Stream3D: Sequential Multi-View 3D Generation via Evidential Memory

Kaichen Zhou[2,3*] Zeyang Bai[1,2*] Xinhai Chang[2*]
Mengyu Wang[3†] Paul Pu Liang[2†] Fangneng Zhan[1†]

[1]World Mind Lab, HKUST [2]Media Lab and EECS, MIT
[3]Kempner Institute, Harvard University

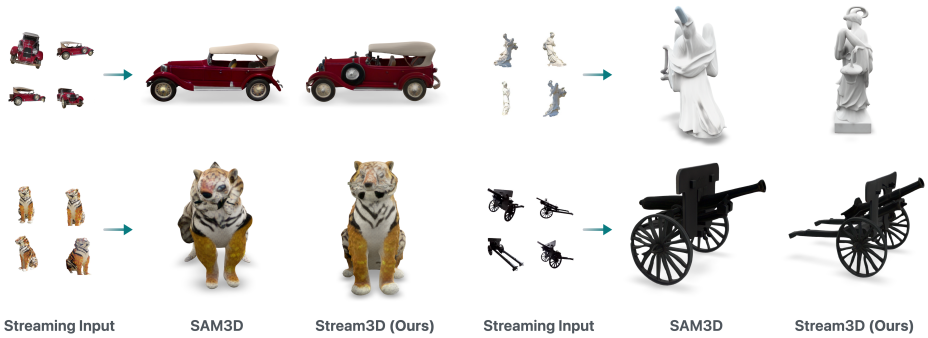


Fig. 1: Stream3D takes streaming input views as additional conditioning signals to improve the performance of pretrained single-view-conditioned 3D generation models without retraining pretrained weights. Compared with SAM-3D, incorporating views from the input stream can substantially improve 3D generation quality.

Abstract. View-conditioned 3D generators such as SAM 3D, TRELLIS and Hunyuan3D produce high-quality object 3D representations from a single view, but real-world visual observation often arrives as long monocular streams. Naively applying these generators to each streaming frame independently leads to severe temporal inconsistency in the generated results. To address this problem, we propose Stream3D, the first training-free streaming mechanism that turns a frozen view-conditioned 3D generator into a streaming generator with constant cross-chunk memory. Stream3D achieves this by maintaining a compact evidential memory, which selectively caches the most informative historical frames based on a proposed evidence score mechanism. As the stream progresses, the memory dynamically updates to retain a fixed number of informative frames, preventing the memory footprint from growing linearly with sequence length. This also prevents degradation over long sequences and keeps the underlying generator completely unchanged without retraining, architectural modifications, or auxiliary losses. Evaluated on both realistic and synthetic streaming benchmarks, Stream3D outperforms latent-transport baselines, including KV-cache reuse and flow-based feature editing, across

*Equal contribution as first authors. †Joint supervision.

both photometric and geometric metrics. More details can be found at: Link.

Keywords: Streaming 3D Generation · Evidential Memory · View-Conditioned Generation

1 Introduction

Object-centric 3D generation is becoming a practical building block for vision and robotics [35, 41, 63]. Recent systems such as SAM 3D Objects [9] and TRELLIS.2 [81] can take an image and reconstruct corresponding Gaussian splat [26] or 3D mesh within seconds. However real capture devices produce long monocular streams: a phone circling an object [49] or a robot observing a scene while moving [21]. Naively applying a single-view 3D generator frame by frame yields temporally inconsistent reconstructions and exploits only partial observations [9, 81]. Feeding all frames jointly through multi-diffusion [4] or multi-view fusion [34] is computationally expensive and can become infeasible for long streams, while processing fixed-size chunks with flow-matching-based editing [31] discards historical context needed for global consistency.

An alternative approach is to adapt techniques from streaming video generation or 3D reconstruction [19, 27, 32, 94]. For instance, state-transport mechanisms like KV banks [32] or FlowEdit-style velocity edits [31] could be introduced to propagate information across chunks. However, transporting latent states in this manner is also problematic. As the orientation, shape, and scale of these states are deeply entangled within the generator, they are difficult to align across frames, leading to severe error accumulation during the streaming process [34]. Furthermore, as these methods transport and accumulate state over time, their memory footprint naturally grows with sequence length [10]. Such state propagation may also introduce compounding errors over long streams, echoing the error accumulation and memory bottlenecks observed in autoregressive video diffusion models [71]. Thus, the mechanism intended to preserve consistency can become a source of degradation when the transported state is misaligned or stale.

To solve this problem, we propose **Stream3D**, a training-free streaming mechanism that turns a frozen view-conditioned 3D generator into a long-stream generator without modifying its weights or architecture. Rather than transporting latent states across chunks, Stream3D retains *evidential views*: observed frames that the generator itself identifies as reliable conditioning signals through its cross-attention maps. Specifically, during a lightweight warmup pass, we compute an evidence score for each query token and incoming views, measuring whether the token attends to that views both strongly and selectively. These scores update a token-level *Adaptive Evidential Memory*, which stores only a fixed number of high-evidence frame indices per query token, keeping memory constant with stream length. This token-level design allows different spatial regions of the 3D volume to retrieve different historical observations, rather than relying on a single global frame subset or an accumulated latent state. As new views arrive,

the retained evidence for each query token can only be maintained or improved, yielding a simple non-degradation property in evidence space. For generation, token-level evidence is aggregated into frame-level ownership scores, and the top-K frames are selected as a bounded conditioning bundle for *Evidence-Based Multi-Generation*. In this way, Stream3D avoids latent-space alignment, preserves long-range visual evidence, and enables coherent streaming 3D generation while leaving the original generator unchanged. We evaluate Stream3D as a lightweight mechanism mechanis around pre-trained SAM 3D on long monocular streams as shown in Fig. 1. Our contributions are fourfold:

