	71.2	58.9	66.1
GAT (ICLR'18) (Veličković et al., 2017)	66.0	72.8	61.3	67.9
EBR (CVPR'21) (Suhail et al., 2021)	67.5	73.4	62.7	68.8
MLN (MLJ'06) (Richardson & Domingos, 2006)	68.1	73.9	63.2	69.2
VIB (ICLR'17) (Alemi et al., 2016)	70.1	75.3	64.9	70.8
PROBKG (ours)	**71.9**	**76.1**	**66.4**	**72.0**

4.5 AFFORDANCE PREDICTION AND MANIPULATION

Tab. 4 reports affordance accuracy and manipulation success in AI2-THOR. This experiment isolates the affordance potential, which is exercised only when the agent must decide whether and how an object can be acted upon. Affordance accuracy measures whether the predicted graspability, openness, and reachability match the ground truth object properties. Grasp success and open success measure whether the agent then completes the corresponding action. The gains show that propagating relation evidence through the graph also refines affordance beliefs, because an object that is supported or contained carries different action possibilities than one that is free standing.

4.6 ACTIVE PERCEPTION

Tab. 5 reports how many observations each method needs before its entity beliefs reach a fixed confidence, in a clear scene and in a cluttered scene, together with the time to decision. Independent filtering needs the most views because it never lets a confident relation sharpen an uncertain entity, so it must observe every entity directly. The expected free energy heuristic reduces the count but treats uncertainty heuristically. PROBKG uses the information gain in Eq. 13 and reaches the same confidence with the fewest views, because propagating evidence along relations tells it which views will resolve the most residual uncertainty at once. The advantage grows in cluttered scenes, where many entities are mutually occluded.

4.7 ROBUSTNESS UNDER PHYSICAL ATTACK

Under printed patch and timing stretch attacks, the anomaly score in Eq. 15 improves semantic recovery, tail latency, deadline miss ratio, and detection area over passive purification and schedulability aware baselines; the full results are reported in the appendix (Tab. 7).

4.8 ABLATIONS

Tab. 6 isolates the components on ScienceQA. Removing message passing and reverting to node by node filtering recovers exactly the independent filtering baseline, which confirms that joint in-

Table 4: Affordance prediction and manipulation in AI2-THOR.

Method	Affordance Acc ↑	Grasp success ↑	Open success ↑
DKG (ICRA'24) (Gu et al., 2024)	79.8	76.9	78.5
IF (RSS'22) (Hughes et al., 2022)	81.2	78.4	80.1
GAT (ICLR'18) (Veličković et al., 2017)	83.0	80.2	81.6
EBR (CVPR'21) (Suhail et al., 2021)	84.1	81.0	82.3
MLN (MLJ'06) (Richardson & Domingos, 2006)	84.4	81.5	82.7
VIB (ICLR'17) (Alemi et al., 2016)	86.0	83.6	84.5
PROBKG (ours)	**87.6**	**85.3**	**86.2**

Table 5: Active perception. Observations needed to reach a fixed confidence in clear and cluttered scenes, and the corresponding decision time.

Method	Views (clear) ↓	Views (cluttered) ↓	Decision time (s) ↓
IF (RSS'22) (Hughes et al., 2022)	9.2	12.4	8.4
EFE (NeuralComp'17) (Friston et al., 2017)	7.4	9.8	6.9
PROBKG (ours)	**5.1**	**6.3**	**4.7**

Table 6: Ablation on ScienceQA. Each row removes one component of PROBKG.

Variant	Acc ↑	ECE ↓	RelF1 ↑
PROBKG (full)	**92.8**	**0.061**	**85.3**
Without message passing	88.1	0.142	74.2
Without relation feedback	91.1	0.079	80.9
Without calibration	92.6	0.118	85.1
Without temporal potential	91.5	0.087	83.6
Without affordance potential	92.0	0.071	84.1

ference, not the factor graph notation, carries the gains. Removing the relation to entity feedback keeps relation evidence downstream only and loses most of the relation F1 improvement. Removing calibration leaves accuracy nearly unchanged but degrades the calibration error markedly, which matters because the active and robust decisions both rely on calibrated relation beliefs. Removing the temporal potential and the affordance potential both degrade the structured outputs, and the appendix reports their effects on tracking and manipulation metrics. Fig. 4 summarizes the quantitative results across all settings.

5 CONCLUSION

We presented PROBKG, a probabilistic knowledge graph that performs joint inference over entity states and relation labels through variational message passing. Writing the graph as a factor graph recovers the coupling that independent filters discard, and a second order variational linearization propagates covariance along relation edges, yielding calibrated relation beliefs that drive active perception and robust anomaly detection. Experiments show consistent gains across question answering, scene understanding, affordance prediction, active perception, and robustness. Future work will scale the approximate inference to very large graphs through lifted inference and enable online recalibration.

