Remember with Confidence: Uncertainty Quantification for Spatio-temporal Memory with Probabilistic Guarantees

Harry Zhang* **Nicolas Gorlo*** **Luca Carlone**
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{harryz,ngorlo,lcarlone}@mit.edu

Abstract

Long-horizon robot operation requires spatio-temporal memory to record the environment state and recall it for downstream reasoning. Scene graphs and retrieval-augmented systems ground VLM descriptions to persistent 3D entities with rich semantic descriptions. However, VLM captions are noisy and viewpoint-inconsistent, and existing systems treat them as an oracle with no mechanism to detect unreliable stored descriptions. We introduce object-level semantic uncertainty for multi-view VLM memory: a score that measures object-centric cross-view semantic scatter of captions and identifies semantically unresolved objects. Then, we include our uncertainty scores in an advanced spatial-semantic memory system, that we dub UQ-DAAAM. UQ-DAAAM uses this score to actively refine uncertain objects under a fixed query budget by selecting high-quality views and fusing the resulting multi-view captions into a single object description. We also derive probabilistic guarantees showing that higher-quality candidate views (as selected by our approach) are more likely to reduce uncertainty. Our experiments show that uncertainty quantification can make embodied 4D memory systems more reliable and more effective. In particular, on the OC-NaVQA benchmark, UQ-DAAAM achieves substantially larger uncertainty reduction and better spatio-temporal question answering performance than baselines.

1 Introduction

Safe long-horizon robot operation benefits from a persistent semantic memory of objects, agents, and regions encountered over time, which can be queried later for downstream reasoning and decision-making. Recent work has shown that 3D scene graphs [13, 20, 15, 51, 24] can serve this role by grounding rich natural-language descriptions to persistent 3D entities observed by a robot. A parallel line of work [2, 52, 54] skips per-object 3D grounding and instead stores per-frame observations or captions in a retrieval database. Storing articulate captions has the advantage that the memory enabling the robot to reason about its spatio-temporal environment is directly interpretable by a human operator. In practice, semantic annotations of these systems come from vision-language models (VLMs), which produce open-vocabulary descriptions from camera images and possibly object masks. In the robot perception loop, VLMs convert raw visual observations into descriptions retrievable at query time that support downstream applications such as spatio-temporal question answering [13, 2] and task grounding [61, 33, 13].

However, VLM outputs are often noisy and semantically inconsistent across viewpoints [4, 7]. An object observed far away, from an occluded angle, or under difficult lighting conditions can result

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

in a caption that confidently states incorrect attributes of the object. Existing memory systems treat VLM annotations as an oracle [15, 2, 54], with no mechanism to detect if a stored description is unreliable. For instance, in a memory system like DAAAM [13], which assigns each object a small fixed set of representative views and stores their captions in its 4D scene graph representation: if those views happen to be poor, the resulting memory entry is silently incorrect.

We address this gap with **UQ-DAAAM**, an uncertainty-aware extension of DAAAM that gives each object in DAAAM's 4D scene graph an explicit semantic uncertainty score, and uses it for targeted refinement under a fixed budget of additional VLM queries. Our uncertainty metric measures *cross-view semantic scatter*, the log-volume spanned by the per-view caption embeddings of an object, which represents how much the VLM's descriptions disagree across views, following the volumetric semantic measurement framework of Lau et al. [27] and Zhang & Carlone [58]. This score identifies objects which descriptions are semantically unresolved and cannot yet be trusted for downstream reasoning. UQ-DAAAM then performs *active refinement*: it allocates the extra query budget preferentially to uncertain objects, selects additional views using DAAAM's view-quality heuristic [13], which scores each (frame, object) pair by visibility and pixel coverage, and fuses the resulting multi-view descriptions into a consolidated object description with reduced uncertainty.

Our analysis separates the uncertainty update from the view-selection heuristic. A standard Schur-complement identity yields the exact one-step uncertainty change and a deterministic criterion for reduction, independently of the view-quality score. Under an additional first-order stochastic dominance assumption, higher-quality views are at least as likely to reduce uncertainty. This comparative result supports the acquisition rule; the usefulness of the quality score remains an empirical property of the captioner and environment.

We evaluate UQ-DAAAM on the OC-NaVQA benchmark [13] for large-scale spatio-temporal question answering. UQ-DAAAM substantially outperforms other baselines on uncertainty reduction rate and downstream task performance, while keeping the underlying scene graph infrastructure and VLM unchanged. In summary, we contribute (i) an object-level uncertainty score over multi-view captions, (ii) a budgeted active-refinement and fusion strategy, (iii) an exact incremental uncertainty update and a comparative, FOSD-based probabilistic justification for view selection, and (iv) an integrated system, UQ-DAAAM, which achieves state-of-the-art results on OC-NaVQA.

2 Related Work

Spatio-temporal robot memory. 3D scene graphs [3, 43, 19, 20, 13] and retrieval-augmented memory systems [2, 52, 54] provide structured representations for long-horizon robot operation, with open-vocabulary semantics attached via VLM features [22, 51, 24] or per-object VLM captions [15]. DAAAM [13] combines both, by annotating a 4D scene-graph with localized VLM captions from a sparse set of representative view. Across all these systems, VLM outputs are treated as oracle annotations [15, 2, 44, 12, 18]; the few exceptions apply confidence only at the task-output level [41, 42], not per-node memory. We instead quantify and reduce semantic uncertainty at the object level, enabling the system to detect and resolve unreliable memory entries before they propagate to downstream reasoning.

Uncertainty quantification. Classical uncertainty quantification (UQ) methods such as MC Dropout [11], Deep Ensembles [26], and conformal prediction [1, 5] address epistemic or aleatoric uncertainty but do not directly apply to multimodal large language models (MLLMs), where token-level confidence is poorly aligned with semantic correctness [8, 27, 7]. Recent MLLM uncertainty work falls into three groups: input-level approaches that model uncertainty in multimodal embeddings [48–50]; output-level approaches that estimate uncertainty from response inconsistency via entropy, self-consistency, or embedding-space dispersion [38, 10, 35, 27, 7]; and methods based on conformal prediction or external verification [39, 46, 23]. VL-Uncertainty [59] estimates hallucination risk using semantic response variability under meaning-preserving visual and textual prompt perturbations. Our setting instead aggregates observations of the same object across physical viewpoints and uses the resulting uncertainty to allocate refinement queries. Orthogonally, hallucination mitigation methods improve training data or model architectures [30, 55, 47, 56], but residual uncertainty remains pervasive [4]. Our uncertainty is represented at the object level and captures semantic disagreement across views, following the theoretical framework of UMPIRE [27] and FUSE [58].

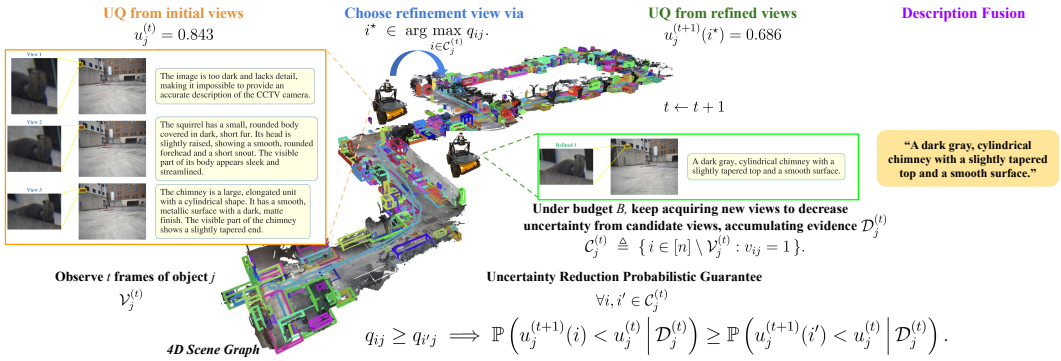


Figure 1: Overview of UQ-DAAAM. Starting from DAAAM's 4D scene graph, we first compute an object-level uncertainty score from the captions associated with the initial selected views. If an object is uncertain, additional refinement views are chosen from the feasible set using the view-quality heuristic q_{ij}, motivated by our comparative FOSD-based analysis. Refinement seeks to reduce uncertainty under a fixed budget. Captions from all views are fused into a single description.

3 Problem Formulation

We study *semantic memory construction* for embodied agents. Given an RGB-D stream $\mathcal{F} = \{f_1, \ldots, f_n\}$, an object-centric perception system [13] yields object observations (or *fragments*) $\mathcal{O} = \{o_1, \ldots, o_m\}$ tracked across frames. The system constructs a structured semantic memory grounded in a geometric reconstruction of the spatio-temporal environment by associating each o_j with natural-language descriptions from a selected view set $\mathcal{V}_j$; since descriptions can vary across views, the initial memory may leave some objects semantically ambiguous. We define an object-level uncertainty score $u_j = \mathcal{U}(o_j, \mathcal{V}_j)$ to quantify this ambiguity. Under a fixed additional query budget, *active refinement* then acquires further views for uncertain objects, allocating queries to the fragments and viewpoints most likely to resolve remaining ambiguity for downstream reasoning.

4 Method

Background. We build upon DAAAM [13], which constructs 4D scene graphs in real time from an RGB-D stream. DAAAM tracks object fragments $\{o_j\}_{j=1}^m$ across a rolling window of candidate frames $\{f_i\}_{i=1}^n$, maintaining a visibility matrix $v_{ij} \in \{0,1\}$ and a heuristic view-quality score $q_{ij} \in [0,1]$ per (frame, fragment) pair. In DAAAM [13], q_{ij} is large when object j is centered and large in frame i, and vice versa. A binary linear program (BLP) assigns each fragment to *a single* selected frame via a binary decision variable y_{ij}, maximizing $\sum_{i,j} q_{ij} y_{ij}$ subject to visibility and frame-budget constraints. The selected (frame, fragment) pairs are captioned by the localized captioning model, DAM [29]. A mapping backend builds a scene graph from the RGB-D sensor data and the annotated fragments. Each object forms a node in the scene graph. The captions then form a per-node description history.

From single-view to multi-view assignment. For object-level uncertainty quantification, we need a measure of view-dependent disagreement. A single caption per fragment provides no signal in that sense. We therefore relax the BLP so that each fragment is covered by multiple views rather than exactly one. Letting $y_{ij} \in \{0,1\}$ indicate whether frame f_i is one of the representative views for o_j, we replace the single-assignment constraint with $\sum_{i=1}^n y_{ij} = r \;\; \forall j \in [m]$, where r is the number of representative views per object (we use $r = 3$). Here, $x_i \in \{0,1\}$ indicates whether frame f_i is selected for annotation, whereas $y_{ij} \in \{0,1\}$ indicates whether object o_j is assigned to that frame. We retain $y_{ij} \leq v_{ij}$, which enforces visibility, and $y_{ij} \leq x_i$, which permits an assignment only when its frame is selected. The latter couples object assignments to frame selection; it does not itself impose a budget. We retain the quality-maximizing assignment objective $\sum_{i,j} q_{ij} y_{ij}$. For each object o_j, we run DAM on every view in $\mathcal{V}_j = \{f_i : y_{ij} = 1\}$, producing the multi-view caption set $\{c_{ij}\}_{i \in \mathcal{V}_j}$ used by the rest of the method.

4.1 Uncertainty Quantification

To quantify object-centric semantic uncertainty, we adapt the idea of volumetric semantic measurement [58, 27] to the multi-view robot memory construction setting of DAAAM. In our setting, each object is observed from multiple selected views, and each view yields a natural-language description produced by a VLM. Uncertainty should increase both when these view-conditioned descriptions are semantically inconsistent and when individual descriptions are internally less coherent. Our formulation therefore combines two signals: *cross-view semantic scatter* and *per-view incoherence*.

Multi-view semantic evidence for an object. Consider an object fragment o_j. Suppose DAAAM selects a set of representative views $\mathcal{V}_j = \{f_{i_1}, \dots, f_{i_r}\}$ for this object, where r is the number of selected views. For each selected view $f_i \in \mathcal{V}_j$, we query the VLM on the localized object crop and obtain a caption c_{ij}. We then embed each caption into a semantic vector space using a text encoder $\phi(\cdot)$ and normalize the result: $\boldsymbol{r}_{ij} \triangleq \frac{\phi(c_{ij})}{\|\phi(c_{ij})\|_2} \in \mathbb{R}^d$. Stacking the normalized caption embeddings column-wise gives the object-level *semantic evidence* matrix $\boldsymbol{R}_j \triangleq [\boldsymbol{r}_{1j}, \dots, \boldsymbol{r}_{rj}] \in \mathbb{R}^{d\times r}$. This matrix captures the semantic content of the object as perceived from the selected views.

Incoherence weighting. Following [58, 27], semantic diversity alone is insufficient for measuring uncertainty, since a set of captions could be similar but each incoherent. We therefore assign each caption c_{ij} an incoherence weight $\gamma_{ij} \triangleq \exp(\alpha(1-p_{ij}))$, where $p_{ij} \in [0,1]$ is the average per-token probability and $\alpha \geq 0$ is a scaling hyperparameter. A more coherent caption yields smaller γ_{ij}; a less coherent one inflates it. The weights are collected in $\boldsymbol{\Gamma}_j \triangleq \mathrm{diag}(\gamma_{1j}, \dots, \gamma_{rj}) \in \mathbb{R}^{r\times r}$.

The formulation also accommodates alternative per-view incoherence weights. In Table 2, we compare the default token-probability weighting with a combined weighting that incorporates predictive entropy and CLIP image–text grounding.

Uncertainty measure via semantic scatter. To combine cross-view semantics and per-view incoherence, we construct an incoherence-adjusted semantic evidence matrix $\boldsymbol{Z}_j \triangleq \boldsymbol{R}_j\boldsymbol{\Gamma}_j \in \mathbb{R}^{d\times r}$. Equivalently, the ℓ-th column of $\boldsymbol{Z}_j$ is the weighted embedding $\boldsymbol{z}_{ij} = \gamma_{ij}\boldsymbol{r}_{ij}$.

Definition 4.1 (Semantic scatter and uncertainty score)**.** *For an object fragment o_j, given incoherence-adjusted semantic evidence $\boldsymbol{Z}_j$, we define the semantic scatter matrix as $\boldsymbol{S}_j \triangleq \boldsymbol{Z}_j^\top\boldsymbol{Z}_j + \lambda\boldsymbol{I}_r = \boldsymbol{\Gamma}_j\boldsymbol{R}_j^\top\boldsymbol{R}_j\boldsymbol{\Gamma}_j + \lambda\boldsymbol{I}_r$, where $\lambda > 0$ is a small regularization constant. Our object-level uncertainty score is then defined as the log-determinant of the semantic scatter matrix*

$$u_j \triangleq \log\det \boldsymbol{S}_j. \tag{1}$$

The determinant of $\boldsymbol{Z}_j^\top\boldsymbol{Z}_j$ measures the squared volume spanned by $\boldsymbol{Z}_j$'s column vectors. $\det(\boldsymbol{S}_j)$, the semantic volume generated by the weighted multi-view caption embeddings, becomes large when (i) the captions from different views point in different semantic directions (increasing cross-view dispersion), or (ii) one or more captions are highly incoherent, inflating their weighted contribution through $\boldsymbol{\Gamma}_j$. Consequently, our uncertainty score u_j increases when the object's interpretation is unstable across views or unreliable within views.[2].