- **Streaming 3D generation.** We pioneer the task of extending frozen view-conditioned 3D generators to long monocular streams, producing temporally consistent 3D generations while keeping memory bounded and avoiding retraining.
- **Adaptive Evidential Memory.** We introduce a compact memory mechanism that stores token-level evidential views rather than transporting latent states. The memory is constant and training-free, with a footprint that remains constant with stream length.
- **Evidence-Based Multi-Generation.** We propose an evidence-guided generation strategy that aggregates token-level memory into a bounded conditioning bundle and runs the frozen generator on the selected views. This enables different spatial regions to draw from the most reliable historical observations without modifying the underlying generator.
- **Superior streaming performance.** We show that Stream3D outperforms latent-transport and fixed-view baselines across photometric and geometric metrics, while avoiding long-horizon degradation and maintaining a constant memory footprint.

2 Related Work

2.1 3D Generation

Recent object-centric 3D generators can be understood along three main design axes: *output representation*, *learning regime*, and *input format*. Along the first axis, different methods adopt different 3D representations, including triplane- and Neural Radiance Field (NeRF)-based backbones [1,6,20,35,49,66,72,87], surface meshes [18,41,65,76–78,82,83], 3D Gaussian splats [22,26,63,64,84,86], and native structured latent volumes [9,37,38,79,81,89]. Along the second axis, existing approaches cover closed-form regressors [1,20,22,35,41,63,66,72,76–78,81,83,84], 3D latent diffusion models [9,25,37,38,53,79,89], and SDS-based optimization methods [7,39,48,54–56,62,64,65,68,75,86]. Some layout-aware variants [2,9,34] further model the spatial arrangement of multiple objects within a scene. The third axis, *input format*, is the one most relevant to our work. Despite their differences in representation and learning paradigm, most existing methods assume a fixed and limited input interface: they take either a single image or a small, predefined set of images as input. Multi-view diffusion methods [29,36,42,43,45,58,59,67]

relax this setting by first hallucinating additional views before reconstructing 3D geometry. Video-conditioned reconstruction models [12, 33, 70, 74, 85, 93] and 4D-aware generators [3, 11, 24, 57, 60, 88, 90, 92] extend the input format further to short temporal clips. More recently, MV-SAM3D [34] enables multi-view fusion directly in the latent space of a 3D generator. Nevertheless, these methods still operate on a fixed input bundle determined in advance, rather than supporting truly open-ended streaming inputs. None of these methods natively supports unbounded online streams. In practice, long-stream processing is typically handled through ad hoc latent-transport schemes, which propagate intermediate states across chunks. However, such designs usually incur memory growth with sequence length and lead to performance degradation over long horizons. In this work, we adapt a frozen view-conditioned generator [9] into a streaming generator that maintains constant cross-chunk memory, independent of stream length, and agnostic to the model backbone.

2.2 Reconstruction from Streaming Inputs

Reconstructing 3D geometry from streaming inputs has traditionally been studied in monocular SLAM [5, 13, 15–17, 21, 28, 47, 50–52, 61, 91], which incrementally estimates camera motion and scene structure from video. Recent methods have extended modern feed-forward reconstruction models to online settings. Notably, Spann3R [69] augments a DUSt3R-style encoder [74] with a token-addressable spatial memory, enabling online pointmap fusion over long image streams. Similarly, SLAM3R [44], also built upon DUSt3R, introduces a real-time end-to-end dense reconstruction system that directly predicts 3D pointmaps from RGB videos. Point3R [80] further incorporates an explicit geometry-aligned spatial pointer memory, together with 3D hierarchical RoPE and an adaptive fusion mechanism. However, DUSt3R itself remains inherently two-view, restricting each inference step to a fixed image pair and making large-scale fusion dependent on iterative matching and optimization. VGGT-SLAM [46, 70] addresses this limitation by adopting the more powerful VGGT transformer, which supports image sets of arbitrary length. CUT3R [73] instead adopts an RNN-style formulation for causal pointmap prediction from unstructured image streams. However, it compresses all past observations into a limited recurrent state, which can hinder long-range memorization and fine-grained multi-view fusion. Following the design philosophy of modern large language models, StreamVGGT [94] and Stream3R [32] employ causal transformers to implicitly cache historical visual tokens. In addition, because CUT3R suffers from severe drift on long streaming inputs, TTT3R [8] proposes a simple empirical state-update rule to improve sequence-length generalization. Stream3D departs from these reconstruction-centered approaches: while online reconstruction focuses on aggregating geometry already observed in the input stream, streaming 3D generation must also infer and synthesize unseen structure under temporal and geometric consistency constraints.

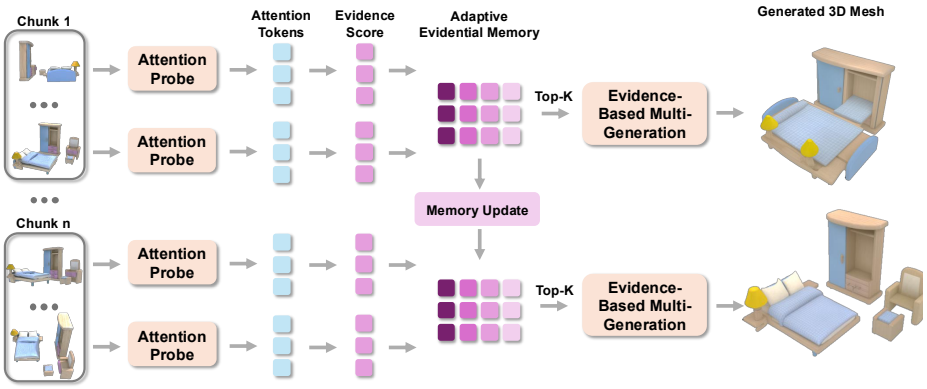


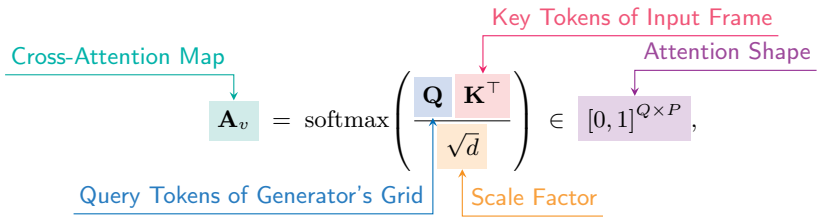
Fig. 2: Framework of Stream3D. Given a streaming video, Stream3D processes frames chunk by chunk. A lightweight warmup pass extracts token-wise evidence score from cross-attention, which is stored to vote for informative frames to update the evidential memory, i.e., **Adaptive Evidential memory**. Then, the top-K informative frames are passed to the **Evidence-Based Multi-Generation** for 3D asset generation. By retaining only compact evidential memory rather than latent states, Stream3D achieves stable long-horizon generation with a constant memory footprint.

3 Method

We extend view-conditioned 3D generators, SAM 3D Objects [9] and TRELLIS.2 [81] — to handle long streaming inputs without retraining shown in Fig. 2. The *Adaptive Evidential Memory* (Sec. 3.2) enables efficient long-range memory by retaining token-level evidence from past views, while *Evidence-Based Multi-Generation* (Sec. 3.3) leverages this memory to produce temporally consistent and geometrically coherent 3D generations throughout the streaming process.

3.1 Problem Setup

3D Generation Preliminary. Let f_θ denote a frozen view-conditioned 3D generator that, given a single input frame v and an initial noise prior $z_0 \sim \mathcal{N}(0, I)$, produces a 3D sample (Gaussian splat [26], mesh, or latent volume) $\hat{y} = f_\theta(v; z_0)$. Recent 3D generation models, i.e., SAM 3D [9], TRELLIS [82], TRELLIS.2 [81], Hunyuan3D 2.0 [89] and CraftsMan3D [37], instantiate f_θ as a two-stage pipeline — a structure stage (SS) producing a coarse occupancy / latent grid, followed by a texture or appearance stage (SLAT) — with at least one cross-attention layer of the form:



(1)

where Q is the number of query tokens in the generator's voxel grid (e.g., $16^3 = 4096$ in SAM 3D's structure stage), P is the number of key tokens in the input frame v, and d is the feature dimension of each query/key token. In its native form, 3D generator f_θ accepts a single frame and produces a single reconstruction; it has no mechanism for incorporating multi-view evidence.

Streaming 3D Generation. Our method extends 3D generator f_θ to accept a stream of condition views $\mathcal{V} = \{v_1, \ldots, v_T\}$ by fusing the per-view forward passes through a confidence-weighted Multi-Diffusion-style aggregation in 3D latent space. The stream is partitioned into overlapping chunks $\mathcal{C}_k$ of size C. At chunk k, we wish to produce a 3D sample $\hat{y}_k$ that is consistent with all previous chunks, while maintaining a memory that does not scale linearly with stream length T.

3.2 Adaptive Evidential Memory

To maintain an efficient memory footprint that does not scale linearly with the total sequence length T while preserving generation quality over long streams, we introduce an *Adaptive Evidential Memory* mechanism, shown on the left of Fig. 3. The right side of Fig. 3 shows that Stream3D improves as more frames are observed and then reaches a stable performance plateau. Rather than blindly retaining all historical frames, this approach dynamically filters and preserves only the most informative views by evaluating their relevance at a per-token level. Our mechanism operates in three distinct stages. (1) First, we compute a token-wise **Evidence Score** via a lightweight attention probe to measure the significance of each incoming view. (2) Second, we update a fixed-capacity global **Memory** that persistently tracks the highest-scoring frames for each individual query token. (3) Finally, we perform **Conditioning-view selection** by allowing the tokens to "vote" for their preferred frames, aggregating these local preferences to select an optimal, bounded-size condition set for the full generation pass.

Evidence Score. At each chunk, we run a small number of flow/denoising steps of the structure-stage generator on the chunk's conditioning set, using a *frozen* prior z_0 sampled once at chunk 0 and reused for all subsequent chunks. From a fixed cross-attention layer L, we extract the cross-attention map in Eq. equation 1. For each non-special query token q and each view v, we compute a per-token evidence score $\mathbf{M}_v[q]$ that measures how informative view v is for reconstructing token q.

The evidence score should capture two complementary properties: the total amount of attention assigned to view v and how spatially concentrated that attention is within the view. We therefore define a normalized attention distribution over the P patch tokens of view v:

$$\mathbf{H}_v[q] = -\frac{1}{\log(P)} \sum_{i=1}^{P} \tilde{\mathbf{A}}_v[q,i] \log \tilde{\mathbf{A}}_v[q,i], \qquad \tilde{\mathbf{A}}_v[q,i] = \frac{\mathbf{A}_v[q,i]}{\sum_{j=1}^{P} \mathbf{A}_v[q,j]}. \quad (2)$$

Here, $\mathbf{H}_v[q] \in [0,1]$ is the normalized entropy of the attention distribution, where lower entropy indicates that the query token attends to a more localized set of

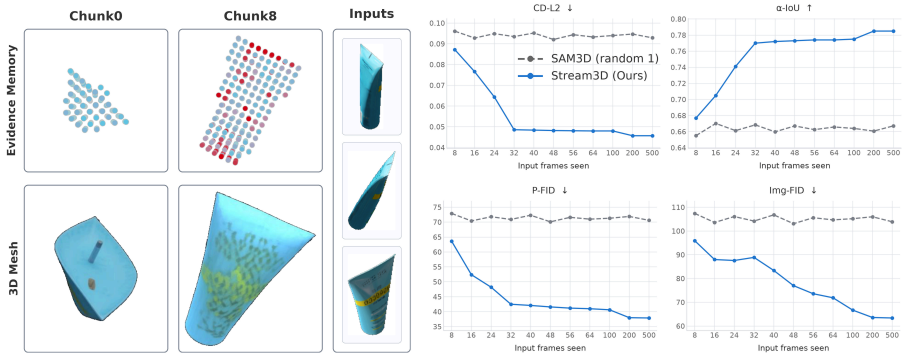


Fig. 3: Adaptive Evidential Memory. Left: Given a streaming sequence, our method updates a compact token-level memory by retaining historical views with the strongest evidence for each query token. The color transition from blue to red denotes increasing frame indices, from earlier to later observations. Over time, the memory incorporates evidence from increasingly diverse viewpoints, and the reconstruction improves from early to later chunks, consistent with the long-horizon geometry and appearance curves. Right: Compared with SAM3D, Stream3D consistently improves performance as the number of observed chunks increases, and then reaches a stable plateau.

patches. We then define the evidence score as

$$\mathbf{M}_v[q] = \left(1 + \beta \left(\sum_{i=1}^{P} \mathbf{A}_v[q,i] - \frac{1}{N_c}\sum_{u=1}^{N_c}\sum_{i=1}^{P} \mathbf{A}_u[q,i]\right)\right) \cdot \left(1 - \mathbf{H}_v[q]\right), \quad (3)$$

where N_c is the number of views in the current chunk and $\beta = 8$ is a scaling hyperparameter. The first factor measures the relative cross-attention mass assigned to view v compared with the chunk average, while the second factor rewards spatially concentrated attention.

This design follows a scale–concentration decomposition. A view provides strong evidence for token q only when it receives above-average attention mass and that attention is focused on a compact, spatially specific set of patches. In contrast, high-entropy attention spread over many patches is treated as weak evidence, even if the total attention mass is large, since it may correspond to diffuse background response or non-specific visual context. Reusing the frozen prior z_0 is also intentional. It makes evidence scores comparable across chunks: changes in the attention maps are driven by differences in the conditioning views rather than by a fresh noise sample at each chunk.

Memory. The cross-chunk memory is represented by two matrices $\mathbf{M}, \mathbf{F} \in Q \times D$, where Q is the number of query tokens, and D is the number of frames to retain, i.e., memory depth. $\mathbf{M} \in \mathbb{R}^{Q\times D}$ is the *evidence memory*, which stores the D token-wise highest evidence scores over all observed views; $\mathbf{F} \in \mathbb{R}^{Q\times D}$ is the *frame-index memory*, which stores their corresponding global frame indices. During the streaming process, each token's top-D list is updated by merging in the new candidates from newly arriving frames. Frames that never enter any token's list are discarded immediately.

Conditioning-view Selection. At a certain chunk, the runtime view set $\mathcal{V}^\star$ selected for the full forward pass is obtained by aggregating the per-token top-D lists into a per-frame *token-ownership count*, then taking the **top-K** frames by that count. Concretely, we define the ownership count of frame f_r as the number of (token, rank) slots it occupies anywhere in $\mathbf{F}$:

$$n_{f_r} \;=\; \big|\{(q,j) : \mathbf{F}[q,j] = f_r,\; 1 \le q \le Q,\; 1 \le j \le D\}\big|, \tag{4}$$

and select the K frames with the largest counts for $\mathcal{V}^\star$. Notably, there are two distinct ranks D and K. D is the *memory depth*: how many candidate frames each token retains in its sorted list; K is the *bundle size*: how many frames the downstream generator consumes per forward pass. Here, D controls how robust the evidence score computation is.

3.3 Evidence-Based Multi-Generation

Generation via multi-view flow-matching fusion. The selected views $\mathcal{V}^\star$ are passed to the generator f_θ, which performs multi-view-conditioned diffusion: at each step t, the generator computes a velocity $V_\theta(z_t, v)$ for each $v \in \mathcal{V}^\star$ and fuses the per-view scores into a single update on the shared latent z_t. Following the standard multi-diffusion fusion rule, the fused velocity at each query token q is a view-weighted average:

$$\bar{V}_\theta(z_t)[q] \;=\; \sum_{v \in \mathcal{V}^\star} \overline{M}_v[q]\, V_\theta(z_t, v)[q], \qquad \sum_{v \in \mathcal{V}^\star} \overline{M}_v[q] = 1, \tag{5}$$

where $\overline{M}_v[q]$ is the normalized per-token, per-view evidence score. The full forward pass is $\hat{y} = f_\theta(\mathcal{V}^\star; z_k)$ with z_k drawn freely per chunk.

Algorithm and properties. Algorithm 1 summarizes the per-chunk procedure. The method involves three integer hyperparameters (L, D, K), a fixed prior z_0, and no learned parameters. Processing each chunk requires a single warmup forward pass to extract cross-attention at layer L, followed by memory updates, and a full forward pass of f_θ using the selected conditioning frame bundle.

Summary. The Adaptive Evidential Memory has two structural properties that clearly distinguish it from latent-transport schemes. First, its cross-chunk memory footprint is bounded by at most $2 \times Q \times D$ scalars, and is therefore constant with respect to the stream length T. Second, for every query token q and rank j, the cached evidence score $\mathbf{M}[q,j]$ is monotonically non-decreasing over time, since each update retains only the top-D entries from the union of the previous cache and the newly arrived candidates. As a result, the conditioning bundle selected for 3D generation can only stay the same or improve in terms of evidence quality as streaming progresses, while the memory cost remains fixed.

4 Experiment

We evaluate Stream3D on long-stream 3D generation, where the model receives a sequence of continuously arriving posed images and must maintain a coherent

Algorithm 1 Streaming inference with Adaptive Evidential Memory.

Require: frozen generator f_θ, stream chunks $\{\mathcal{C}_k\}_{k\geq 0}$, attention layer L, query-token range $[a, b]$, memory depth D, bundle size K.

Sample $z_0 \sim \mathcal{N}(0, I)$ ▷ frozen prior for evidence probing
Initialize evidence memory $\mathcal{M} \leftarrow \mathbf{0}^{Q\times D}$
Initialize frame-index memory $\mathcal{F} \leftarrow \mathbf{0}^{Q\times D}$
for $k = 0, 1, 2, \ldots$ **do**
    Receive current chunk $\mathcal{C}_k$
    Run warmup flow/denoising steps of f_θ on $\mathcal{C}_k$ using the frozen prior z_0
    Extract cross-attention maps $\{\mathbf{A}_v\}_{v\in\mathcal{C}_k}$ from layer L via Eq. 1
    Compute per-view evidence scores $\mathbf{m}_v \in [0, 1]^Q$ over patch tokens for each $v \in \mathcal{C}_k$
    Update $(\mathcal{M}, \mathcal{F})$ by row-wise top-D merging over query tokens
    Compute frame ownership counts $\{n_f\}$ from $\mathcal{F}$ via Eq. 4
    Select the conditioning bundle $\mathcal{V}_k^\star$ from the top-K frames
    Sample a fresh generation prior $z_k \sim \mathcal{N}(0, I)$
    Generate $\hat{y}_k \leftarrow f_\theta(\mathcal{V}_k^\star; z_k)$ with evidence-weighted fusion in Eq. 5
    return $\hat{y}_k$
end for

3D representation over time. Unlike standard multi-view reconstruction, this setting stresses two properties simultaneously: the method must exploit newly observed views to improve geometry and appearance, while preserving long-range consistency without reprocessing the entire stream. Our experiments are designed to answer three questions: (i) whether Stream3D improves streaming 3D generation quality over single-view and multi-view baselines, (ii) whether token-wise evidential memory provides better long-range consistency than existing streaming alternatives, and (iii) how memory size and view-selection strategy affect performance.

4.1 Experimental Setup

Implementation: Our experiments are run on NVIDIA H100 GPU. We use SAM3D with its standard settings as the underlying generation backbone. We adopt Depth Anything 3 [40] to estimate the camera pose and depth of initial input to get a point map, which are fed into SAM 3D for 3D generation. We set $K = 8$ and $D = 5$ for efficiency, and provide ablation studies to justify this choice. In all experiments, we evaluate multi-view generation models using streams of length 100.

Dataset: We evaluate on two complementary benchmarks. The first is the GSO benchmark [14], which contains high-quality scanned objects with ground-truth 3D assets. Following prior multi-view generation and reconstruction protocols, we render posed image streams from each object and evaluate both novel-view appearance and geometric accuracy. This controlled setting allows us to measure whether the generated 3D content remains faithful to the underlying object as the stream length increases. The second benchmark is NAVI [23], which provides

Fig. 4: Qualitative results on GSO and NAVI. Stream3D produces more consistent and geometrically faithful 3D generations than single-view and multi-view diffusion baselines.

more complex object-centric view sequences with diverse camera trajectories and challenging viewpoint variation.

Baseline: We compare Stream3D with representative single-view, multi-view, and streaming baselines. Single-view baselines, including SAM3D, and TRELLIS, generate 3D content from individual observations and therefore provide a lower bound on streaming consistency. EscherNet [30] serves as a strong multi-view baseline that benefits from multiple posed observations but is not designed for unbounded streaming inputs. StreamVGGT [95] serves as a reconstruction baseline with streaming input. We also compare against TRELLIS+M.D. [82] and TRELLIS.2+M.D. [81] which aggregate multiple views through full-context or multi-window diffusion but becomes expensive as the number of frames grows.

Metrics. We report appearance and geometry metrics. For appearance, we use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS) and Fréchet Inception Distance (FID) on held-out novel views. These metrics evaluate whether the generated representation renders images that are both photometrically accurate and perceptually faithful. For geometry, we report Patch Fréchet Inception Distance (PFID), Chamfer Distance (CD) and Intersection over Union (IOU). These metrics measure absolute depth quality, relative geometric consistency, and fine-grained 3D accuracy.

4.2 Main Results

Table 1 reports results on GSO and NAVI, and Fig. 4 shows qualitative results on GSO. Across both datasets, Stream3D achieves the strongest overall performance on appearance and geometry metrics. On GSO, Stream3D(SAM3D) obtains the best CD, IoU, P-FID, PSNR, SSIM, and LPIPS, while also remaining competitive on Image FID. On NAVI, Stream3D achieves the best CD and Image FID, with

Table 1: Main results on GSO and NAVI. Stream3D consistently improves geometry and appearance over single-view and multi-view generation baselines.

Data	Method	Geometry			Appearance			
		CD↓	IOU↑	PFID↓	PSNR↑	SSIM↑	LPIPS↓	Image FID↓
GSO	StreamVGGT	0.064	0.549	75.520	10.800	0.787	0.258	189.180
	EscherNet + NeuS	0.063	0.664	49.570	14.790	0.840	0.151	120.990
	TRELLIS.2	0.146	0.433	140.513	12.281	0.830	0.201	141.006
	TRELLIS+M.D.	0.053	0.750	44.515	15.933	0.854	0.149	**60.091**
	TRELLIS.2+M.D.	0.093	0.648	87.720	14.082	0.839	0.176	98.255
	SAM3D	0.094	0.664	71.263	14.178	0.848	0.178	105.197
	Stream3D(TRELLIS.2)	0.087	0.692	77.730	14.213	0.845	0.169	83.279
	Stream3D(SAM3D)	**0.048**	**0.775**	**40.641**	**16.145**	**0.866**	**0.139**	66.711
NAVI	TRELLIS.2	0.071	0.818	42.045	20.362	0.934	0.075	88.067
	TRELLIS+M.D.	0.055	0.810	44.297	20.426	0.935	0.078	87.654
	TRELLIS.2+M.D.	0.066	0.791	50.839	19.914	0.934	0.078	89.755
	SAM3D	0.068	0.799	47.989	19.445	0.930	0.086	89.593
	Stream3D(TRELLIS.2)	0.045	**0.833**	**40.431**	**20.880**	0.936	**0.072**	85.546
	Stream3D(SAM3D)	**0.043**	0.825	42.151	20.720	**0.939**	0.073	**81.173**

Table 2: Ablation studies of different streaming strategies.

Data	Method	Geometry			Appearance			
		CD↓	IOU↑	PFID↓	PSNR↑	SSIM↑	LPIPS↓	Image FID↓
GSO	SAM3D + FlowEdit	0.090	0.668	76.445	14.343	0.850	0.178	98.643
	SAM3D + KV-Cache	0.084	0.682	67.292	14.482	0.852	0.171	83.353
	MV-SAM3D + Last Chunk	0.113	0.630	86.044	13.906	0.848	0.184	104.228
	MV-SAM3D with K random views	0.064	0.676	68.534	14.828	0.859	0.156	83.039
	MV-SAM3D + Visibility	0.063	0.716	58.881	*15.05*	*0.855*	*0.163*	81.219
	MV-SAM3D + VLM	0.059	0.710	47.990	*15.23*	*0.855*	*0.157*	78.530
	MV-SAM3D + Stream3D	0.053	0.760	47.260	15.892	0.864	0.1449	77.750
	Ours (Fast)	0.051	0.759	43.940	15.510	0.861	0.152	80.610
	Ours	**0.048**	**0.775**	**40.641**	**16.145**	**0.866**	**0.139**	**66.711**

consistently strong appearance scores. These results indicate that the improvement is not limited to rendering quality but also reflects better 3D structure. We draw two conclusions. (1) Compared with single-view baselines such as SAM3D, TRELLIS.2, and TRELLIS.2, Stream3D substantially improves both appearance and geometry. This confirms that streaming observations provide useful information that cannot be recovered from a single image alone. Single-view methods often generate plausible visible regions but hallucinate unobserved geometry inconsistently. (2) Compared with multi-view baselines such as TRELLIS+M.D. and TRELLIS.2+M.D., Stream3D shows stronger long-stream behavior. Multi-view diffusion improves consistency with a bounded view set, but does not provide an explicit mechanism for compact long-range memory. Stream3D addresses this by retaining token-level evidence and selecting a bounded conditioning bundle, maintaining high reconstruction quality while keeping memory usage constant.

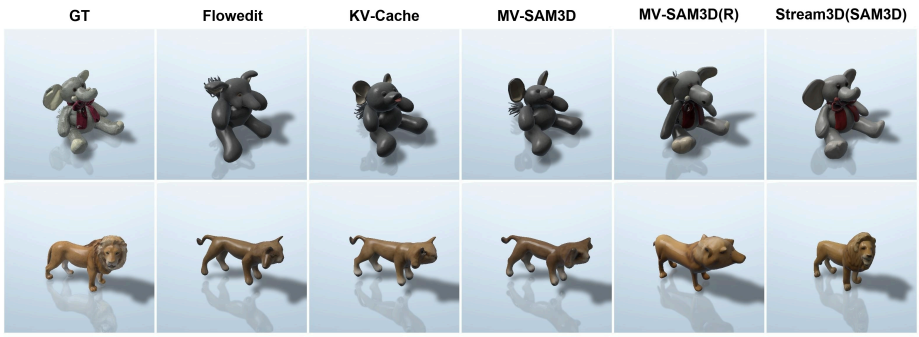


Fig. 5: Qualitative result of our Ablation studies. FlowEdit denotes SAM3D with FlowEdit, and KV-Cache denotes SAM3D with KV-cache reuse. MV-SAM3D denotes MV-SAM3D applied to the last input chunk, while MV-SAM3D(R) denotes MV-SAM3D with K randomly selected views.

4.3 Streaming Baseline Comparison

Table 2 compares Stream3D with several streaming alternatives, and Fig. 5 reports qualitative results. (1) We also compare against cache- and transport-based streaming baselines. KV-cache reuse efficiently carries information across chunks, but cached states are not explicitly filtered by geometric reliability and can accumulate stale or ambiguous evidence. FlowEdit improves short-range consistency through local latent editing, but it operates on fixed-size chunks and loses long-range observations outside the editing window. In contrast, Stream3D does not transport all past latent states. It keeps a compact persistent memory and updates it according to token-level evidence. (2) We first compare against MV-SAM3D-style fixed-view selection to test whether a compact set of representative frames can replace streaming memory. Random view sampling is unstable because it may discard observations that are critical for reconstructing specific regions. Visibility- and VLM-based selection improve over random sampling, but they still operate at the frame level. This reveals the limitation of global view selection: different spatial tokens may require evidence from different historical views. In contrast, Stream3D maintains token-level evidential memory, allowing each spatial region to retain the observations that best support its generation. (3) Finally, MV-SAM3D+Stream3D uses our selected views with MV-SAM3D fusion. Its improvement over other MV-SAM3D variants shows that evidential selection is useful by itself, while the remaining gap to Stream3D shows that evidence-weighted fusion also matters. (4) To further demonstrate the generalizability of Stream3D, we also implement it on the shortcut version of SAM3D, denoted as Ours (Fast). The results show that even with the shortcut sampler, Ours (Fast) still achieves strong performance. Overall, Stream3D provides a more flexible and scalable mechanism for long-stream generation, preserving global consistency without storing or jointly fusing all past frames.

4.4 Ablation Studies

We conduct ablation studies in Tab. 3 to analyze the effects of evidence stage, evidence normalization, memory depth, and TopK view selection. First, we study which generation stage should be used for evidence computation. Following [9], we denote the first generation stage as SS and the second as SLAT. Deeper stages produce more reliable geometry-aware evidence: SS stabilizes around stage 9 and SLAT around stage 6. Among the tested combinations, SS9-SLAT6 achieves strong geometry and appearance performance, indicating that later transformer features provide better multi-view correspondence for streaming view selection. Second, we compare evidence scoring variants. Our normalized Evidence score outperforms entropy-only scoring (ENT) and the unnormalized variant Evidence(N). ENT provides weaker geometric alignment, while Evidence(N) is less stable and leads to higher perceptual errors. This shows that normalization is important for balancing confidence across candidate views and avoiding dominance by uncalibrated attention magnitudes. Third, we evaluate memory depth D. Increasing D from 1 to 3 or 5 improves long-range consistency by retaining multiple high-evidence historical views. However, further increasing D to 7 does not consistently help, likely because weaker or stale evidence can enter the memory. Overall, $D = 5$ provides a stable trade-off between reconstruction quality, appearance, and memory cost. Finally, we analyze TopK view selection. Using too few views ($K = 4$) reduces multi-view coverage and degrades performance. Once enough informative views are selected, performance becomes relatively stable: increasing K from 8 to 16 yields only moderate gains. We therefore use $K = 8$ as a practical default, balancing quality and efficiency.

4.5 Efficiency and Scalability

Stream3D adds one warmup probe per chunk to extract cross-attention evidence. As shown in Tab. 4, this probe introduces a fixed overhead, but the dominant cost remains the frozen multi-view generation backbone. Under the same view budget, the generation time of Stream3D is close to MV-SAM3D, indicating that the evidential-memory update does not change the main computational bottleneck. The persistent memory itself is lightweight. Tab. 5 shows that it stores only the top-D evidence scores and frame indices for each query token, so its size scales as $O(QD)$ rather than $O(T)$. In practice, reconstructing the selected bundle also requires the RGB frames and metadata, such as poses and depths, for frames currently referenced by memory. However, unreferenced frames can be discarded in an online setting, so the persistent cache remains bounded by the retained evidence rather than the full stream length.

Table 5: Evidence memory footprint.

D	Evidence memory	Shape
1	96 KB	4096×2
2	192 KB	4096×4
3	288 KB	4096×6
4	384 KB	4096×8
5	480 KB	4096×10

Table 3: Ablation studies on module selection and hyperparameters. "Stage" denotes the step at which evidence scores are computed. "E" denotes the evidence computation strategy. "D" denotes the memory depth. "K" denotes the number of selected Top-K views. Our setting is highlighted.

Stage	E	D	K	Geometry			Appearance			
				CD-L2↓	α-IoU↑	P-FID↓	PSNR↑	SSIM↑	LPIPS↓	Img-FID↓
SS6-SLAT3	Evi	5	8	0.057	0.744	48.62	15.577	0.859	0.1531	72.16
SS6-SLAT6	Evi	5	8	0.057	0.742	49.09	15.536	0.858	0.1528	70.97
SS9-SLAT3	Evi	5	8	0.048	0.775	41.36	16.136	0.866	0.1406	68.28
SS9-SLAT6	Evi	5	8	0.048	0.775	40.64	16.145	0.866	0.1396	66.71
SS9-SLAT9	Evi	5	8	0.048	0.776	41.05	16.119	0.865	0.1403	67.88
SS9-SLAT6	Evi(U)	5	8	0.056	0.749	44.98	15.730	0.863	0.1486	74.08
SS9-SLAT6	ENT	5	8	0.053	0.760	47.26	15.892	0.864	0.1449	67.75
SS9-SLAT6	Evi	1	8	0.051	0.769	41.15	15.952	0.866	0.1406	67.93
SS9-SLAT6	Evi	3	8	0.049	0.773	**39.47**	16.067	0.866	0.1415	66.45
SS9-SLAT6	Evi	7	8	0.049	0.772	40.19	16.040	0.863	0.1425	67.50
SS9-SLAT6	Evi	5	4	0.055	0.755	45.47	15.741	0.863	0.1478	70.20
SS9-SLAT6	Evi	5	12	0.048	0.775	40.12	16.167	0.865	0.1411	66.29
SS9-SLAT6	Evi	5	16	**0.048**	**0.779**	41.48	**16.169**	**0.867**	**0.1393**	**65.69**

Table 4: Efficiency. Full-length efficiency on a clean single-H100 setting.

Model	Warmup / chunk (s)	Generation total (s)	Total / sequence (s)	Per frame (s)	Peak GPU (GB)
Stream3D	6.9	955	1062	10.6	18.4
Stream3D-Fast	**6.7**	**486**	**590**	**5.9**	**18.3**
MV-SAM3D	n/a	945	945	9.45	18.4

5 Conclusion

We introduced Stream3D, a training-free framework that extends frozen view-conditioned 3D generators to long monocular streams. Instead of transporting latent states across chunks, Stream3D uses the generator's cross-attention maps to identify historical views that provide reliable conditioning evidence for each 3D query token. This evidence is stored in a compact Adaptive Evidential Memory and used by Evidence-Based Multi-Generation to select a bounded view bundle for each chunk. As a result, Stream3D keeps memory constant with stream length while leaving the original generator unchanged. Experiments on GSO and NAVI show that Stream3D improves both appearance and geometry over single-view generators, multi-view diffusion baselines, and streaming alternatives such as KV-cache reuse and flow-based feature editing. These results support our central claim: streaming 3D generation is better addressed as an evidence selection problem than as a latent transport problem. By preserving token-level evidential views rather than unstable latent states, Stream3D provides a scalable path toward coherent long-horizon 3D generation from continuously arriving visual observations. Limitations are provided in the appendix.

References

1. 3DTopia: Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM (2023) 3
2. Anciukevičius, T., Xu, Z., Fisher, M., Henderson, P., Bilen, H., Mitra, N.J., Guerrero, P.: Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12608–12618 (2023) 3
3. Bahmani, S., Skorokhodov, I., Rong, V., Wetzstein, G., Guibas, L., Wonka, P., Tulyakov, S., Park, J.J., Tagliasacchi, A., Lindell, D.B.: 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7996–8006 (2024) 4
4. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023) 2
5. Campos, C., Elvira, R., Rodriguez, J.J.G., Montiel, J.M.M., Tardos, J.D.: Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam. IEEE Transactions on Robotics **37**(6), 1874–1890 (2021) 4
6. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022) 3
7. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22246–22256 (2023) 3
8. Chen, X., Chen, Y., Xiu, Y., Geiger, A., Chen, A.: Ttt3r: 3d reconstruction as test-time training. arXiv preprint arXiv:2509.26645 (2025) 4
9. Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., et al.: Sam 3d: 3dfy anything in images. arXiv preprint arXiv:2511.16624 (2025) 2, 3, 4, 5, 13
10. Chen, Y., Wang, L., Huang, W., Yang, S., Zhang, B., Xiao, Y., Chu, R., Mao, W., Hu, Q., Liu, S., et al.: Longlive-2.0: An nvfp4 parallel infrastructure for long video generation. arXiv preprint arXiv:2605.18739 (2026) 2
11. Chen, Z., Wang, Y., Wang, F., Wang, Z., Liu, H.: V3d: Video diffusion models are effective 3d generators. arXiv preprint arXiv:2403.06738 (2024) 4
12. Chen, Z., Tan, H., Zhang, K., Bi, S., Luan, F., Hong, Y., Li, F., Xu, Z.: Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. arXiv preprint arXiv:2410.12781 (2024) 4
13. Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O.: Monoslam: Real-time single camera slam. IEEE transactions on pattern analysis and machine intelligence **29**(6), 1052–1067 (2007) 4
14. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. IEEE (2022) 9
15. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(3), 611–625 (2018) 4
16. Engel, J., Schöps, T., Cremers, D.: Lsd-slam: Large-scale direct monocular slam. In: European conference on computer vision. pp. 834–849. Springer (2014) 4

17. Forster, C., Pizzoli, M., Scaramuzza, D.: Svo: Fast semi-direct monocular visual odometry. In: 2014 IEEE international conference on robotics and automation (ICRA). pp. 15–22. IEEE (2014) 4
18. Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., Litany, O., Gojcic, Z., Fidler, S.: Get3d: A generative model of high quality 3d textured shapes learned from images. In: Advances in Neural Information Processing Systems. vol. 35, pp. 31841–31854 (2022) 3
19. Henschel, R., Khachatryan, L., Hayrapetyan, D., Poghosyan, H., Tadevosyan, V., Wang, Z., Navasardyan, S., Shi, H.: Streamingt2v: Consistent, dynamic, and extendable long video generation from text. arXiv preprint arXiv:2403.14773 (2024) 2
20. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023) 3
21. Huang, H., Li, L., