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

Harry Zhang* Nicolas Gorlo* Luca Carlone
Massachusetts Institute of Technology * Equal contribution

NeurIPS 2026

Motivation

- **Scene Graphs** can be combined with VLMs to form **persistent 4D robot memory** support complex downstream visual question answering tasks (*Gorlo et al., 2025*).
- VLM captions can be confidently wrong, yet persist as object descriptions in robot memory.
- An incorrect memory entry can affect later spatial and temporal questions.
- A limited query budget requires choosing which objects to refine and which views to inspect.

Contributions

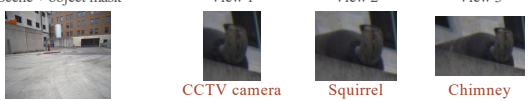
- A mathematical model that quantifies **per-object uncertainty** via semantic scatter across views and incoherence within each caption.
- Refine uncertain objects with additional views **under a small budget** and fuse the resulting descriptions.
- Derive a **quick, exact** online update and a conditional guarantee for quality-guided acquisition.
- Update the robot memory in **real time** (>=10hz).

Per-Object Uncertainty Quantification

Lower uncertainty: traffic sign (normalized UQ 0.513)



Higher uncertainty: chimney (normalized UQ 0.843)



$\mathbf{R}_j = [\,\mathbf{r}_{1j}\ \mathbf{r}_{2j}\ \mathbf{r}_{3j}\,]$: **one caption embedding per column**.

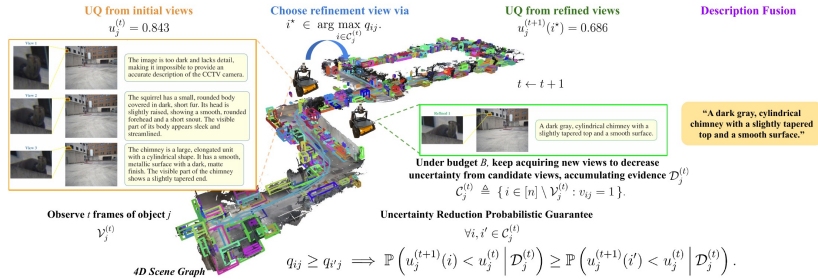
$\mathbf{r}_{ij} = \phi(c_{ij})/\|\phi(c_{ij})\|_2$, with caption c_{ij} and text encoder ϕ.

$\mathbf{\Gamma}_j = \mathrm{diag}(\gamma_{1j}, \gamma_{2j}, \gamma_{3j}), \quad \gamma_{ij} = e^{\alpha(1-p_{ij})}$.

$\mathbf{\Gamma}_j$ **rescales the columns** of $\mathbf{R}_j$. Here p_{ij} is mean token probability and $\alpha \geq 0$ controls weighting. Lower p_{ij} gives a larger incoherence weight.

$$u_j = \log\det(\mathbf{\Gamma}_j \mathbf{R}_j^\top \mathbf{R}_j \mathbf{\Gamma}_j + \lambda \mathbf{I}).$$

UQ-DAAAM: Approach Overview



Use three initial views to score each object. For uncertain entries, caption up to two unused high-quality views, recompute uncertainty, and fuse the captions. Stop when uncertainty crosses the threshold or the budget ends.

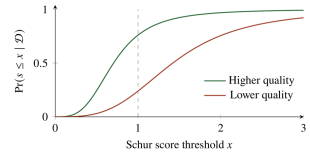
Uncertainty Reduction Theoretical Guarantee

1. q_{ij}: quality of object j in view i

Lower q_{ij} Higher q_{ij}



2. FOSD assumption (illustration)



FOSD assumes that higher quality shifts the Schur-score distribution toward smaller values.

3. Schur score s: multiplicative change in the scatter determinant

With object/time indices omitted, $\mathbf{Z}$ stacks the current weighted caption embeddings, $\mathbf{r}$ is the new unit embedding, γ its incoherence weight, and $\lambda > 0$ the regularizer. $\mathbf{S}_{\text{new}}$ is the scatter matrix after adding $\gamma\mathbf{r}$; $\mathcal{D}$ denotes current evidence.

$$\mathbf{S} = \mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{I}, \qquad \mathbf{b} = \mathbf{Z}^\top(\gamma\mathbf{r}), \qquad \boxed{s \equiv \gamma^2 + \lambda - \mathbf{b}^\top\mathbf{S}^{-1}\mathbf{b} = \frac{\det \mathbf{S}_{\text{new}}}{\det \mathbf{S}} > 0}.$$

$$\Delta u = \log s \implies \Delta u < 0 \iff s < 1$$

$$q_{ij} \geq q_{i'j} \implies \mathbb{P}\Big(\Delta u_j(i) < 0 \mid \mathcal{D}_j^{(t)}\Big) \geq \mathbb{P}\Big(\Delta u_j(i') < 0 \mid \mathcal{D}_j^{(t)}\Big)$$

Results on OC-NaVQA

Method	Accuracy ↑	Position ↓	Time ↓
DAAAM	0.711	41.75	1.792
UQ-DAAAM	**0.761**	**37.84**	**1.589**

All split. Position [m], time [min]. Same QA model.

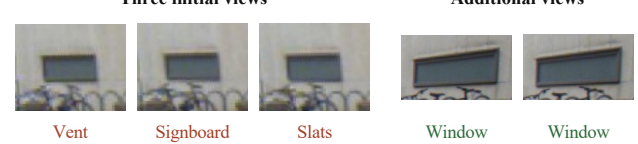
VQA comparison

"What's the color of the sculpture of the horse?"

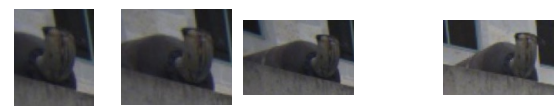
DAAAM
Incorrect: "I didn't see the horse."

UQ-DAAAM
Correct: "Brown."

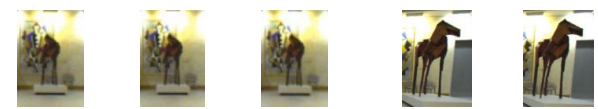
Additional Views Clarify Descriptions



Refined: a dark-framed window with tinted glass.
Additional views resolve the conflicting initial descriptions.



Refined: a dark gray cylindrical chimney with a tapered top.
One additional view clarifies the object's shape.



Refined: an animal sculpture, described as "giraffe or horse."
The fused caption preserves the remaining ambiguity.