Captain Safari: A World Engine with Pose-Aligned 3D Memory

Supplementary Material

7. Additional Ablations on Memory Retrieval

To further validate the design of our pose-conditioned memory retrieval, we compare our cross-attention-based retrieval with a simpler *nearest-pose* baseline. In the nearest-pose variant, rather than aggregating pose-aligned world tokens via cross-attention, the model simply uses the memory feature m_τ corresponding to the past frame whose camera pose is closest to the query pose p_t.

As shown in Table 3, replacing our retrieval mechanism with nearest-pose memory degrades performance across all metrics compared to *Captain Safari*. More importantly, the nearest-pose variant even underperforms the DiT baseline (w/o *Mem.*). This is because a single nearest pose is often unaligned with the entire clip's viewpoint span under the aggressive 6-DoF motion of FPV drone flights, and this misalignment misguides the denoising process.

Table 3. **Additional ablation on memory retriever.** A single nearest pose misguides the denoising process under aggressive FPV motion, resulting in worse performance than the DiT baseline (w/o *Mem.*) and Captain Safari.

Ablation	Video Quality		3D Consistency		Trajectory Following	
	FVD↓	LPIPS↓	MEt3R↓	Recon.↑	AUC@30↑	CosSim↑
w/o $Mem.$	**998.47**	**0.504**	0.3720	0.912	0.193	0.508
Nearest-pose	1042.56	0.517	0.3810	0.836	0.185	0.492
Captain Safari	1023.46	0.512	**0.3690**	**0.968**	**0.200**	**0.563**

8. Complexity and Scalability Analysis

To address potential concerns regarding the computational overhead of retrieving and conditioning on the world memory, we provide a detailed complexity analysis here.

A core advantage of *Captain Safari* is its highly scalable design, which stems from *decoupling* the memory retrieval process from the iterative DiT denoising loop. During inference, we first process the local memory $\mathcal{M}_{\text{local}}$ and query it in a *single pass* to obtain the pose-aligned world tokens w_t. Because w_t has a *fixed size* (controlled by the number of learnable queries M), the subsequent denoising loop only cross-attends to this compact set of tokens.

Assuming generating a video sequence of length $\mathcal{T}$ requires S denoising steps. Standard self-attention in a DiT costs $\mathcal{O}(S\mathcal{T}^2)$. As shown in Table 4, our formulation adds a one-time retrieval cost of $\mathcal{O}(|\mathcal{M}_{\text{local}}|)$, plus a cross-attention term $\mathcal{O}(S\mathcal{T}|w_t|)$ in the denoising loop. Since $|w_t|$ is fixed and very small compared to the full history, the cost of scaling the memory length $|\mathcal{M}_{\text{local}}|$ only impacts the one-time retrieval. Consequently, our computational scaling with respect to memory length is *essentially constant* during the heavy iterative denoising phase. This contrasts sharply with standard cross-attention approaches that attend to the full memory at every denoising step and scale linearly, as well as early-concatenation approaches that scale quadratically.

Table 4. **Complexity comparison.** By decoupling memory retrieval and using fixed-size world tokens, our denoising overhead remains essentially constant with respect to $|\mathcal{M}_{\text{local}}|$.

Complexity $\mathcal{O}(\cdot)$	DiT Base	Captain Safari	Full Cross-Attn	Concatenation
Retrieving cost	N/A	$\lvert\mathcal{M}_{\text{lc}}\rvert$	N/A	N/A
Denoising cost	$S\mathcal{T}^2$	$S\mathcal{T}^2+S\mathcal{T}\lvert w_t\rvert$	$S\mathcal{T}^2+S\mathcal{T}\lvert\mathcal{M}_{\text{lc}}\rvert$	$S(\mathcal{T}+\lvert\mathcal{M}_{\text{lc}}\rvert)^2$
Scaling w.r.t $\mathcal{M}_{\text{lc}}$	N/A	$\approx$ **Constant**	Linear	Quadratic

9. Dynamic Scene Generation and Additional Qualitative Results

One concern regarding world engines is their ability to model the real world, which is inherently dynamic rather than purely static. While our implicit memory and trajectory formulation provide a strong geometric prior for stable backgrounds, our generator is built on top of powerful diffusion priors. Consequently, *Captain Safari* goes beyond rigid 3D rendering and remains highly capable of synthesizing dynamic content.

As demonstrated in our supplementary materials[3], the model can generate plausible moving elements such as moving vehicles and natural phenomena (e.g., ocean waves) with stable 3D environments.

For extensive additional qualitative results, including high-resolution video comparisons, and demonstrations of dynamic capabilities under aggressive 6-DoF motion, strong parallax, and complex outdoor layouts, please visit our project page[4] and the supplementary website[5].

10. Data Usage and Ethics

To comply with platform policies (e.g., AirVuz and YouTube) and respect copyright ownership, we do not redistribute the raw video files. Instead, *OpenSafari* is released as a comprehensive, open-source data curation pipeline. We provide the full suite of downloading, preprocessing, semantic filtering, and camera reconstruction scripts[6]. Researchers can use this pipeline to fetch the original URLs and automatically reproduce our carefully verified, geometry-annotated dataset for non-commercial, academic research purposes.

[3] johnson111788.github.io/captain-safari-supp/
[4] johnson111788.github.io/open-safari/
[5] johnson111788.github.io/captain-safari-supp/
[6] github.com/johnson111788/Captain-Safari/