4.2 Active View Selection

Assume a calibrated threshold value τ: if u_j exceeds τ, then it implies that the agent is uncertain about the semantic interpretation of o_j. We let the agent actively acquire additional views for the same object fragment in an online manner: it acquires further views until the uncertainty falls below τ or a budget runs out. In the following discussion, we assume that the agent has already selected t views for object o_j and is considering which additional view to acquire next. We denote $u_j^{(t)}$ as the uncertainty score computed from the current t views. If $u_j^{(t)} > \tau$, then the agent is uncertain about o_j based on the current evidence, and we wish to assign additional views to lower $u_j^{(t+1)}$.

[2]We provide a more extensive rationale behind object-level uncertainty in Appendix A

BOED-inspired active view selection. Let θ_j denote the latent semantic state of object o_j, and let $\mathcal{V}_j^{(t)} \subseteq [n]$ denote the set of view indices selected for o_j after t acquisitions, so that $\boldsymbol{Z}_j^{(t)} \in \mathbb{R}^{d\times t}$ stacks the weighted embeddings $\{\boldsymbol{z}_{ij}\}_{i\in\mathcal{V}_j^{(t)}}$ and $\boldsymbol{S}_j^{(t)} \triangleq (\boldsymbol{Z}_j^{(t)})^\top \boldsymbol{Z}_j^{(t)} + \lambda \boldsymbol{I}_t$. We write $\mathcal{D}_j^{(t)}$ for the evidence accumulated after these t views, and define the feasible candidate views as the set of views where o_j is visible in the rolling window but have not yet been selected:

$$\mathcal{C}_j^{(t)} \triangleq \{\, i \in [n] \setminus \mathcal{V}_j^{(t)} : v_{ij} = 1 \,\}. \tag{2}$$

Inspired by Bayesian optimal experimental design (BOED) [6], we regard each candidate view as an experiment and its future caption as an outcome. Our utility is the expected reduction of the semantic scatter score. We do not specify a Bayesian generative model or compute a posterior over θ_j; consequently, this utility is not a Bayesian information gain.

Definition 4.2 (Exact one-step utility and acquisition rule)**.** *The one-step utility of candidate view i is defined as the expected uncertainty reduction $\Delta_j(i \mid \mathcal{D}_j^{(t)}) \triangleq \mathbb{E}\left[u_j^{(t)} - u_j^{(t+1)}(i) \,\middle|\, \mathcal{D}_j^{(t)}\right]$, where $u_j^{(t+1)}(i)$ is the uncertainty score after acquiring view i. The next refinement view should therefore be chosen by*

$$i_j^\star \in \arg\max_{i\in\mathcal{C}_j^{(t)}} \Delta_j(i \mid \mathcal{D}_j^{(t)}). \tag{3}$$

One-step update via augmented scatter matrix. Given $\boldsymbol{r}_{ij}$ as the normalized VLM output embedding when presented with view i for object j, $\gamma_{ij} > 0$ as its incoherence weight, and weighted candidate embedding $\boldsymbol{z}_{ij} \triangleq \gamma_{ij}\, \boldsymbol{r}_{ij} \in \mathbb{R}^d$, we define the similarity vector $\boldsymbol{b}_{ij}^{(t)} \triangleq (\boldsymbol{Z}_j^{(t)})^\top \boldsymbol{z}_{ij} \in \mathbb{R}^t$. Denoting the semantic scatter matrix obtained via t currently collected views of object j as $\boldsymbol{S}_j^{(t)}$, we can apply matrix algebra to write out the augmented scatter matrix after adding view i as

$$\boldsymbol{S}_j^{(t+1)}(i) = \begin{bmatrix} \boldsymbol{S}_j^{(t)} & \boldsymbol{b}_{ij}^{(t)} \\ (\boldsymbol{b}_{ij}^{(t)})^\top & \gamma_{ij}^2 + \lambda \end{bmatrix}. \tag{4}$$

With this formulation, we can derive a closed-form expression for the one-step uncertainty update. The underlying Schur-complement identity is standard. Our contribution is its use for incremental object-level uncertainty updates within the memory-refinement loop, reusing previously computed caption embeddings. For sake of clarity, we first define two intermediate random variables.

Definition 4.3 (Redundancy score and Schur score)**.** *For any candidate view $i \in \mathcal{C}_j^{(t)}$ acquired for object j with t existing views, we define the redundancy score $\eta_{ij}^{(t)}$ and Schur score $s_{ij}^{(t)}$ as*

$$\eta_{ij}^{(t)} \triangleq (\boldsymbol{b}_{ij}^{(t)})^\top (\boldsymbol{S}_j^{(t)})^{-1} \boldsymbol{b}_{ij}^{(t)}, \quad s_{ij}^{(t)} \triangleq \gamma_{ij}^2 + \lambda - \eta_{ij}^{(t)}. \tag{5}$$

The redundancy score $\eta_{ij}^{(t)}$ measures how much newly acquired information is already known and is large when the new weighted caption embedding is well explained by the currently selected views. The Schur score $s_{ij}^{(t)}$ is the Schur complement of $\boldsymbol{S}_j^{(t)}$ in $\boldsymbol{S}_j^{(t+1)}(i)$ and captures the amount of new information that view i would add after accounting for redundancy.

Theorem 4.4 (Closed-form one-step uncertainty update)**.** *For any candidate view $i \in \mathcal{C}_j^{(t)}$ acquired for object j with t existing views, the one-step uncertainty update is given by*

$$u_j^{(t+1)}(i) - u_j^{(t)} = \log\left(s_{ij}^{(t)}\right). \tag{6}$$

Theorem 4.5 (Criterion for one-step uncertainty reduction)**.** *For any candidate view $i \in \mathcal{C}_j^{(t)}$ acquired for object j with t existing views, we have the following uncertainty reduction criterion:*

$$u_j^{(t+1)}(i) < u_j^{(t)} \quad \Longleftrightarrow \quad \eta_{ij}^{(t)} > \gamma_{ij}^2 + \lambda - 1.$$

Intuitively, to reduce uncertainty, the redundancy (agreement with existing evidence) η_{ij} of a new view needs to overcome both regularization and its own incoherence.

BOED-inspired proxy selection using the view-quality score. So far, our analysis suggests that for a candidate view $i \in \mathcal{C}_j^{(t)}$, the exact one-step uncertainty reduction is $u_j^{(t)} - u_j^{(t+1)}(i) = -\log s_{ij}^{(t)}$, and the expected uncertainty-reduction utility is thus $\Delta_j(i) \triangleq \mathbb{E}\left[-\log s_{ij}^{(t)} \,\middle|\, \mathcal{D}_j^{(t)}\right]$. Evaluating these expectations requires a model of future caption outcomes or additional sampling for each candidate.

We avoid this expense by using the existing view-quality score q_{ij} as a proxy. For a fixed object o_j, we expect higher-quality views to result in smaller Schur scores, and thus larger one-step uncertainty reductions. Towards this end, we make the following assumption on the relationship between q_{ij} and $s_{ij}^{(t)}$.

Assumption 4.6 (First-order stochastic dominance)**.** *For each object o_j, the conditional distribution of the random Schur score $s_{ij}^{(t)}$ is stochastically nonincreasing in q_{ij}. That is, for any two candidates $i, i' \in \mathcal{C}_j^{(t)}$,*

$$q_{ij} \geq q_{i'j} \quad \Longrightarrow \quad \mathbb{P}\left(s_{ij}^{(t)} \leq x \,\middle|\, \mathcal{D}_j^{(t)}\right) \geq \mathbb{P}\left(s_{i'j}^{(t)} \leq x \,\middle|\, \mathcal{D}_j^{(t)}\right) \qquad \forall x > 0.$$

Equivalently, conditional on the current evidence, a higher-quality view produces a Schur score that is smaller in the sense of first-order stochastic dominance (FOSD).

This assumption is natural because the Schur score $s_{ij}^{(t)}$ decreases when the future caption is less incoherent (smaller γ_{ij}) and more explained by the current evidence (larger $\eta_{ij}^{(t)}$). Thus, a higher-quality view is assumed to shift the distribution of $s_{ij}^{(t)}$ toward smaller values.

Proposition 4.7 (Monotonicity of expected reduction under FOSD)**.** *Suppose the FOSD assumption holds for object o_j. Then for any two candidates $i, i' \in \mathcal{C}_j^{(t)}$, $q_{ij} \geq q_{i'j} \quad \Longrightarrow \quad \Delta_j(i) \geq \Delta_j(i')$. Consequently, any maximizer of q_{ij} over $\mathcal{C}_j^{(t)}$ is also a maximizer of the expected uncertainty-reduction utility $\Delta_j(i)$ and candidate views can be ranked by q_{ij}, yielding the proxy acquisition rule:*

$$i_j^\star \in \arg\max_{i \in \mathcal{C}_j^{(t)}} q_{ij}. \tag{7}$$

Theorem 4.8 (Comparative FOSD-based probabilistic guarantee)**.** *Suppose that for two feasible candidate views $i, i' \in \mathcal{C}_j^{(t)}$, $q_{ij} \geq q_{i'j}$ and the corresponding Schur scores $s_{ij}^{(t)}, s_{i'j}^{(t)}$ satisfies Assumption 4.6. Then*

$$\mathbb{P}\left(u_j^{(t+1)}(i) < u_j^{(t)} \,\middle|\, \mathcal{D}_j^{(t)}\right) \geq \mathbb{P}\left(u_j^{(t+1)}(i') < u_j^{(t)} \,\middle|\, \mathcal{D}_j^{(t)}\right).$$

This is a comparative guarantee conditional on Assumption 4.6: it orders the probabilities of uncertainty reduction, but does not ensure that either candidate reduces uncertainty. It also does not guarantee caption correctness or downstream task performance. The view-selection ablations provide empirical support for the quality heuristic, but do not establish FOSD for every object and candidate pair. In contrast, the exact update and reduction criterion in Theorems 4.4 and 4.5 require no assumption about q_{ij}.

Proposition 4.9 (Absolute reduction bound from Schur-score moments)**.** *Let $\mu_{ij} = \mathbb{E}[s_{ij}^{(t)} \mid \mathcal{D}_j^{(t)}]$ and $\sigma_{ij}^2 = \mathrm{Var}(s_{ij}^{(t)} \mid \mathcal{D}_j^{(t)}) < \infty$. If $\mu_{ij} < 1$, then*

$$\mathbb{P}\Big(u_j^{(t+1)}(i) < u_j^{(t)} \,\Big|\, \mathcal{D}_j^{(t)}\Big) \geq \frac{(1-\mu_{ij})^2}{\sigma_{ij}^2 + (1-\mu_{ij})^2}.$$

This follows from Cantelli's inequality and Theorem 4.4. Unlike the comparative guarantee, this bound does not require FOSD or a view-quality surrogate. It concerns a decrease in the uncertainty score, rather than crossing the resolution threshold or correcting a factual error.

Roles of uncertainty and view quality. The uncertainty score u_j determines which objects receive refinement and when refinement stops. The quality score q_{ij} ranks candidate views for an object already selected for refinement. These signals serve different purposes: clear views can yield contradictory captions, while low-quality views can yield mutually consistent captions.

Method	Descriptive Question Accuracy ↑				Position Error [m] ↓				Temporal Error [min] ↓			
	Short	Medium	Long	All	Short	Medium	Long	All	Short	Medium	Long	All
DAAAM	0.679	0.753	0.690	0.711	22.07	41.69	53.47	41.75	3.669	0.914	1.438	1.792
DAAAM-RR	0.683	0.751	0.694	0.716	21.52	41.19	53.34	41.78	3.748	0.918	1.428	1.769
DAAAM-QR	0.715	0.760	0.748	0.738	17.83	29.47	57.26	40.14	2.284	0.905	1.408	1.734
DAAAM-NC [23]	0.681	0.755	0.692	0.713	21.88	41.53	53.39	41.62	3.648	0.912	1.434	1.783
DAAAM-LN [34]	0.682	0.756	0.694	0.714	21.74	41.45	**53.31**	41.55	3.623	0.910	1.430	1.776
DAAAM-Ent [25]	0.703	0.763	0.734	0.733	20.13	36.82	55.29	40.33	2.891	0.907	1.368	1.718
DAAAM-ES [7]	0.714	0.769	0.748	0.743	19.28	34.61	55.87	39.69	2.703	**0.903**	1.341	1.683
UQ-DAAAM (Ours)	**0.731**	**0.778**	**0.786**	**0.761**	**15.46**	**25.35**	58.94	**37.84**	**1.948**	0.910	**1.275**	**1.589**

Table 1: Results on OC-NaVQA benchmark [13, 2]. All models use GPT-5-mini.

Overall active view refinement procedure. Backed by our theoretical analysis, for each object o_j with $u_j > \tau$, we allocate a per-object refinement budget B. We collect feasible additional views $\mathcal{C}_j^{(t)}$ for o_j via Equation (2). Next, we rank the candidates in $\mathcal{C}_j^{(t)}$ by the heuristic view-quality score q_{ij} and select the top-B views. The VLM is then queried on these additional views sequentially, and each resulting caption is incorporated into the object's multi-view semantic representation. After each update, the object-level uncertainty is recomputed. Refinement stops when the uncertainty falls below τ, the budget B is exhausted, or no feasible candidate views remain. Note that our refinement procedure can be understood as an "overt attention" process to reduce uncertainty.

4.3 Multi-view Descriptions Aggregation

After uncertainty-driven view selection, each object is associated with a small set of captions obtained from multiple views. We fuse these captions into a single object description using a constrained LLM-based aggregation step. Captions are ordered by acquisition stage (initial views first, refinement views last) and the LLM is instructed to privilege later (refinement) captions when resolving conflicts. When disagreement exists only among initial views, the fused caption preserves the alternatives using cautious language (e.g., *"white or gray"*) rather than collapsing ambiguity prematurely. The LLM is further constrained to use only information present in at least one source caption, preventing hallucinated facts. The full prompt specification and output format are given in Appendix H; the overall UQ-DAAAM procedure is summarized in Appendix I.