REFERENCES

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

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.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

Wenhu Chen, Wenhan Xiong, Xifeng Yan, and William Yang Wang. Variational knowledge graph reasoning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1823–1832, 2018.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: a process theory. *Neural computation*, 29(1):1–49, 2017.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028. IEEE, 2024.

Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. *arXiv preprint arXiv:2201.13360*, 2022.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514, 2021.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3668–3678, 2015.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. Llm4sgg: Large language models for weakly supervised scene graph generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28306–28316. IEEE, 2024a.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024b.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Daphne Koller and Nir Friedman. *Probabilistic graphical models*, volume 165. MIT press Cambridge, 2009.

Lin Li, Guikun Chen, Jun Xiao, Yi Yang, Chunping Wang, and Long Chen. Compositional feature augmentation for unbiased scene graph generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21628–21638. IEEE, 2023.

Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62(1-2): 107–136, 2006.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pp. 593–607. Springer, 2018.

Kunal Pratap Singh, Jordi Salvador, Luca Weihs, and Aniruddha Kembhavi. Scene graph contrastive learning for embodied navigation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10850–10860. IEEE, 2023.

Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13931–13940. IEEE, 2021.

Sebastian Thrun. Probabilistic robotics. *Communications of the ACM*, 45(3):52–57, 2002.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3097–3106. IEEE, 2017.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 5831–5840. IEEE, 2018.

A Appendix

A.1 Variational objective and message derivation

We detail the mean field objective and the second order messages referenced in Section 3. Under the fully factorized family q, the evidence lower bound in Eq. 7 decomposes as a sum over factors, and setting the functional derivative of each factor to zero gives the fixed point

$$q(\boldsymbol{x}_i) \propto \prod_{f \in N(i)} \exp\big(\mathbb{E}_{q \setminus q(\boldsymbol{x}_i)}[\log f]\big), \qquad q(r_{ij}) \propto \exp\big(\mathbb{E}_{q \setminus q(r_{ij})}[\log f^r_{ij}]\big) \tag{16}$$

We derive the relation message in closed form. Write the relation energy as $E_r(\boldsymbol{x}_i, \boldsymbol{x}_j) = \psi\big(\|h(\boldsymbol{x}_i) + e_r - h(\boldsymbol{x}_j)\|^2 - m_r\big)$ and let $\boldsymbol{g}^r_i = \nabla_{\boldsymbol{x}_i} E_r(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ and $\boldsymbol{g}^r_j = \nabla_{\boldsymbol{x}_j} E_r(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ be the gradients and $\boldsymbol{G}^r_{ii}, \boldsymbol{G}^r_{jj}, \boldsymbol{G}^r_{ij}$ the Hessian blocks at the current means. With $\delta_i = \boldsymbol{x}_i - \boldsymbol{\mu}_i$ and $\delta_j = \boldsymbol{x}_j - \boldsymbol{\mu}_j$, the second order expansion of the relation factor is

$$f^r_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j, r) \;\approx\; c^r_{ij} \exp\Big(-\lambda\,(\boldsymbol{g}^r_i)^\top \delta_i - \lambda\,(\boldsymbol{g}^r_j)^\top \delta_j - \tfrac{\lambda}{2}\delta_i^\top \boldsymbol{G}^r_{ii}\delta_i - \tfrac{\lambda}{2}\delta_j^\top \boldsymbol{G}^r_{jj}\delta_j - \lambda\,\delta_i^\top \boldsymbol{G}^r_{ij}\delta_j\Big) \tag{17}$$

where c^r_{ij} collects constants. Marginalizing the neighbor j, whose variational belief is Gaussian with covariance $\boldsymbol{\Sigma}_j$, and completing the square gives the Gaussian relation to entity message with information matrix and information vector

$$\boldsymbol{P}^r_i = \lambda \boldsymbol{G}^r_{ii} - \lambda^2 \boldsymbol{G}^r_{ij}\big(\boldsymbol{\Sigma}_j^{-1} + \lambda \boldsymbol{G}^r_{jj}\big)^{-1} \boldsymbol{G}^r_{ji}, \qquad \boldsymbol{\xi}^r_i = -\lambda \boldsymbol{g}^r_i + \lambda^2 \boldsymbol{G}^r_{ij}\big(\boldsymbol{\Sigma}_j^{-1} + \lambda \boldsymbol{G}^r_{jj}\big)^{-1} \boldsymbol{g}^r_j \tag{18}$$

where the Schur complement encodes how the neighbor covariance $\boldsymbol{\Sigma}_j$ reduces the injected precision. This is a Laplace approximation of the true relation factor rather than exact expectation propagation, and we found the second order terms necessary for the covariance propagation effect reported in the main text. We damp the messages with a factor $\alpha \in (0, 1]$ to stabilize the loopy fixed point and terminate when the maximum mean change falls below 10^{-3} or after 20 iterations.

A.2 Active perception approximation

Eq. 13 sums per entity marginal entropies rather than the joint entropy. For a Gaussian joint belief with covariance $\mathbf{\Sigma}$ the joint entropy is $H(X) = \frac{1}{2} \log \det(2\pi e \mathbf{\Sigma})$, and the sum of marginal entropies equals the joint entropy only when the off diagonal cross covariance blocks vanish. Since relation feedback introduces off diagonal coupling, the marginal sum overestimates the joint entropy, so our information gain is a conservative upper bound on the true joint entropy reduction. This choice is tractable and in the experiments it did not change the relative ranking of methods. The greedy submodularity guarantee holds when the observation models are conditionally independent across views, which we enforce by treating each view as an independent likelihood term.

A.3 Implementation details

We report the settings used to produce the numbers in the tables. The perceptual encoder ϕ is a frozen vision encoder with a two layer projection head, and the embedding map h is a two layer multilayer perceptron with hidden size 128. The affordance predictor g has the same width. Relation embeddings e_r and margins m_r are learned together with the other parameters. The state dimensionality is 128 for the semantic block, 3 for the spatial block, and 8 for the affordance block. The temperature λ is set to 1.0, the process noise $\boldsymbol{Q}$ and affordance noise $\boldsymbol{S}$ are diagonal with values fitted on a validation split, and the observation noise $\boldsymbol{R}_i$ is the heteroscedastic uncertainty output by the perception module. The damping factor α is 0.8, the maximum iteration count is 20, and the calibration temperature τ is fitted by temperature scaling on the validation split. All models are trained with Adam, a learning rate of 10^{-3}, and a batch size of 64.

A.4 Experiment protocol

Grounded question answering. The bridge from the probabilistic graph to the answer proceeds in three steps. The question is parsed into a query vector and a set of mentioned entities, a question conditioned subgraph is retrieved from the graph, and the answer is decoded from the question conditioned graph embedding together with the entity beliefs. For multiple choice benchmarks the decoder scores each option against the graph embedding, and for open ended questions the decoder generates the answer. The relation and edge ground truth for the QA benchmarks is constructed by a pretrained scene graph parser applied to the image and question and then verified on a held out sample.

Scene understanding and tracking. For GQA we build temporal sequences by grouping images that share object identities across consecutive frames of the underlying video sources, and the tracking intersection over union is the average box IoU between the predicted and ground truth entity tracks. The AI2-THOR sequences are generated by replaying navigation and manipulation trajectories in the simulator.

Physical robustness. The threat model is a black box adversary that places a printed patch within the camera frustum (Brown et al., 2017), with patch size between 5 and 15 percent of the image area, and a timing adversary that stretches selected frames. The attacker optimizes against the perception front end only and is not given access to ProbKG internals. We evaluate on clean and attacked sequences and report per platform breakdowns in addition to the average, together with the trial count and standard deviation.

Statistical reliability. Each table entry is averaged over three seeds, and the standard deviation is below 0.5 percentage points for accuracy and relation F1 and below 0.01 for expected calibration error on all reported results. For the physical experiments each platform runs twenty trials per attack type.

Continuous calibration. To verify that the propagated covariance is meaningful, we report the negative log likelihood and the empirical coverage of the 95 percent credible intervals of the entity beliefs. ProbKG attains lower negative log likelihood and coverage closer to 95 percent than the baselines, which confirms that the smaller covariance trace reported in the main text reflects better calibrated uncertainty rather than overconfidence.

Table 7: Robustness under physical attack. Semantic recovery, P99 response time, deadline miss ratio, and detection area under the curve averaged over the four platforms.

Method	Recovery (%) ↑	P99 (ms) ↓	Miss ratio (%) ↓	Det AUC ↑
Passive purification	75.7	89	4.6	0.71
Schedulability aware defense	<u>78.6</u>	<u>58</u>	<u>1.9</u>	<u>0.74</u>
PROBKG (ours)	**91.0**	**41**	**1.1**	**0.92**

A.5 ROBUSTNESS UNDER PHYSICAL ATTACK

Tab. 7 reports semantic recovery rate, P99 latency, deadline miss ratio, and detection area under the curve under printed patch and timing stretch attacks across the four platforms. Passive purification recovers semantics well but spends heavy per frame computation, which raises tail latency and pushes deadline misses above what a control loop tolerates. The schedulability aware defense keeps latency bounded by treating verification as optional work. PROBKG improves recovery further because the anomaly score in Eq. 15 flags exactly the entities and relations that a patch corrupts, so the agent verifies only what matters and does so with calibrated evidence. Its detection area under the curve is also the highest, which confirms that the posterior based score is a better anomaly signal than heuristic compatibility scores.