We are grateful to all the reviewers for the insightful comments and suggestions. We are delighted that the **novelty** and **efficacy** of our method, as well as its potential to stimulate further research, have been acknowledged by all.

[R1] Technical writing. We are very sorry for any confusion our writing may have caused and are grateful to R1 for the suggestions. We read through the paper in its entirety and have identified areas where we will improve our writing (*e.g.*, deep auto-encoders (DAEs) with noise and the culling step in Sec. 3.2-3.3). We will clarify in the final version.

[R1] Quantization. We note quantization in our method is an application-specific design choice rather than a limitation. When compute power and memory allow, finer quantization can be used to obtain better localization accuracy (see comparison to ORB-SLAM and others in [11]). In our case, relatively coarse quantization is sufficient for scene synthesis, where the global scene representation is more crucial. We will conduct more experiments with different quantization levels (experiments running as we write this) and include results in the final version.

[R1] Evaluation setting & downstream benefits. We evaluated on navigation and exploration tasks using standard metrics (*e.g.*, trajectory and position errors as in MapNet). We thank R2 for the suggested maze experiment which we are currently conducting. In the meantime, results for a similar experiment, measuring the predicted global scene quality (SSIM) w.r.t. agent steps & fraction of scene observed, are in Figure I (b). Please note the improved SSIM over time.

[R1] Stochastic Hallucinations. Our method can be used with or without noise input depending on the application (*e.g.*, no noise for deterministic navigation planning or heavy noise for image dataset augmentation). Figure I (a) shows how our proposed model predicts different global properties for identical trajectories with different hallucinatory noise (with convergence as observations accumulate). We apologize for the confusion and will clarify the DAE architecture.

[R2] GTM-SM Comparison. We used a GTM-SM implementation available on GitHub and were able to reproduce results reported in the paper. We thank R2 for suggesting comparison with custom versions of GTM-SM with pose information. We show results in Table I. While the performance of these variants is much improved (when compared to those in our submission), we still observe the superiority of our method w.r.t. prediction of unseen regions.

[R2] Hallucination Benefits. R2 is right that hallucinations are more reliable when target scenes have learnable priors (*e.g.*, structure of faces). Hallucination of uncertain content can be of lower quality due to the trade-off between representing uncertainties w.r.t. missing content and unsure localization (giving blurred results), and synthesizing detailed (but likely incorrect) images. Soft registration and hallucinations' statistical nature can add "uncertainty", whereas our generative components partially compensate for them (*e.g.*, our choice of GAN to improve sampling). For data generation use-cases, relaxing hallucination constraints and scaling up $\mathcal{L}_{hallu}$ & $\mathcal{L}_{anam}$ can improve image detail, at the price of possible memory corruption (we focused on consistency rather than high-resolution hallucinations).

[R2] Larger Environments. One interesting direction for future work could be the use of pyramidal/multi-scale memory maps for refined registration/synthesis or for capturing larger scenes.

[R3] Feature Culling. We are very sorry for unclear exposition (some explanation is in supp. material). Inspired by *culling* in computer graphics, our process extracts features from memory to sample the view from a requested viewpoint, ignoring features (-1 in Eq. 3) outside the agent's field of view at that position. We will clarify in the final version.

[R3] Unclear Terms. We thank R3 for the suggestions and will include them in the final version (Fig. 4 caption, replacing "patch" & "view field", "83 × 83px", *etc.*). Please note t is defined L61 and L111 and s in L114 and L159.

Table I: Comparison w.r.t. $\boldsymbol{A^s_{cel}}$, editing GTM-SM to leverage ground-truth locations $\boldsymbol{l_t}$.

Methods	Average Position Error			Absolute Trajectory Error			Anam. Metr.		Hall. Metr.	
	Med. ↘	Mean ↘	Std. ↘	Med. ↘	Mean ↘	Std. ↘	L1 ↘	SSIM ↗	L1 ↘	SSIM ↗
GTM-SM trained with L1 loss between s_t and l_t	**1.0px**	1.03px	1.23px	0.79px	0.87px	0.86px	0.13	0.64	0.15	0.40
GTM-SM fed with l_t as s_t (no localization)	0px (NA – poses passed as inputs)			0px (NA – poses passed as inputs)			0.08	0.76	0.13	0.43
Ours	**1.0px**	**0.68px**	**1.02px**	**0.49px**	**0.60px**	**0.64px**	**0.06**	**0.80**	**0.09**	**0.72**

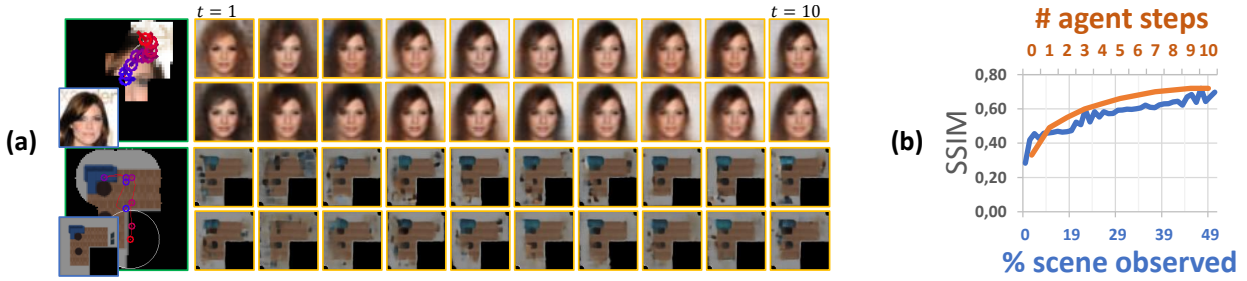

Figure I: (a) Global sampling for the same trajectories but different noise passed to the hallucinatory DAE at each step t; (b) SSIM of the global scene representation w.r.t. agent steps and scene observed for A^s_{cel}.