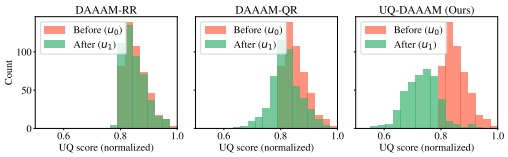
5 Experiments

In our implementation, we set $r = 3$ to gather diverse perspectives on each object to measure the semantic scatter as the uncertainty. We use a budget of $B = 2$ for refinement, and the uncertainty threshold τ is set to the top 20-th percentile of the object uncertainty distribution, selected on a held-out calibration dataset, from which we also derive the statistics to normalize the uncertainty scores. In the calculation of u, we use a small regularization constant of $\lambda = 1e - 8$ to ensure numerical stability of the log-determinant. We set the temperature for caption generation to 1. For caption fusion, we use GPT-5-mini. Visualizations and analysis of qualitative results are in Appendix Q.

5.1 Budgeted Refinement on OC-NaVQA

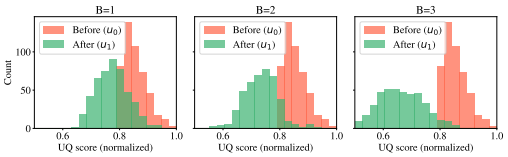
We evaluate UQ-DAAAM on spatio-temporal question answering (SQA) using the CODa dataset and the object-centric NaVQA (OC-NaVQA) benchmark introduced by DAAAM [13]. DAAAM motivates OC-NaVQA as a large-scale, long-horizon extension of NaVQA [2]: it re-annotates spatial queries with actual object positions, improves label quality with a custom 3D labeling tool, and evaluates over the full sequence context rather than a restricted observation window, resulting in settings of up to 35.8 minutes and 1.64 km of traveled distance.

Following DAAAM [13], we report three metrics on their corresponding question categories. *Question accuracy* is the fraction of descriptive (binary yes/no) questions answered correctly. *Positional error* measures the distance, in meters, between the predicted and annotated object locations for spatial questions. *Temporal error* measures the absolute difference, in minutes, between predicted and annotated times or durations for temporal questions. Higher accuracy and lower positional and temporal errors indicate better performance. All methods use the same reasoning backend and eval-



	DAAAM-RR	DAAAM-QR	UQ-DAAAM
URR	5.6%	34.2%	92.3%
Mean Δ_u	-0.006	-0.036	-0.121

Figure 2: Uncertainty reduction comparison across refinement strategies. *Top:* Uncertainty reduction distributions. *Bottom:* URR and mean uncertainty reduction at fixed budget $B = 2$.



	$B = 1$	$B = 2$	$B = 3$
URR	68.9%	92.3%	94.4%
Mean Δ_u	-0.079	-0.121	-0.169

Figure 3: Uncertainty reduction comparison across budgets. *Top:* per-budget uncertainty reduction curves. *Bottom:* URR and mean uncertainty reduction.

uation protocol, isolating differences in memory construction and refinement. We report the main comparison at a fixed refinement budget in Table 1. Specifically, we compare:

- **DAAAM**: the original pipeline with no post hoc refinement;
- **DAAAM-RR**: after the initial DAAAM pass, additional feasible object-view queries are allocated uniformly at random on objects under the same extra budget;
- **DAAAM-QR**: after the initial DAAAM pass, the extra budget is allocated only for low-quality objects via DAAAM's original view-quality score q_{ij}, without uncertainty;
- **UQ-DAAAM**: our full method, which uses object-level uncertainty to prioritize refinement and then updates memory with the additional multi-view semantic evidence. In addition, we compare against 4 different UQ methods implemented on top of DAAAM:
- **DAAAM-QR + Best**: uses quality-based object and view selection, as in DAAAM-QR, but retains the single highest-quality caption instead of fusing the available captions. This baseline examines the performance of quality-only refinement without multi-view caption fusion.
- **DAAAM-NC** [23], which uses neighborhood consistency;
- **DAAAM-LN** [34], which uses next-token log likelihood as an uncertainty signal;
- **DAAAM-Ent** [25], which uses entropy as an uncertainty signal; and
- **DAAAM-ES** [7], which uses evidence score as an uncertainty signal. The details for the baselines can be found in Appendix J.

As the results show in Table 1, UQ-DAAAM outperforms the baselines across all three metrics, with the largest gains in question accuracy and positional error. This demonstrates that uncertainty-aware refinement can more effectively allocate a fixed semantic-query budget to improve the quality.

5.2 Uncertainty Reduction

In addition to downstream task performance, we directly evaluate whether refinement achieves its immediate objective: reducing object-level semantic uncertainty. For each object o_j, we compute its uncertainty before refinement, u_j^{pre}, and after refinement, u_j^{post}. We focus in particular on the subset of objects that are initially marked as uncertain, i.e., $\mathcal{U}_\tau^{\text{pre}} = \{\, j \in [m] : u_j^{\text{pre}} > \tau \,\}$. We then measure how effectively each method reduces uncertainty within this set.

Our primary metric is the *uncertainty reduction rate*, defined as the percentage of initially uncertain objects whose uncertainty falls below the threshold after refinement: $\text{URR} = \frac{\left|\{\, j \in \mathcal{U}_\tau^{\text{pre}} : u_j^{\text{post}} \le \tau \,\}\right|}{|\mathcal{U}_\tau^{\text{pre}}|}$ This metric directly captures how often refinement succeeds in converting an uncertain object into a semantically resolved one. We compare the same DAAAM-RR, DAAAM-QR, and UQ-DAAAM variants from Section 5.1; uncertainty-as-error-predictor results are deferred to Appendix J.

The uncertainty-reduction results in Figure 2 show a clear advantage for UQ-DAAAM over both non-uncertainty-based refinement baselines. At a fixed budget $B = 2$, UQ-DAAAM achieves a URR of 92.3%, compared to 34.2% for quality-based refinement and only 5.6% for random refinement. UQ-DAAAM shifts the post-refinement distribution markedly to the left, showing a much larger and more consistent reduction in uncertainty across objects. Overall, these results indicate that the proposed uncertainty-aware refinement is substantially more effective.

	Budget			Object				Weight		View			Fusion					Threshold			
	B=1	B=2	B=3	LQ	CD	UW	Full	Single	Comb	R	q_{ij}	Dir	Best	IO	RO	UC	C	10	20	30	40
Accuracy	0.742	0.761	**0.774**	0.748	0.737	0.752	**0.761**	0.761	0.763	0.741	**0.761**	0.754	0.754	0.728	0.739	0.747	**0.761**	0.748	0.761	0.767	**0.770**
Position [m]	38.46	37.84	**36.99**	38.63	39.25	38.47	**37.84**	37.84	37.79	39.18	**37.84**	38.39	38.41	40.12	39.47	38.89	**37.84**	39.31	37.84	37.21	**37.09**
Temporal [min]	1.712	1.589	**1.539**	1.651	1.703	1.638	**1.589**	1.589	1.586	1.681	**1.589**	1.624	1.631	1.762	1.714	1.668	**1.589**	1.674	1.589	1.558	**1.551**

Table 2: Ablation study across six design axes. Each group varies one design axis while holding others fixed.

5.3 Ablations

We perform a comprehensive ablation study to separate the effect of uncertainty estimation, object selection, view refinement, and caption fusion.

Performance versus budget. We vary the additional refinement budget B and plot uncertainty reduction performance as a function of the number of extra caption queries. This reveals the refinement-efficiency regime in which UQ-DAAAM provides the largest gains. Results shown in Figure 3 suggest that increasing the refinement budget consistently improves uncertainty reduction, but with diminishing returns. As the per-object budget increases from $B = 1$ to $B = 3$, the post-refinement uncertainty distribution shifts progressively leftward, indicating that additional queried views provide increasingly stronger semantic disambiguation. Notably, most of the gain is already achieved by $B = 2$, suggesting that a small number of extra views is sufficient to resolve uncertainty for the majority of objects, while larger budgets provide smaller but still measurable additional benefit. The downstream OC-NaVQA results are consistent with the uncertainty-reduction analysis. As shown in "Budget" of Table 2, larger refinement budgets lead to better task performance, and all budgeted UQ-DAAAM variants outperform vanilla DAAAM. As the per-object refinement budget increases, all three metrics improve, with the largest gain from $B = 1$ to $B = 2$, consistent with the uncertainty-reduction results. Overall, these results show that reducing object-level uncertainty translates into measurable downstream gains in spatio-temporal question answering.

Refinement object selection rule. We ablate the rule used to choose which objects receive additional refinements (Table 2, "Object"): lowest-quality (**LQ**), highest cross-view disagreement via caption embedding dispersion (**CD**), uncertainty without incoherence weighting (**UW**), and full uncertainty (**Full**). Full leads across all metrics; UW comes closest, confirming that incoherence weighting adds a consistent but modest gain over dispersion alone, while CD trails, indicating that raw caption disagreement without visual grounding is a poor proxy for semantic unreliability.

Incoherence weighting. We compare the default token-probability weighting (**Single**) with **DAAAM-Comb** (**Comb**), which combines token probability, predictive entropy, and CLIP image–text grounding. The comparison in the "Weight" group of Table 2 gives similar downstream scores for the two variants, with small numerical improvements in positional and temporal error for Comb. These results suggest limited additional benefit from the richer weighting in this setting; the reported means alone do not establish statistical equivalence.

Refinement view selection heuristic. Conditioned on the same object selected for refinement, we ablate the rule used to choose the next additional view. We compare three policies: selecting a random visible view (**R**), selecting the visible view with the highest heuristic quality score q_{ij}, and a direction-based heuristic [32] that prioritizes views according to the object's viewing direction (**Dir**). This comparison isolates whether refinement gains arise simply from better geometric visibility, from DAAAM's original quality heuristic, or from choosing views that are expected to be semantically more informative. As shown under the "View" tab in Table 2, random view selection (R) noticeably degrades performance. Dir closely tracks q_{ij}, suggesting that geometric visibility already recovers most of the gain, but the quality-weighted score gives additional filtering.

Fusion strategy. We ablate how multi-view captions are consolidated into a final object description in order to separate the effect of semantic fusion from the effects of uncertainty estimation and refinement. We compare five variants: no fusion, which uses the single best caption only (**Best**); initial-only fusion, which aggregates only captions from the initial selected views (**IO**); refined-only

fusion, which aggregates only captions acquired during refinement (**RO**); unconstrained LLM fusion, which merges all captions without explicit conflict-handling rules (**UC**); and our constrained stage-aware fusion, which prefers refined captions when they conflict with initial ones while preserving unresolved alternatives when disagreement remains within the initial set (**C**). This comparison tests whether the gains of UQ-DAAAM come merely from acquiring more views, or from consolidating multi-view semantic evidence in a way that is conservative, conflict-aware, and aligned with the refinement process. As shown under the "Fusion" tab in Table 2, C achieves the best performance. The monotone improvement from IO $\rightarrow$ RO $\rightarrow$ UC $\rightarrow$ Best $\rightarrow$ C shows that each level of conflict-awareness and stage-sensitivity provides additional gain, and that the benefit of refinement is substantially eroded when captions are fused naively. The quality-only DAAAM-QR + Best baseline achieves an accuracy of 0.736, a positional error of 39.68 m, and a temporal error of 1.781 min on the All split. By comparison, the Best variant of UQ-DAAAM in Table 2 achieves 0.754, 38.41 m, and 1.631 min, respectively. Both retain a single highest-quality caption, so, under matched query budgets and evaluation settings, this comparison supports the benefit of uncertainty-guided object selection independently of caption fusion.

Refinement threshold. We ablate the effect of the uncertainty threshold used to trigger refinement, which determines how aggressively the system allocates additional semantic-query budget. Specifically, we use the 10th, 20th, 30th, and 40th quantiles of the object-level uncertainty distribution as refinement thresholds, so that progressively lower thresholds cause a larger fraction of objects to be refined. This experiment exposes the tradeoff between semantic-query cost and downstream gain: stricter thresholds focus refinement on only the most uncertain objects, while more permissive thresholds increase coverage at the cost of additional queries. As shown under the "Threshold" tab in Table 2, performance improves substantially from the 10th to the 20th quantile, and continues to improve at the 30th quantile. The 40th quantile yields only marginal further gain, indicating diminishing returns beyond the 30th percentile and confirming that the default 20 offers a favorable cost-performance tradeoff.

5.4 Memory and Runtime Details

Across 7 CODa sequences (98,866 frames), UQ-DAAAM runs at 13.5 Hz on a single RTX 4090, comfortably exceeding the 10 Hz sensor rate. The cross-view BLP solver adds a mean latency of only 1.49% of total frame processing time. Each object receives 3.44 ± 0.68 cross-view captions on average; uncertainty is computed via `numpy.linalg.slogdet` on CPU with negligible overhead. Of the 18,298 objects annotated across all sequences, 6,324 (34.6%) were selected for refinement, receiving 1.32 ± 0.47 additional views each. Active view selection reduced the mean UQ score from 0.56 to 0.32, indicating substantially increased caption agreement. GPU memory remains within the 24 GB envelope of the RTX 4090, as the UQ and BLP components are CPU-only and the DAM-3B is shared between initial annotation and refinement.

6 Conclusion

We presented UQ-DAAAM, an uncertainty-aware extension of DAAAM for 4D spatial-temporal memory. Our key idea is to endow object-centric memory with an explicit notion of semantic uncertainty, so that the system can recognize when multi-view descriptions remain ambiguous and selectively acquire additional evidence. To this end, we introduced an object-level uncertainty measure that combines cross-view semantic disagreement with per-view incoherence, a refinement strategy that allocates a query budget to uncertain objects and informative views, and a fusion mechanism that consolidates refined multi-view descriptions into stable object summaries. Experiments show that UQ-DAAAM improves uncertainty reduction and spatio-temporal reasoning over baselines. We show qualitative examples in Appendix Q and list the limitations of the work in Appendix R.

Acknowledgements

This work was partially funded by ONR RAPID Program and by DCIST.

References

[1] A.N. Angelopoulos and S. Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[2] A. Anwar, J. Welsh, J. Biswas, S. Pouya, and Y. Chang. ReMEmbR: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2025.

[3] I. Armeni, Z. He, J. Gwak, A. Zamir, M. Fischer, J. Malik, and S. Savarese. 3D scene graph: A structure for unified semantics, 3D space, and camera. In *Intl. Conf. on Computer Vision (ICCV)*, pp. 5664–5673, 2019.

[4] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M.Z. Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[5] R.F. Barber, E.J. Candes, A. Ramdas, and R.J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

[6] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical science*, pp. 273–304, 1995.

[7] C. Chen, K. Liu, Z. Chen, Y. Gu, Y. Wu, M. Tao, Z. Fu, and J. Ye. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*, 2024.

[8] S-H. Chou, S. Chandhok, J. Little, and L. Sigal. Mm-r3: On (in-) consistency of vision-language models (vlms). In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4762–4788, 2025.

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[11] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Intl. Conf. on Machine Learning (ICML)*, pp. 1050–1059. PMLR, 2016.

[12] M.F. Ginting, D-K. Kim, X. Meng, A.M. Reinke, B.J. Krishna, N. Kayhani, O. Peltzer, D. Fan, A. Shaban, S-K. Kim, M. Kochenderfer, A. Agha-mohammadi, and S. Omidshafiei. Enter the mind palace: Reasoning and planning for long-term active embodied question answering. In *Conference on Robot Learning (CoRL)*, 2025.

[13] N. Gorlo, L. Schmid, and L. Carlone. Describe anything anywhere at any moment. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[14] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

[15] Q. Gu, A. Kuwajerwala, S. Morin, K.M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, C. Gan, C.M. de Melo, J.B. Tenenbaum, A. Torralba, F. Shkurti, and L. Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, May 2024.

[16] C. Guo, G. Pleiss, Y. Sun, and K.Q. Weinberger. On calibration of modern neural networks. In *Intl. Conf. on Machine Learning (ICML)*, pp. 1321–1330, 2017.

[17] P. He, X. Liu, J. Gao, and W. Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[18] D. Honerkamp, M. Büchner, F. Despinoy, T. Welschehold, and A. Valada. Language-grounded dynamic scene graphs for interactive object search with mobile manipulation. *IEEE Robotics and Automation Letters (RA-L)*, 2024.

[19] N. Hughes, Y. Chang, and L. Carlone. Hydra: a real-time spatial perception engine for 3D scene graph construction and optimization. In *Robotics: Science and Systems (RSS)*, 2022.

[20] N. Hughes, Y. Chang, S. Hu, R. Talak, R. Abdulhai, J. Strader, and L. Carlone. Foundations of spatial perception for robotics: Hierarchical representations and real-time systems. *Intl. J. of Robotics Research*, 2024.

[21] E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[22] K.M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, S. Li, G. Iyer, S. Saryazdi, N. Keetha, A. Tewari, J.B. Tenenbaum, C.M. de Melo, M. Krishna, L. Paull, F. Shkurti, and A. Torralba. Conceptfusion: Open-set multimodal 3d mapping. In *Robotics: Science and Systems (RSS)*, 2023.

[23] Z. Khan and Y. Fu. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10854–10863, 2024.

[24] S. Koch, N. Vaskevicius, M. Colosi, P. Hermosilla, and T. Ropinski. Open3DSG: Open-vocabulary 3D scene graphs from point clouds with queryable objects and open-set relationships. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[25] L. Kuhn, Y. Gal, and S. Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

[26] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[27] G.K.R. Lau, H. Dao, N.K.H. Lin, and B.K.H. Low. Uncertainty quantification for mllms. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025.

[28] L. Li, J. Lei, Z. Gan, and J. Liu. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2042–2051, 2021.

[29] L. Lian, Y. Ding, Y. Ge, S. Liu, H. Mao, B. Li, M. Pavone, M-Y. Liu, T. Darrell, A. Yala, and Y. Cui. Describe anything: Detailed localized image and video captioning. In *Intl. Conf. on Computer Vision (ICCV)*, 2025.

[30] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang. Aligning large multi-modal model with robust instruction tuning. *CoRR*, 2023.

[31] H. Liu, C. Li, Q. Wu, and Y.J. Lee. Visual instruction tuning. In *Conf. on Neural Information Processing Systems (NeurIPS)*, 2023.

[32] X. Liu, A. Prabhu, F. Cladera, I.D. Miller, L. Zhou, C.J. Taylor, and V. Kumar. Active metric-semantic mapping by multiple aerial robots. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3282–3288. IEEE, 2023.

[33] D. Maggio, Y. Chang, N. Hughes, M. Trang, D. Griffith, C. Dougherty, E. Cristofalo, L. Schmid, and L. Carlone. Clio: Real-time task-driven open-set 3D scene graphs. *IEEE Robotics and Automation Letters (RA-L)*, 9(10):8921–8928, 2024.

[34] A. Malinin and M. Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

[35] P. Manakul, A. Liusie, and M.J.F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

[36] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

[37] J. Ni, G.H. Abrego, N. Constant, J. Ma, K. Hall, D. Cer, and Y. Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the association for computational linguistics: ACL 2022*, pp. 1864–1874, 2022.

[38] A. Nikitin, J. Kossen, Y. Gal, and P. Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37: 8901–8929, 2024.

[39] V. Quach, A. Fisch, T. Schuster, A. Yala, J.H. Sohn, T.S. Jaakkola, and R. Barzilay. Conformal language modeling. *arXiv preprint arXiv:2306.10193*, 2023.

[40] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Intl. Conf. on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

[41] A.Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley, Z. Xu, D. Sadigh, A. Zeng, and A. Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Conference on Robot Learning (CoRL)*, 2023.

[42] A.Z. Ren, J. Clark, A. Dixit, M. Itkina, A. Majumdar, and D. Sadigh. Explore until confident: Efficient exploration for embodied question answering. In *Robotics: Science and Systems (RSS)*, 2024.

[43] A. Rosinol, A. Gupta, M. Abate, J. Shi, and L. Carlone. 3D dynamic scene graphs: Actionable spatial perception with places, objects, and humans. In *Robotics: Science and Systems (RSS)*, 2020. doi: 10.15607/RSS.2020.XVI.079. URL `http://news.mit.edu/2020/robots-spatial-perception-0715`.

[44] S. Saxena, B. Buchanan, C. Paxton, P. Liu, B. Chen, N. Vaskevicius, L. Palmieri, J. Francis, and O. Kroemer. Grapheqa: Using 3d semantic scene graphs for real-time embodied question answering. In *Conference on Robot Learning (CoRL)*, 2025.

[45] L. Schmid, M. Abate, Y. Chang, and L. Carlone. Khronos: A unified approach for spatio-temporal metric-semantic SLAM in dynamic environments. In *Robotics: Science and Systems (RSS)*, 2024.

[46] J. Su, J. Luo, H. Wang, and L. Cheng. Api is enough: Conformal prediction for large language models without logit-access. *arXiv preprint arXiv:2403.01216*, 2024.

[47] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024.

[48] D. Tran, J. Liu, M.W. Dusenberry, D. Phan, M. Collier, J. Ren, K. Han, Z. Wang, Z. Mariet, H. Hu, et al. Plex: Towards reliability using pretrained large model extensions. *arXiv preprint arXiv:2207.07411*, 2022.

[49] U. Upadhyay, S. Karthik, M. Mancini, and Z. Akata. Probvlm: Probabilistic adapter for frozen vison-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1899–1910, 2023.

[50] A. Venkataramanan, P. Bodesheim, and J. Denzler. Probabilistic embeddings for frozen vision-language models: Uncertainty quantification with gaussian process latent variable models. *arXiv preprint arXiv:2505.05163*, 2025.

[51] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. *Robotics: Science and Systems (RSS)*, 2024.

[52] Q. Xie, S.Y. Min, P. Ji, Y. Yang, T. Zhang, A. Bajaj, R. Salakhutdinov, M. Johnson-Roberson, and Y. Bisk. Embodied-RAG: General non-parametric embodied memory for retrieval and generation, 2024. URL `https://arxiv.org/abs/2409.18313`.

[53] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 2025.

[54] Y. Yang, H. Yang, J. Zhou, P. Chen, H. Zhang, Y. Du, and C. Gan. 3d-mem: 3d scene memory for embodied exploration and reasoning. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 17294–17303, 2025.

[55] Q. Yu, J. Li, L. Wei, L. Pang, W. Ye, B. Qin, S. Tang, Q. Tian, and Y. Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12944–12953, 2024.

[56] B. Zhai, S. Yang, C. Xu, S. Shen, K. Keutzer, and M. Li. Halle-switch: Controlling object hallucination in large vision language models. *arXiv preprint arXiv:2310.01779*, 2023.

[57] A. Zhang, C. Eranki, C. Zhang, J-H. Park, R. Hong, P. Kalyani, L. Kalyanaraman, A. Gamare, A. Bagad, M. Esteva, et al. Toward robust robot 3-d perception in urban environments: The ut campus object dataset. *IEEE Trans. Robotics*, 40:3322–3340, 2024.

[58] H. Zhang and L. Carlone. Fuse: Quantifying uncertainty in multimodal llms by bayesian fusing epistemic and aleatoric uncertainty. *Work in progress*, 2026.

[59] R. Zhang, H. Zhang, and Z. Zheng. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024.

[60] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conf. on Computer Vision (ECCV)*, pp. 1–21. Springer, 2022.

[61] Z. Zhang, Z. Zhu, P. Li, T. Liu, X. Ma, Y. Chen, B. Jia, S. Huang, and Q. Li. Task-oriented sequential grounding in 3D scenes, 2024. URL `https://arxiv.org/abs/2408.04034`.

[62] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang. Fast segment anything, 2023.

Supplementary Material

A Object-level Uncertainty Matters in DAAAM

DAAAM is designed to build grounded semantic memory from a sparse set of representative views. However, not all objects are equally easy to describe from a limited number of views. Some objects remain semantically stable after only a few observations, while others exhibit persistent ambiguity due to clutter, viewpoint effects, or fine-grained attributes. The uncertainty score u_j provides a principled way to detect these difficult cases. Rather than treating all objects uniformly, the system can identify which objects are already semantically well characterized and which objects require further refinement. This makes uncertainty quantification a key interface between memory construction and active view acquisition: it summarizes the quality of the current semantic evidence for each object and supports downstream decisions about where additional sensing or captioning effort should be spent.

B Augmented Block Form after Adding a New View

Recall

$$\boldsymbol{Z}_j^{(t)} = \big[\boldsymbol{z}_{j1}, \ldots, \boldsymbol{z}_{jt}\big] \in \mathbb{R}^{d\times t}$$

denote the current weighted embedding matrix, where each column is a weighted caption embedding, and let

$$\boldsymbol{S}_j^{(t)} = (\boldsymbol{Z}_j^{(t)})^\top \boldsymbol{Z}_j^{(t)} + \lambda \boldsymbol{I}_t$$

be the corresponding regularized Gram matrix. For a candidate view i, let

$$\boldsymbol{z}_{ij} = \gamma_{ij}\boldsymbol{r}_{ij} \in \mathbb{R}^d$$

be the weighted embedding that would be obtained if view i were queried. After adding this candidate, the augmented weighted embedding matrix becomes

$$\boldsymbol{Z}_j^{(t+1)}(i) = \big[\, \boldsymbol{Z}_j^{(t)}, \boldsymbol{z}_{ij} \,\big] \in \mathbb{R}^{d\times(t+1)}.$$

Therefore,

$$\boldsymbol{S}_j^{(t+1)}(i) = \big(\boldsymbol{Z}_j^{(t+1)}(i)\big)^\top \boldsymbol{Z}_j^{(t+1)}(i) + \lambda \boldsymbol{I}_{t+1}.$$

Expanding the product gives

$$\big(\boldsymbol{Z}_j^{(t+1)}(i)\big)^\top \boldsymbol{Z}_j^{(t+1)}(i) = \begin{bmatrix} (\boldsymbol{Z}_j^{(t)})^\top \\ \boldsymbol{z}_{ij}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{Z}_j^{(t)} & \boldsymbol{z}_{ij} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{Z}_j^{(t)})^\top \boldsymbol{Z}_j^{(t)} & (\boldsymbol{Z}_j^{(t)})^\top \boldsymbol{z}_{ij} \\ \boldsymbol{z}_{ij}^\top \boldsymbol{Z}_j^{(t)} & \boldsymbol{z}_{ij}^\top \boldsymbol{z}_{ij} \end{bmatrix}.$$

Now define

$$\boldsymbol{b}_{ij}^{(t)} \triangleq (\boldsymbol{Z}_j^{(t)})^\top \boldsymbol{z}_{ij} \in \mathbb{R}^t.$$

Since $\|\boldsymbol{r}_{ij}\|_2 = 1$ and $\boldsymbol{z}_{ij} = \gamma_{ij}\boldsymbol{r}_{ij}$, we also have

$$\boldsymbol{z}_{ij}^\top \boldsymbol{z}_{ij} = \|\boldsymbol{z}_{ij}\|_2^2 = \gamma_{ij}^2.$$

Adding the regularization term $\lambda \boldsymbol{I}_{t+1}$ blockwise yields

$$\boldsymbol{S}_j^{(t+1)}(i) = \begin{bmatrix} (\boldsymbol{Z}_j^{(t)})^\top \boldsymbol{Z}_j^{(t)} + \lambda \boldsymbol{I}_t & \boldsymbol{b}_{ij}^{(t)} \\ (\boldsymbol{b}_{ij}^{(t)})^\top & \gamma_{ij}^2 + \lambda \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}_j^{(t)} & \boldsymbol{b}_{ij}^{(t)} \\ (\boldsymbol{b}_{ij}^{(t)})^\top & \gamma_{ij}^2 + \lambda \end{bmatrix}.$$

Thus, the updated regularized Gram matrix has the claimed augmented block form.

C Proof of Theorem 4.4

Theorem 4.4 (Closed-form one-step uncertainty update)**.** *For any candidate view $i \in \mathcal{C}_j^{(t)}$ acquired for object j with t existing views, the one-step uncertainty update is given by*

$$u_j^{(t+1)}(i) - u_j^{(t)} = \log\Big(s_{ij}^{(t)}\Big). \tag{6}$$

Proof. Recall that

$$\boldsymbol{S}_j^{(t+1)}(i) = \begin{bmatrix} \boldsymbol{S}_j^{(t)} & \boldsymbol{b}_{ij}^{(t)} \\ (\boldsymbol{b}_{ij}^{(t)})^\top & \tilde{\gamma}_{ij}^2 + \lambda \end{bmatrix},$$

where $\boldsymbol{S}_j^{(t)} \in \mathbb{R}^{t\times t}$ is positive definite because

$$\boldsymbol{S}_j^{(t)} = (\boldsymbol{Z}_j^{(t)})^\top \boldsymbol{Z}_j^{(t)} + \lambda I_t, \qquad \lambda > 0.$$

Hence $(\boldsymbol{S}_j^{(t)})^{-1}$ exists. For a block matrix of the form

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix},$$

with $\boldsymbol{A}$ invertible, the determinant can be written using the Schur complement of $\boldsymbol{A}$:

$$\det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \det(\boldsymbol{A})\ \det(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}).$$

In our case, identify

$$\boldsymbol{A} = \boldsymbol{S}_j^{(t)}, \qquad \boldsymbol{B} = \boldsymbol{b}_{ij}^{(t)}, \qquad \boldsymbol{C} = (\boldsymbol{b}_{ij}^{(t)})^\top, \qquad \boldsymbol{D} = \gamma_{ij}^2 + \lambda.$$

Therefore, the Schur complement of $\boldsymbol{S}_j^{(t)}$ in $\boldsymbol{S}_j^{(t+1)}(i)$ is

$$\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B} = \gamma_{ij}^2 + \lambda - (\boldsymbol{b}_{ij}^{(t)})^\top (\boldsymbol{S}_j^{(t)})^{-1} \boldsymbol{b}_{ij}^{(t)}.$$

Since D is a scalar, its determinant is just itself, so applying the block determinant identity yields

$$\det \boldsymbol{S}_j^{(t+1)}(i) = \det \boldsymbol{S}_j^{(t)} \Big(\gamma_{ij}^2 + \lambda - (\boldsymbol{b}_{ij}^{(t)})^\top (\boldsymbol{S}_j^{(t)})^{-1} \boldsymbol{b}_{ij}^{(t)} \Big).$$

Taking logarithms on both sides gives

$$\log\det \boldsymbol{S}_j^{(t+1)}(i) = \log\det \boldsymbol{S}_j^{(t)} + \log\Big(\gamma_{ij}^2 + \lambda - (\boldsymbol{b}_{ij}^{(t)})^\top (\boldsymbol{S}_j^{(t)})^{-1} \boldsymbol{b}_{ij}^{(t)} \Big).$$

Finally, since

$$u_j^{(t)} = \log\det \boldsymbol{S}_j^{(t)}, \qquad u_j^{(t+1)}(i) = \log\det \boldsymbol{S}_j^{(t+1)}(i),$$

we obtain

$$u_j^{(t+1)}(i) - u_j^{(t)} = \log\Big(\tilde{\gamma}_{ij}^2 + \lambda - (\boldsymbol{b}_{ij}^{(t)})^\top (\boldsymbol{S}_j^{(t)})^{-1} \boldsymbol{b}_{ij}^{(t)} \Big),$$

which proves the claim.

Since $\boldsymbol{S}_j^{(t)} \succ 0$, the Schur complement formula gives

$$\det \boldsymbol{S}_j^{(t+1)}(i) = \det \boldsymbol{S}_j^{(t)} \Big(\gamma_{ij}^2 + \lambda - (\boldsymbol{b}_{ij}^{(t)})^\top (\boldsymbol{S}_j^{(t)})^{-1} \boldsymbol{b}_{ij}^{(t)} \Big). \tag{8}$$

Taking logarithms on both sides yields the claim. □

D Proof of Theorem 4.5

Theorem 4.5 (Criterion for one-step uncertainty reduction)**.** *For any candidate view $i \in \mathcal{C}_j^{(t)}$ acquired for object j with t existing views, we have the following uncertainty reduction criterion:*

$$u_j^{(t+1)}(i) < u_j^{(t)} \quad \Longleftrightarrow \quad \eta_{ij}^{(t)} > \gamma_{ij}^2 + \lambda - 1.$$

Proof. By Theorem 4.4,

$$u_j^{(t+1)}(i) - u_j^{(t)} = \log\big(\gamma_{ij}^2 + \lambda - \eta_{ij}^{(t)}\big).$$

Since $\log(\cdot)$ is strictly increasing, this quantity is negative if and only if

$$\gamma_{ij}^2 + \lambda - \eta_{ij}^{(t)} < 1,$$

which is equivalent to

$$\eta_{ij}^{(t)} > \gamma_{ij}^2 + \lambda - 1.$$ □

E Proof of Proposition 4.7

Proposition 4.7 (Monotonicity of expected reduction under FOSD)**.** *Suppose the FOSD assumption holds for object o_j. Then for any two candidates $i, i' \in \mathcal{C}_j^{(t)}$, $q_{ij} \geq q_{i'j} \implies \Delta_j(i) \geq \Delta_j(i')$. Consequently, any maximizer of q_{ij} over $\mathcal{C}_j^{(t)}$ is also a maximizer of the expected uncertainty-reduction utility $\Delta_j(i)$ and candidate views can be ranked by q_{ij}, yielding the proxy acquisition rule:*

$$i_j^\star \in \arg\max_{i \in \mathcal{C}_j^{(t)}} q_{ij}. \tag{7}$$

Proof. Let

$$X \triangleq s_{ij}^{(t)}, \qquad Y \triangleq s_{i'j}^{(t)},$$

viewed as positive random variables conditional on the current evidence $\mathcal{D}_j^{(t)}$. By assumption,

$$X \preceq_{\mathrm{FOSD}} Y,$$

meaning that their conditional distribution functions satisfy

$$F_X(x) \triangleq \mathbb{P}\Big(X \leq x \,\Big|\, \mathcal{D}_j^{(t)}\Big) \geq \mathbb{P}\Big(Y \leq x \,\Big|\, \mathcal{D}_j^{(t)}\Big) \triangleq F_Y(x) \qquad \forall x > 0.$$

Equivalently, for every threshold $x > 0$, X is conditionally more likely than Y to take smaller values.

Now consider the function

$$g(s) = -\log s, \qquad s > 0.$$

Since $\log s$ is strictly increasing on $(0, \infty)$, g is strictly decreasing on $(0, \infty)$. Because X and Y are positive almost surely, $g(X)$ and $g(Y)$ are well defined.

A standard property of first-order stochastic dominance states that if $X \preceq_{\mathrm{FOSD}} Y$, then for every decreasing measurable function g for which the expectations exist,

$$\mathbb{E}\Big[g(X)\Big|\mathcal{D}_j^{(t)}\Big] \geq \mathbb{E}\Big[g(Y)\Big|\mathcal{D}_j^{(t)}\Big].$$

Applying this with $g(s) = -\log s$, we obtain

$$\mathbb{E}\Big[-\log s_{ij}^{(t)}\Big|\mathcal{D}_j^{(t)}\Big] \geq \mathbb{E}\Big[-\log s_{i'j}^{(t)}\Big|\mathcal{D}_j^{(t)}\Big].$$

By definition of the exact BOED utility,

$$\Delta_j(i) = \mathbb{E}\Big[-\log s_{ij}^{(t)}\Big|\mathcal{D}_j^{(t)}\Big], \qquad \Delta_j(i') = \mathbb{E}\Big[-\log s_{i'j}^{(t)}\Big|\mathcal{D}_j^{(t)}\Big].$$

Therefore,

$$\Delta_j(i) \geq \Delta_j(i'),$$

which proves the claim. □

F Proof of Theorem 4.8

Theorem 4.8 (Comparative FOSD-based probabilistic guarantee)**.** *Suppose that for two feasible candidate views $i, i' \in \mathcal{C}_j^{(t)}$, $q_{ij} \geq q_{i'j}$ and the corresponding Schur scores $s_{ij}^{(t)}, s_{i'j}^{(t)}$ satisfies Assumption 4.6. Then*

$$\mathbb{P}\left(u_j^{(t+1)}(i) < u_j^{(t)} \,\middle|\, \mathcal{D}_j^{(t)}\right) \geq \mathbb{P}\left(u_j^{(t+1)}(i') < u_j^{(t)} \,\middle|\, \mathcal{D}_j^{(t)}\right).$$

Proof. Recall from the one-step update formula that for any candidate view i,

$$u_j^{(t+1)}(i) - u_j^{(t)} = \log s_{ij}^{(t)},$$

where the Schur complement satisfies $s_{ij}^{(t)} > 0$ almost surely, since it is the Schur complement of the positive definite matrix $\boldsymbol{S}_j^{(t+1)}(i)$ with respect to the positive definite principal block $\boldsymbol{S}_j^{(t)}$.

Because the logarithm is strictly increasing on $(0, \infty)$, we have

$$u_j^{(t+1)}(i) < u_j^{(t)} \quad \Longleftrightarrow \quad u_j^{(t+1)}(i) - u_j^{(t)} < 0 \quad \Longleftrightarrow \quad \log s_{ij}^{(t)} < 0 \quad \Longleftrightarrow \quad s_{ij}^{(t)} < 1.$$

Therefore, conditional on the current evidence $\mathcal{D}_j^{(t)}$,

$$\mathbb{P}\Big(u_j^{(t+1)}(i) < u_j^{(t)} \,\Big|\, \mathcal{D}_j^{(t)}\Big) = \mathbb{P}\Big(s_{ij}^{(t)} < 1 \,\Big|\, \mathcal{D}_j^{(t)}\Big).$$

Likewise,

$$\mathbb{P}\Big(u_j^{(t+1)}(i') < u_j^{(t)} \,\Big|\, \mathcal{D}_j^{(t)}\Big) = \mathbb{P}\Big(s_{i'j}^{(t)} < 1 \,\Big|\, \mathcal{D}_j^{(t)}\Big).$$

Now let

$$F_i(x) \triangleq \mathbb{P}\Big(s_{ij}^{(t)} \le x \,\Big|\, \mathcal{D}_j^{(t)}\Big), \qquad F_{i'}(x) \triangleq \mathbb{P}\Big(s_{i'j}^{(t)} \le x \,\Big|\, \mathcal{D}_j^{(t)}\Big)$$

denote the conditional distribution functions of the two Schur complements. By the assumed first-order stochastic dominance,

$$F_i(x) \ge F_{i'}(x) \qquad \forall x > 0.$$

In particular, evaluating this inequality at $x = 1$ gives

$$\mathbb{P}\Big(s_{ij}^{(t)} \le 1 \,\Big|\, \mathcal{D}_j^{(t)}\Big) \ge \mathbb{P}\Big(s_{i'j}^{(t)} \le 1 \,\Big|\, \mathcal{D}_j^{(t)}\Big).$$

Combining this with the equivalence between uncertainty reduction and the event $\{s < 1\}$, we conclude that

$$\mathbb{P}\Big(u_j^{(t+1)}(i) < u_j^{(t)} \,\Big|\, \mathcal{D}_j^{(t)}\Big) \ge \mathbb{P}\Big(u_j^{(t+1)}(i') < u_j^{(t)} \,\Big|\, \mathcal{D}_j^{(t)}\Big),$$

which proves the claim. □

G Redundancy Score Intuition

Recall that

$$\boldsymbol{b}_{ij}^{(t)} = (\boldsymbol{Z}_j^{(t)})^\top \boldsymbol{z}_{ij}, \qquad \eta_{ij}^{(t)} = (\boldsymbol{b}_{ij}^{(t)})^\top (\boldsymbol{S}_j^{(t)})^{-1} \boldsymbol{b}_{ij}^{(t)},$$

where $\boldsymbol{z}_{ij} = \gamma_{ij} \boldsymbol{r}_{ij}$ is the candidate weighted caption embedding and

$$\boldsymbol{S}_j^{(t)} = (\boldsymbol{Z}_j^{(t)})^\top \boldsymbol{Z}_j^{(t)} + \lambda \boldsymbol{I}_t.$$

Substituting $\boldsymbol{b}_{ij}^{(t)} = (\boldsymbol{Z}_j^{(t)})^\top \boldsymbol{z}_{ij}$ gives

$$\eta_{ij}^{(t)} = \boldsymbol{z}_{ij}^\top \boldsymbol{Z}_j^{(t)} (\boldsymbol{S}_j^{(t)})^{-1} (\boldsymbol{Z}_j^{(t)})^\top \boldsymbol{z}_{ij}.$$

Thus, if we define

$$\boldsymbol{P}_j^{(t)} \triangleq \boldsymbol{Z}_j^{(t)} (\boldsymbol{S}_j^{(t)})^{-1} (\boldsymbol{Z}_j^{(t)})^\top \in \mathbb{R}^{d \times d},$$

then

$$\eta_{ij}^{(t)} = \boldsymbol{z}_{ij}^\top P_j^{(t)} \boldsymbol{z}_{ij}.$$

The matrix $\boldsymbol{P}_j^{(t)}$ acts as a *regularized projection-like operator* onto the span of the previously selected weighted embeddings. Consequently, the quadratic form $\eta_{ij}^{(t)}$ measures how much of the candidate vector $\boldsymbol{z}_{ij}$ lies in directions that are already explained by the current evidence. In particular, $\eta_{ij}^{(t)}$ is large when $\boldsymbol{z}_{ij}$ is well aligned with the span of the columns of $Z_j^{(t)}$, and small when $\boldsymbol{z}_{ij}$ points in a semantically new direction.

The term γ_{ij}^2 has an equally direct interpretation. Since $\|\boldsymbol{r}_{ij}\|_2 = 1$,

$$\|\boldsymbol{z}_{ij}\|_2^2 = \|\gamma_{ij} \boldsymbol{r}_{ij}\|_2^2 = \gamma_{ij}^2.$$

Hence γ_{ij}^2 is simply the squared norm of the candidate weighted embedding. Larger γ_{ij} corresponds to greater internal incoherence, so more incoherent candidate captions contribute larger total weighted magnitude.

Combining these two facts, the Schur complement term

$$\gamma_{ij}^2 + \lambda - \eta_{ij}^{(t)}$$

can be interpreted as a *regularized residual energy*: it is the total weighted magnitude of the candidate embedding, plus regularization, minus the portion already explained by the currently selected views. Therefore, uncertainty decreases precisely when the candidate is sufficiently explained by the existing evidence relative to its own incoherence.

Idealized example. This interpretation becomes exact in the idealized case where $\lambda = 0$ and the columns of $\boldsymbol{Z}_j^{(t)}$ are orthonormal. Then

$$\boldsymbol{S}_j^{(t)} = (\boldsymbol{Z}_j^{(t)})^\top \boldsymbol{Z}_j^{(t)} = \boldsymbol{I}_t,$$

so

$$\boldsymbol{P}_j^{(t)} = \boldsymbol{Z}_j^{(t)} (\boldsymbol{Z}_j^{(t)})^\top,$$

which is the orthogonal projector onto $\mathrm{span}(\boldsymbol{Z}_j^{(t)})$. In this case,

$$\eta_{ij}^{(t)} = \boldsymbol{z}_{ij}^\top P_j^{(t)} \boldsymbol{z}_{ij} = \left\| \boldsymbol{P}_j^{(t)} \boldsymbol{z}_{ij} \right\|_2^2,$$

so $\eta_{ij}^{(t)}$ is exactly the squared norm of the component of $\boldsymbol{z}_{ij}$ that lies in the span of the already selected weighted embeddings. If $\boldsymbol{z}_{ij}$ lies entirely in that span, then $\eta_{ij}^{(t)} = \|\boldsymbol{z}_{ij}\|_2^2 = \gamma_{ij}^2$; if it is orthogonal to that span, then $\eta_{ij}^{(t)} = 0$. The general regularized case preserves the same intuition, but with a softened projection induced by $(\boldsymbol{S}_j^{(t)})^{-1}$.

Proposition G.1 (Lower bound from approximate stochastic dominance)**.** *Let*

$$s_{ij}^{(t)} = \tilde{\gamma}_{ij}^2 + \lambda - \eta_{ij}^{(t)}$$

be the random Schur complement for candidate view i, conditional on the current evidence $\mathcal{D}_j^{(t)}$. Suppose there exists a reference random variable $S^\star$ and a constant $\varepsilon \in [0, 1]$ such that

$$\mathbb{P}\Big(s_{ij}^{(t)} \le x \,\Big|\, \mathcal{D}_j^{(t)}\Big) \;\ge\; \mathbb{P}(S^\star \le x) - \varepsilon \qquad \forall x > 0.$$

Then

$$\mathbb{P}\Big(u_j^{(t+1)}(i) < u_j^{(t)} \,\Big|\, \mathcal{D}_j^{(t)}\Big) \;\ge\; \mathbb{P}(S^\star < 1) - \varepsilon.$$

Proof. Since

$$u_j^{(t+1)}(i) - u_j^{(t)} = \log s_{ij}^{(t)},$$

we have

$$u_j^{(t+1)}(i) < u_j^{(t)} \quad \Longleftrightarrow \quad s_{ij}^{(t)} < 1.$$

Therefore

$$\mathbb{P}\Big(u_j^{(t+1)}(i) < u_j^{(t)} \,\Big|\, \mathcal{D}_j^{(t)}\Big) = \mathbb{P}\Big(s_{ij}^{(t)} < 1 \,\Big|\, \mathcal{D}_j^{(t)}\Big).$$

Applying the assumed CDF bound at $x = 1$ yields

$$\mathbb{P}\Big(s_{ij}^{(t)} < 1 \,\Big|\, \mathcal{D}_j^{(t)}\Big) \;\ge\; \mathbb{P}(S^\star < 1) - \varepsilon,$$

which proves the claim. □

Proposition 4.9 (Absolute reduction bound from Schur-score moments)**.** *Let $\mu_{ij} = \mathbb{E}[s_{ij}^{(t)} \mid \mathcal{D}_j^{(t)}]$ and $\sigma_{ij}^2 = \mathrm{Var}(s_{ij}^{(t)} \mid \mathcal{D}_j^{(t)}) < \infty$. If $\mu_{ij} < 1$, then*

$$\mathbb{P}\Big(u_j^{(t+1)}(i) < u_j^{(t)} \,\Big|\, \mathcal{D}_j^{(t)}\Big) \ge \frac{(1 - \mu_{ij})^2}{\sigma_{ij}^2 + (1 - \mu_{ij})^2}.$$

Proof. Since

$$u_j^{(t+1)}(i) < u_j^{(t)} \quad \Longleftrightarrow \quad s_{ij}^{(t)} < 1,$$

it suffices to bound $\mathbb{P}(s_{ij}^{(t)} < 1 \mid \mathcal{D}_j^{(t)})$. By Cantelli's inequality,

$$\mathbb{P}\Big(s_{ij}^{(t)} - \mu_{ij} \geq 1 - \mu_{ij} \,\Big|\, \mathcal{D}_j^{(t)}\Big) \leq \frac{\sigma_{ij}^2}{\sigma_{ij}^2 + (1-\mu_{ij})^2}.$$

Hence

$$\mathbb{P}\Big(s_{ij}^{(t)} < 1 \,\Big|\, \mathcal{D}_j^{(t)}\Big) \geq 1 - \frac{\sigma_{ij}^2}{\sigma_{ij}^2 + (1-\mu_{ij})^2} = \frac{(1-\mu_{ij})^2}{\sigma_{ij}^2 + (1-\mu_{ij})^2}. \qquad \square$$

H LLM-based Captions Fusion Details

After uncertainty-driven view selection, each object is associated with a small set of captions obtained from multiple views. We fuse these captions into a single object description using a constrained LLM-based aggregation step. The prompt is designed to exploit the acquisition structure of our pipeline: the first few captions correspond to the initial representative views, while later captions come from refinement views that are assumed to be resolve uncertainty. Accordingly, the LLM is instructed to regard later captions as more reliable evidence when conflicts arise. If a fact disagrees between initial and refined views, the refined-view version is preferred. If disagreement exists only among the initial views, the conflict is not suppressed; instead, the fused caption retains the alternatives using cautious natural-language constructions such as *"white or gray"* or *"possibly wooden or metal."* This preserves uncertainty rather than collapsing it prematurely into a single claim.

Formally, let $\{c_{j1}, \ldots, c_{jr_j}\}$ denote the captions for object o_j, ordered by acquisition stage, with captions from the potential refinement stage placed after the initial ones. The fusion module takes this ordered caption list as input and prompts the LLM to generate a structured JSON output containing three fields: a fused caption, the facts used in that caption, and the facts omitted during fusion. The LLM is explicitly constrained to satisfy three requirements: (i) it may only use information that appears in at least one source caption, preventing unsupported hallucinations; (ii) it must privilege facts from refined views when they conflict with facts from initial views; and (iii) it must expose unresolved ambiguity from the initial views rather than silently discarding it. In this way, fusion serves as a semantic consolidation layer over the multi-view captions: consistent facts are merged into a unified object description, refined views correct lower-quality initial evidence, and residual ambiguity is preserved in the language when the evidence remains unresolved. We present the detailed prompt in Appendix H. Appendix H shows the system prompt used to fuse multiple captions of the same object into a single concise description while preserving uncertainty and handling conflicts between initial and refined views. We use `gpt-4o-mini` from OpenAI's API for caption fusion, and the prompt is designed to elicit structured JSON output containing the fused caption, t he facts used in that caption, and any omitted facts. The prompt explicitly instructs the LLM to avoid hallucinating new information, to privilege refined views over initial views when conflicts arise, and to surface unresolved ambiguity rather than silently discarding it.

System prompt for caption fusion

```
SYSTEM_PROMPT = """\
You are a scene-description assistant that fuses multiple captions of the same object
    into one.

Caption ordering:
  - Captions are numbered starting from [1]. The first three ([1]-[3]) are initial views.
  - If more than three captions are provided, captions [4] onwards are refined views with
     higher quality.
  - When facts conflict, prefer the version stated in refined views ([4]+) over initial
    views ([1]-[3]).

Handling conflicts:
  - For facts that conflict across initial views only, keep all alternatives using
    natural phrasing
    (e.g. "white or gray", "appears round or oval", "possibly wooden or metal").
  - For facts that conflict between initial and refined views, prefer the refined view's
    version and
    drop the initial view's version.
  - Never silently drop a conflicting fact from the initial views --- surface it as an
    alternative.

Other rules:
  - Never introduce new information not present in any source caption.
  - The fused caption should be concise (one to three sentences).

Respond ONLY with valid JSON matching this schema:
{
  "fused_caption": "<one to three sentences>",
  "used_facts": ["<fact 1>", ...],
  "omitted_facts": ["<fact that was dropped>", ...]
}
"""
```

I Algorithm Details

Algorithm 1 UQ-DAAAM: Uncertainty Quantification and Active Object Refinement

Require: RGB-D frame stream $\mathcal{F} = \{f_1, \dots, f_n\}$, object fragments $\mathcal{O} = \{o_1, \dots, o_m\}$, visibility indicators v_{ij}, DAAAM heuristic scores q_{ij}, refinement budget B, uncertainty threshold τ
Ensure: Refined object memories $\{\mathcal{V}_j, u_j, \hat{c}_j\}_{j=1}^{m}$

Run the standard DAAAM pipeline to construct the initial memory and obtain, for each object o_j, an initial set of selected views $\mathcal{V}_j$, where $|\mathcal{V}_j| = r$, and captions $\{c_{\ell j}\}_{\ell=1}^{r}$
for each object $o_j \in \mathcal{O}$ **do**
    Compute normalized caption embeddings $\boldsymbol{r}_{\ell j} \leftarrow \frac{\phi(c_{\ell j})}{\|\phi(c_{\ell j})\|_2}, \qquad \ell = 1, \dots, r$
    Compute incoherence weights $\gamma_{\ell j} \leftarrow \exp\big(\alpha(1 - p_{\ell j})\big)$
    Form weighted embedding matrix $\boldsymbol{Z}_j \leftarrow [\, \gamma_{1j}\boldsymbol{r}_{1j}, \dots, \gamma_{rj}\boldsymbol{r}_{rj} \,]$
    Form semantic scatter and uncertainty $\boldsymbol{S}_j \leftarrow \boldsymbol{Z}_j^{\top}\boldsymbol{Z}_j + \lambda \boldsymbol{I}, \qquad u_j \leftarrow \log\det \boldsymbol{S}_j$
end for
Identify uncertain objects $\mathcal{U}_\tau \leftarrow \{\, j \in [m] : u_j > \tau \,\}$
for each object o_j with $j \in \mathcal{U}_\tau$ **do**
    Initialize object-level refinement budget $b_j \leftarrow B$
    while $b_j > 0$ **do**
        Form feasible candidate-view set $\mathcal{C}_j \leftarrow \{\, i \in [n] \setminus \mathcal{V}_j : v_{ij} = 1 \,\}$
        if $\mathcal{C}_j = \emptyset$ **then**
            break
        end if
        Select the next refinement view using the proxy rule $i^\star \in \arg\max_{i \in \mathcal{C}_j} q_{ij}$
        Query the VLM on object o_j in view $f_{i^\star}$ to obtain a new caption c
        Update the semantic scatter and uncertainty online via Equation (4)
        Append view: $\mathcal{V}_j \leftarrow \mathcal{V}_j \cup \{i^\star\}$
        Decrease object-level budget: $b_j \leftarrow b_j - 1$
    end while
end for
for each object $o_j \in \mathcal{O}$ **do**
    Fuse the multi-view captions associated with $\mathcal{V}_j$ into a final caption $\hat{c}_j$
end for
return $\{\mathcal{V}_j, u_j, \hat{c}_j\}_{j=1}^{m}$

J Uncertainty Quality Experiments

A central claim of the paper is that the proposed uncertainty score is meaningful. We isolate uncertainty measurement from view selection by evaluating the quality of the uncertainty scores themselves, without using them to select views. In our experiments, we use Llava-v1.5-13b [31], with CLIP-ViT-L-336px [40] as its vision backbone. Following [25, 27], we evaluate our uncertainty metric's performance on three distinct visual question-answering (VQA) benchmark datasets: VQAv2 [14], OKVQA [36], and AdVQA [28]. We compare our metric against a variety of uncertainty-quantification baselines:

- **Neighborhood Consistency** [23]: examines the consistency of the model responses over rephrased questions generated by a small proxy VQG model.
- **LN-Entropy** [34]: normalizes the joint log-probability of each sequence by dividing it by the sequence length sampled via multinomial sampling.
- **Semantic Entropy** [25]: models the uncertainty over different meanings by clustering the generated sequences by [17] and measuring the clusters' entropies.
- **EigenScore** [7]: computes the log determinant of the covariance matrix of response embeddings via SVD.

Let $u^\star_{\text{correct}}$ and $u^\star_{\text{incorrect}}$ denote the uncertainty scores of correct and incorrect responses, respectively. Our target quantity is the pairwise ranking probability $\mathbb{P}(u^\star_{\text{correct}} < u^\star_{\text{incorrect}})$, *i.e.*, how often a randomly chosen correct response receives a lower uncertainty score than a randomly chosen incorrect response. Following [27, 25], we summarize this ranking quality with the **AUROC** of a binary classifier that uses $u^\star$ to predict correctness (lower $u^\star$ indicates higher confidence / more likely correct).

	Method	AUROC (↑)			Pearson Corr. (↑)			Spearman Corr. (↑)			ECE (↓)			AURAC (↑)		
		VQA	OK	AdV	VQA	OK	AdV	VQA	OK	AdV	VQA	OK	AdV	VQA	OK	AdV
Baselines	Neighborhood	0.761	0.512	0.660	0.789	0.577	0.543	0.712	0.580	0.703	0.331	0.502	0.353	0.886	0.629	0.683
	LN-Entropy	0.763	0.698	0.639	0.563	0.856	0.913	0.566	0.699	0.900	0.055	0.057	0.072	0.899	0.741	0.705
	Semantic Entropy	0.840	0.712	0.757	0.909	0.444	0.799	0.897	0.532	0.781	0.056	0.189	0.182	0.902	0.734	0.742
	EigenScore	0.850	0.741	0.766	0.922	0.790	0.872	0.883	0.842	0.811	0.056	0.190	0.211	0.913	0.753	0.753
	Ours	0.853	0.746	0.763	0.902	0.895	0.941	0.852	0.901	0.875	0.041	0.041	0.043	0.862	0.743	0.710

Table 3: Uncertainty evaluation on VQAv2, OKVQA, and AdVQA. We report AUROC, calibration correlations (Pearson, Spearman), Expected Calibration Error (ECE; lower is better), and AURAC (higher is better).

We test whether the (*normalized*) uncertainty score is calibrated in the sense that $\mathbb{P}(\text{correct} \mid u^\star) \approx 1 - u^\star$. We follow the standard binning-based calibration protocol of [16]: we sort and bin instances by $u^\star$ and compute per-bin average uncertainty and empirical accuracy. We report: **Calibration correlation**, the correlation between binned mean uncertainty and binned empirical error. We provide both **Pearson** and **Spearman** correlations. **Expected Calibration Error (ECE)**, which measures the (weighted) average absolute gap between binned predicted confidence and binned accuracy.

As shown in Table 3, our metric consistently outperforms all baselines on both discrimination (AUROC/TPR/FPR) and calibration (correlations/ECE), including multimodal-targeted methods such as UMPIRE and Neighborhood Consistency. On the more challenging datasets (OKVQA and AdVQA), where model outputs exhibit higher diversity and incoherence, our metric remains robust and provides substantially better correctness prediction.

We also evaluate whether our metric can improve *selective answering—i.e.*, abstaining on uncertain queries and answering only when the model is confident.

Let $u_q^\star$ be the uncertainty score for query q (lower $u^\star$ indicates higher confidence). For any threshold θ, the selective policy answers iff $u_q^\star \leq \theta$ and abstains otherwise. Varying θ induces a trade-off between *coverage* (fraction answered) and *accuracy* on the answered subset. Following [10, 27, 21], we plot the Rejection–Accuracy curve: we sort queries by $u^\star$ and progressively reject the most uncertain fraction; at each rejection rate $r \in [0, 1]$ (equivalently, coverage $1 - r$), we compute the accuracy on the remaining answered queries. We summarize the overall benefit of uncertainty-based abstention with the **Area Under the Rejection–Accuracy Curve (AURAC)**, which aggregates performance across all thresholds (higher is better).

As shown in Table 3, our metric consistently achieves the highest AURAC across datasets. This indicates that fusing dual-source uncertainty yields a more reliable abstention signal: the model preferentially rejects queries that are likely to be incorrect, thereby improving accuracy on the answered subset. In practical VQA settings, this supports user-facing behaviors where the system abstains when uncertain rather than returning low-confidence answers.

K Experiments Details

Dataset and preprocessing. We evaluate on the CODa outdoor driving dataset [57], which provides 30,Hz stereo images, LiDAR-projected depth maps, and globally consistent poses. All experiments use the front camera (`cam0`) images, center-cropped and resized to 480×640 pixels. Depth is sourced from the `3d_raw_estimated` LiDAR projection, filtered to the range $[0.05, 20.0]$m. Frames are processed at 10Hz (the sensor publishing rate).

Scene reconstruction. We use Khronos [45] for real-time 3D scene graph construction, with a TSDF voxel size of 0.2,m and truncation distance of 0.6,m. Object fragments are extracted with a minimum volume of 0.005m^3 and maximum of 50m^3, requiring at least 8 observations and a minimum allocation confidence of 0.5. Traversability places are computed with widths in $[0.4, 2.0]$m, with label projection to ground planes enabled.

Segmentation and tracking. We use FastSAM [62] compiled to a TensorRT engine at 480×640 resolution for open-vocabulary instance segmentation, with a minimum mask area of 400 pixels. Object tracking uses ByteTrack [60] with appearance re-identification via a CLIP-based ReID model (TensorRT), a new-track threshold of 0.6, and a track buffer of 30 frames. A track must accumulate at least 8 observations before it qualifies for semantic annotation.

Method	Accuracy ↑	Position [m] ↓	Temporal [min] ↓
DAAAM-RR	0.716 ± 0.014	41.18 ± 2.31	1.769 ± 0.089
DAAAM-QR	0.738 ± 0.011	40.14 ± 1.98	1.734 ± 0.072
DAAAM-NC [23]	0.713 ± 0.016	41.62 ± 2.54	1.783 ± 0.094
DAAAM-LN [34]	0.714 ± 0.015	41.55 ± 2.48	1.776 ± 0.091
DAAAM-Ent [25]	0.733 ± 0.013	40.33 ± 2.17	1.718 ± 0.083
DAAAM-ES [7]	0.743 ± 0.012	39.69 ± 2.02	1.683 ± 0.078
DAAAM-QR+Best	0.736 ± 0.011	39.68 ± 1.92	1.781 ± 0.085
UQ-DAAAM (Ours)	$\mathbf{0.761} \pm 0.009$	$\mathbf{37.84} \pm 1.74$	$\mathbf{1.589} \pm 0.063$

Table 4: Mean ± standard deviation on OC-NaVQA (All split). Results averaged over three seeds.

Semantic annotation. We use DAM-3B [29] as the vision-language model for object captioning, operating in focal-prompt mode with batch sizes of 32–128 masks. Frame selection uses Perception Encoder (PE-Core-L14-336) CLIP features to select the most visually informative frame per track. Per-frame CLIP features (ViT-B/16) are computed every 5 frames for temporal scene understanding. Semantic annotation queries are triggered every 120 frames.

Cross-view UQ. We use the cross-view uncertainty quantification mode, where the assignment BLP [13] assigns r=3 representative frames per tracked fragment (with a minimum of 2 views required). The BLP is solved with GLPK_MI (15s timeout), with a greedy set-cover fallback. Each assigned view produces an independent VLM caption; the log-determinant Gram uncertainty is computed from the r L2-normalised caption embeddings with $\lambda = 10^{-8}$. We additionally incorporate the UMPIRE [27] coherence term with weight $\alpha = 1.0$, scaling each embedding by its incoherence weight $\gamma_i = \alpha(1 - \exp(\ell\ell_i))$ where $\ell\ell_i$ is the per-sequence log-likelihood.

Active view selection. Objects whose initial UQ score exceeds the threshold $\tau = -0.25$ are selected for refinement. Each flagged object receives up to b=2 additional views, chosen to maximally reduce the log-determinant via the rank-1 Schur complement update: $\log\det(\mathbf{G}^{(t+1)}) = \log\det(\mathbf{G}^{(t)}) + \log(c - \mathbf{b}^\top(\mathbf{G}^{(t)})^{-1}\mathbf{b})$, where $\mathbf{b}$ and c are the cross-terms and self-similarity of the new view embedding. The Gram inverse is updated via the Sherman–Morrison formula, avoiding recomputation. Refinement runs asynchronously during pipeline shutdown with a timeout of 300s.

Downstream evaluation. For the NaVQA benchmark [2], we use GPT-5-mini for question answering over the constructed scene graph memory, with sentence embeddings computed by Sentence-T5-XL [37]. We evaluate on 7 CODa sequences (0, 3, 4, 6, 16, 21, 22) totalling 98,866 frames. All experiments are conducted on a single NVIDIA RTX 4090 GPU with 24GB VRAM, using 1 assignment worker and 1 grounding worker.

L Additional View-Selection Results

Table 5 records the additional view-selection results. The uncertainty measurements come from a smaller-scale experiment using the first 5,000 frames of CODa sequence 3, with a per-object refinement budget of $B = 2$ and uncertainty threshold $\tau = -0.25$.

Heuristic	URR ↑	u_{post} ↓	Accuracy ↑	Position [m] ↓	Temporal [min] ↓
Random	0.221	0.630	0.735	40.9	1.74
PoseDiv	0.261	0.634	0.740	40.2	1.71
Sim-CLIP	0.256	0.643	0.739	40.4	1.72
Sharp	0.294	0.611	0.744	39.6	1.69
Div-CLIP	0.347	0.631	0.759	37.9	1.57
q (ours)	0.438	0.565	0.761	37.8	1.59

Table 5: Additional next-view heuristic results.

M Indoor Evaluation on VSI-Bench Source Scenes

We additionally evaluate UQ-DAAAM on four ScanNet scenes [9] from the source data of VSI-Bench [53], comprising 446 mapped objects. This experiment examines uncertainty reduction and budgeted refinement in indoor environments. Our pipeline uses the underlying RGB-D observations and scene geometry; these results are therefore not directly comparable with VSI-Bench's video-only question-answering evaluation.

Scene	Objects	Refined (%)	u_{pre}	u_{post}	Mean $\Delta u \downarrow$	Decrease rate (%) $\uparrow$
`scene0591_01`	185	38.9	0.903	0.798	-0.105	98.6
`scene0651_02`	85	18.8	0.823	0.550	-0.273	100.0
`scene0356_00`	84	28.6	0.866	0.585	-0.281	100.0
`scene0221_01`	92	42.4	0.782	0.488	-0.295	100.0
All	446	33.9	0.857	0.658	-0.200	99.3

Table 6: Uncertainty refinement on four indoor ScanNet scenes from VSI-Bench's source data. Uncertainty scores are min–max normalized. "Refined" denotes the fraction of mapped objects selected for refinement; uncertainty statistics are reported over the refined subset. The decrease rate measures the fraction with $u_{\mathrm{post}} < u_{\mathrm{pre}}$, rather than the fraction crossing the uncertainty threshold used to define URR.

As shown in Table 6, refinement lowers mean normalized uncertainty from 0.857 to 0.658, with a decrease for 99.3% of refined objects. These results provide additional evidence of uncertainty reduction in indoor scenes; they do not independently establish improved caption correctness or downstream question-answering performance.

N External Evaluation of Caption Correctness

To assess whether uncertainty-guided refinement improves caption correctness beyond reducing our internal uncertainty score, we evaluate captions using GPT-4o as an external judge. GPT-4o belongs to a different model family from the DAM-3B captioner. For each evaluated object, the judge checks the caption against a held-out view of that object, using a frame that was not used to produce the caption, and labels factual errors.

The initial-view uncertainty score predicts judge-labeled caption errors with an AUROC of 0.74, compared with a chance level of 0.50. On the initially uncertain subset, refinement reduces the judge-labeled factual-error rate from 37% to 18%, an absolute reduction of 19 percentage points. These results provide evidence that the uncertainty score identifies captions with factual errors and that refinement improves caption correctness under an external evaluator. This evaluation uses automated model judgments rather than human-verified correctness labels.

O Interface to Other Memory Architectures

The uncertainty computation requires an object association, a collection of normalized per-view embeddings, and positive per-view weights. Given these inputs, the weighted Gram matrix, log-determinant score, and Schur-complement update do not depend on DAAAM's scene-graph implementation or on a particular captioning model.

For a memory that retains multiple image observations per object, an analogous score could be constructed from normalized image embeddings and suitable per-view reliability weights. However, variation in image embeddings may reflect appearance and viewpoint changes rather than caption unreliability. Such a transfer therefore requires a new interpretation of the score, threshold calibration, and downstream validation.

Integrating budgeted refinement additionally requires a candidate-view interface and an acquisition policy. The algebraic update transfers to other embeddings, whereas the relationship between view quality and expected uncertainty reduction must be assessed for each representation and task. Evaluation within other navigation and memory systems remains future work.

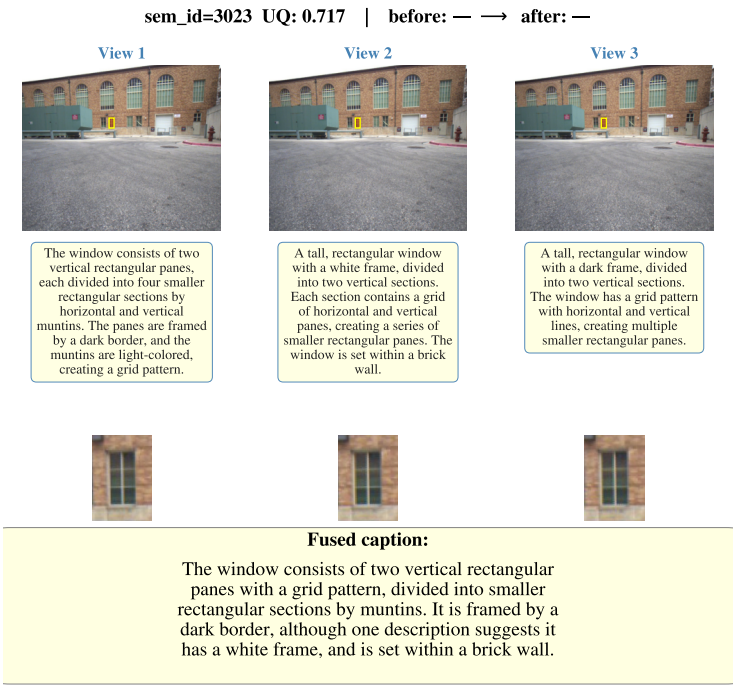
P Memory and Runtime Details

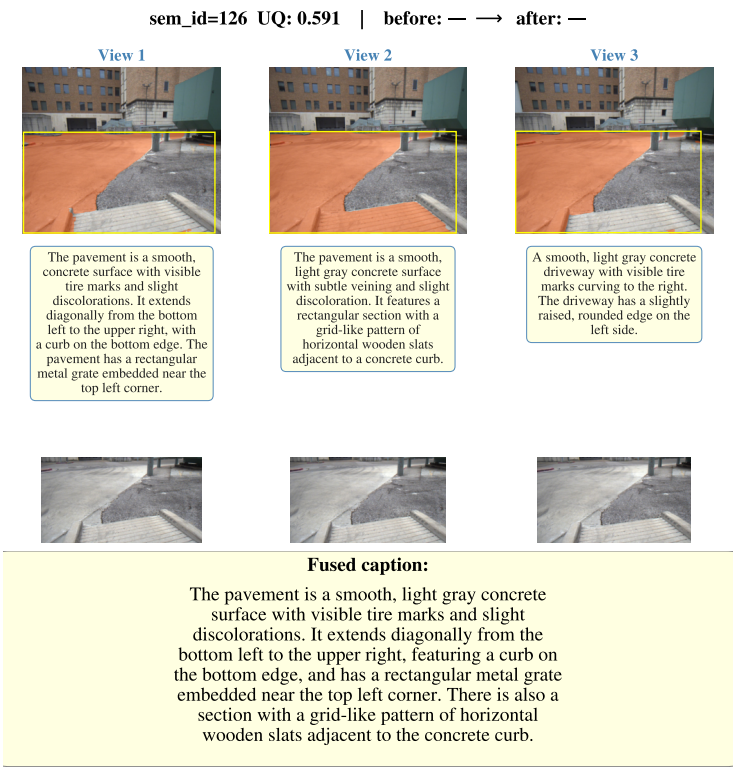
We evaluate the overhead introduced by cross-view uncertainty quantification and active view selection on the same CODa sequences, using a single NVIDIA RTX 4090 GPU. Across 7 sequences (98,866 frames total), UQ-DAAAM achieves a mean frame rate of 13.5 Hz, comfortably exceeding the 10 Hz sensor rate of the CODa dataset and confirming that the UQ extensions preserve real-time operation. The cross-view BLP solver, which assigns $r{=}3$ views per tracked fragment, adds a mean latency of 131.7 ± 132.5 ms per query batch, accounting for only 1.49% of total frame processing time. This is because the BLP runs on CPU via `cvxpy` and executes infrequently (828 calls across 98,866 frames). Each object receives 3.44 ± 0.68 cross-view captions on average, from which the log-determinant Gram uncertainty is computed via `numpy.linalg.slogdet` entirely on CPU with negligible overhead. Of the 18,298 objects annotated across all sequences, 6,324 (34.6%) were selected for active view refinement based on their initial uncertainty, receiving 1.32 ± 0.47 additional views each. Active view selection reduced the mean UQ score from 0.56 to 0.32, indicating substantially increased caption agreement. The additional VLM calls for refinement occur asynchronously in the grounding worker thread during shutdown, with a configurable timeout. GPU memory consumption remains within the 24 GB envelope of the RTX 4090, as the UQ and BLP components are CPU-only and the VLM model (DAM-3B) is shared between initial annotation and refinement.

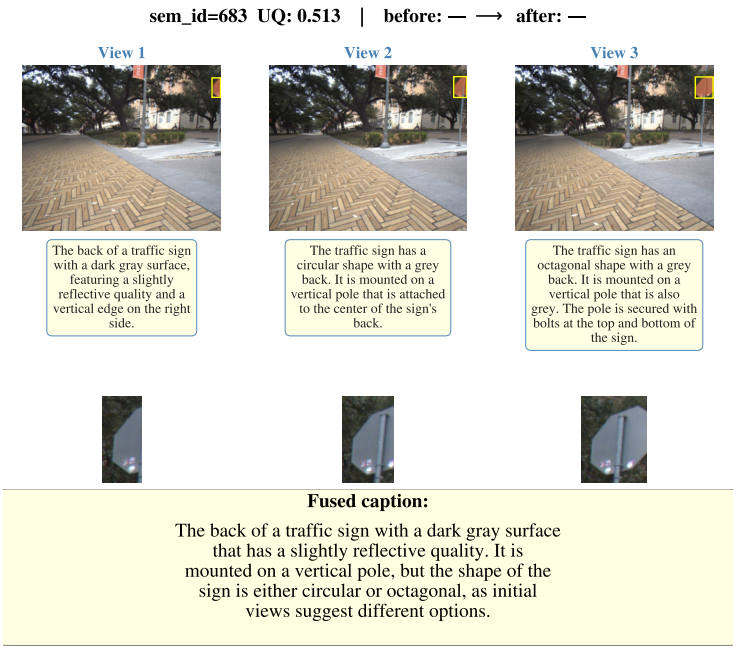
Q Qualitative Results and Visualizations

In the following qualitative examples, we present the captions and uncertainty scores for both initial and refined views of selected objects. We show both objects with low uncertainty, which do not require refinement (before and after view selection are the same), and objects with high uncertainty, where the refined views lead to more consistent captions and reduced uncertainty. In each visualization, we show the raw RGB images, object masks, and zoomed-in views of the masked objects for clarity.

Q.1 Examples of Objects with Low Uncertainty



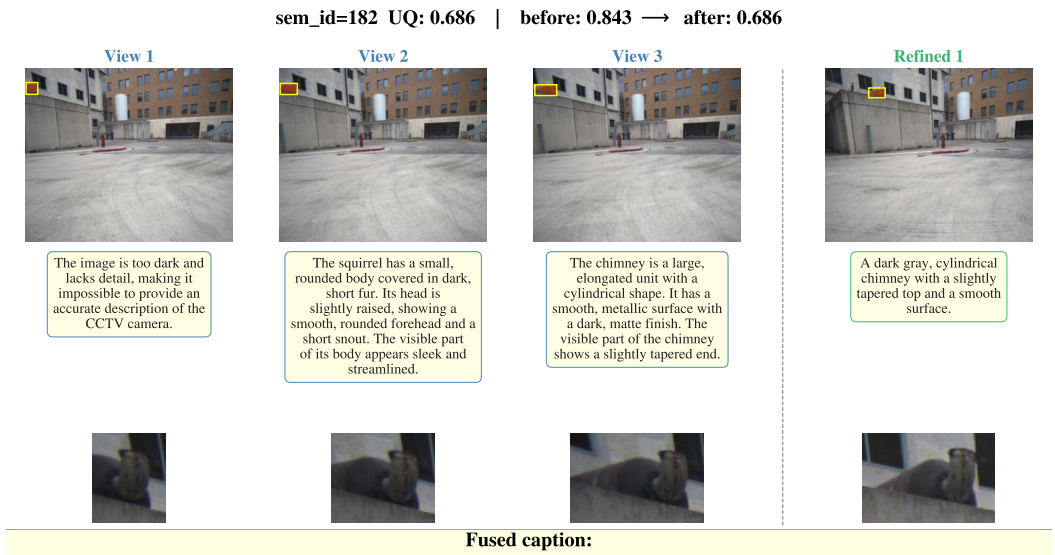




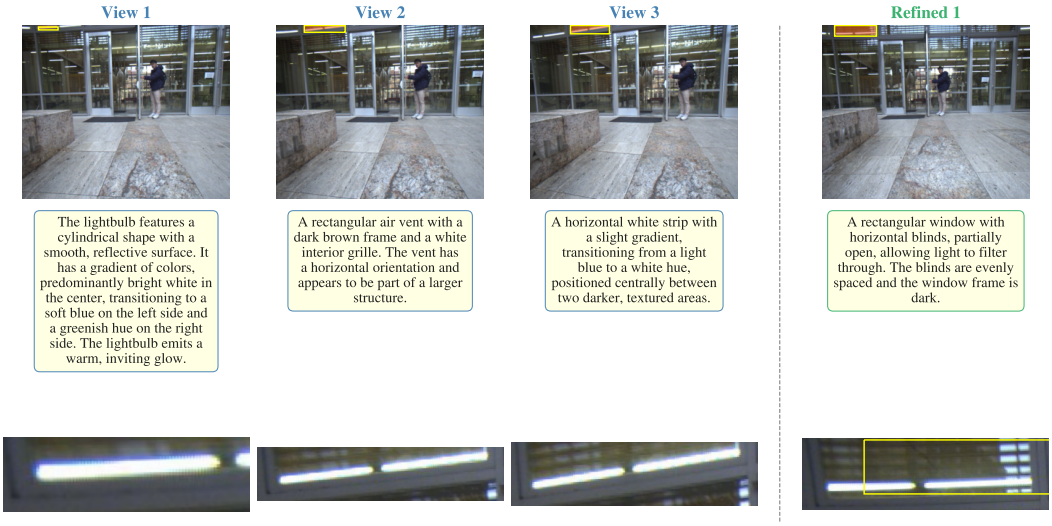
No refinement views are needed for these objects, since the initial views already provide a stable and coherent semantic summary. The captions are consistent across views, and the uncertainty scores are low, indicating that the system is confident in its understanding of these objects.

Q.2 Examples of Objects with High Uncertainty





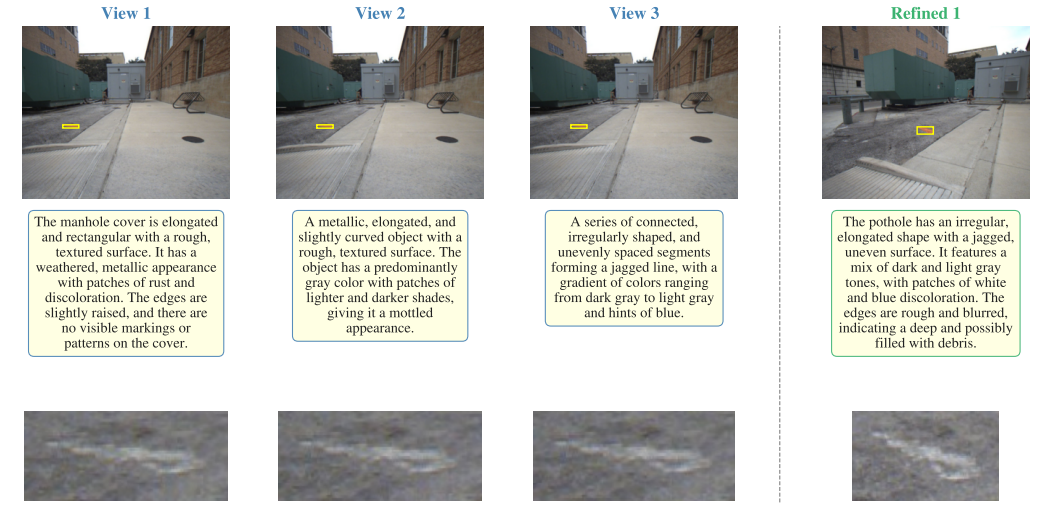
sem_id=1523 UQ: 0.719 | before: 0.836 ⟶ after: 0.719



Fused caption:

The object is a rectangular window with horizontal blinds, partially open, allowing light to filter through. The blinds are evenly spaced and the window frame is dark.

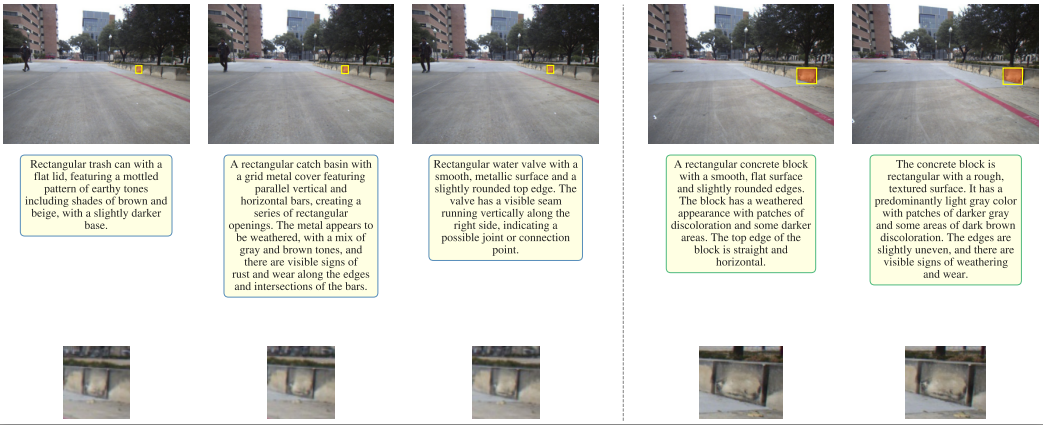
sem_id=29 UQ: 0.747 | before: 0.906 ⟶ after: 0.747



Fused caption:

The pothole has an irregular, elongated shape with a jagged, uneven surface featuring a mix of dark and light gray tones, and patches of white and blue discoloration. The edges are rough and blurred, suggesting it may be filled with debris.

sem_id=305 UQ: 0.725 | before: 0.899 ⟶ after: 0.725



Fused caption:

A rectangular concrete block with a smooth, flat surface and slightly rounded edges. It appears weathered, showing patches of light gray with darker gray and brown discoloration, and has signs of wear.

sem_id=772 UQ: 0.759 | before: 0.897 ⟶ after: 0.759



Fused caption:

The object appears to be a giraffe or a horse with a pattern of dark brown patches outlined by lighter lines, standing with its long neck extended upwards and head turned slightly to the left. Its eyes are soft and gentle with long eyelashes, and its ossicones are short and tufted.



Across these qualitative examples, UQ-DAAAM consistently selects refined views that expose object evidence that is either missing or misleading in the initial views, leading to substantial reductions in uncertainty. In the first example (sem_id=262), the initial views yield conflicting descriptions such as an air vent, a "Farmhouse Market" signboard, and a slatted structure, while the refined views, captured from a closer and more frontal viewpoint, reveal the object clearly as a dark-framed window with a tinted glass pane, reducing uncertainty from 0.853 to 0.311. In the second and third examples (sem_id=182 and sem_id=1523), the original captions are highly unstable and include implausible object hypotheses such as a CCTV camera, a squirrel, a glowing lightbulb, and an air vent; the refined views instead provide more frontal and informative observations that reveal a dark gray cylindrical chimney and a rectangular window with partially open horizontal blinds, lowering uncertainty from 0.843 to 0.686 and from 0.836 to 0.719, respectively. The fourth example (sem_id=29) shows a harder case: the initial views describe the object as a manhole cover, a curved metallic object, or an irregular jagged surface pattern, while the refined view provides a sharper observation of a pothole-like structure, producing a moderate but still meaningful reduction from 0.906 to 0.747. The fifth example (sem_id=305) shows that unstable initial hypotheses of a trash can, a grid metal catch basin,

and a water valve are resolved by the refined views into a coherent description of a weathered rectangular concrete block, reducing uncertainty from 0.899 to 0.725. The sixth example (sem_id=772) illustrates how distance-induced ambiguity, where the initial small crops are described as a person with a lanyard, an impressionistic landscape painting, and a mannequin in a brown coat, is mitigated by refined views that zoom in on the object and reveal an animal sculpture; here uncertainty reduction does not force a single interpretation, as the model remains unsure whether it is a horse or a giraffe, and the fused caption conservatively preserves both possibilities (0.897 to 0.759). The seventh example (sem_id=417) highlights an important failure mode: although the refined view clearly improves the visual clarity of the object, which is actually a bike parking rack, the underlying VLM still misidentifies it as a tall phone booth with a sliding glass door, so the uncertainty only decreases modestly from 0.891 to 0.822. This case shows that refinement can expose a sharper observation while still being bottlenecked by the backbone VLM's recognition ability, and that lower uncertainty does not automatically imply semantic correctness. Finally, the eighth example (sem_id=679) demonstrates that refinement could avoid occlusion: initial captions conflict between a trash can, a ceramic vase, and a signboard due to the occlusion of the metal fences, while the refined views narrow the ambiguity between a dense bush and a tree canopy and reduce uncertainty from 0.914 to 0.750, with the fused caption conservatively preserving both possibilities. Overall, these examples show that UQ-DAAAM tends to select views with better geometry, scale, or visibility, which makes the resulting captions more semantically consistent and leads to measurable uncertainty reduction, while also exposing residual failures that are attributable to the underlying VLM rather than to the uncertainty-driven view selection itself.

R Limitations

UQ-DAAAM inherits several limitations from the underlying DAAAM pipeline while introducing additional ones of its own. First, our uncertainty estimate depends on the quality of the generated object captions and their associated coherence signals. If the VLM systematically hallucinates, misses fine-grained attributes, or produces poorly calibrated confidence estimates, the resulting uncertainty score may not faithfully reflect the true semantic ambiguity of the object. In particular, cross-view agreement does not always imply correctness, since multiple views may consistently support the same mistaken description. Second, the refinement stage introduces additional semantic-query cost. Although our method is designed to allocate this budget selectively, repeated querying and fusion may still be too expensive for highly dynamic platforms or settings with strict latency constraints. Finally, while caption fusion produces a more stable object summary, maintaining multi-view descriptions, uncertainty estimates, and refinement histories for many objects may increase memory overhead over long horizons. Future work should investigate stronger active-view acquisition strategies and memory-bounded summarization mechanisms for long-term deployment.

Temporal state changes. Although the underlying memory retains temporal information, our uncertainty score and acquisition objective contain no explicit temporal term. They cannot distinguish disagreement caused by a genuine state change, such as a door opening, from disagreement caused by viewpoint variation or captioning errors. Time-decayed evidence and change detection that separates successive object states are directions for future work.

Acquisition costs. Our refinement budget counts additional object–view queries and treats them as having equal cost. It does not model differences in inference latency, communication cost, or robot motion. The current procedure selects from available observations; extending it to physical viewpoint acquisition would require a cost-aware objective.

Joint refinement of multiple objects. View selection is performed independently for each object. Consequently, it does not explicitly exploit a view that could reduce uncertainty for several co-visible objects. A joint coverage objective could improve the use of a shared acquisition budget, but would need to balance object coverage against the visibility and detail available for each individual object.

Scope of the acquisition guarantee. The comparative guarantee assumes that view quality induces first-order stochastic dominance of the Schur-score distributions. This relationship may fail for particular objects, captioners, or environments. The empirical success of a view-selection heuristic does not by itself establish this distributional assumption.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: the main claims made in the abstract and introduction are justified by the experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: the paper discusses the limitations of the work in Appendix R.

   Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: the paper provides the full set of assumptions and complete proofs for all theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient information to reproduce the main experimental results, including the experimental setting and details, statistical significance, and compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper will provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides sufficient details on the experimental setting, including data splits, hyperparameters, and optimizer settings.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper provides error bars and statistical significance information for the experiments that support the main claims.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides detailed information on the compute resources used for each experiment, including the type of compute workers, memory, and execution time.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper adheres to the NeurIPS Code of Ethics in all aspects.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work is foundational research and does not have direct societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: The paper properly credits the creators of all assets used and respects the licenses and terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.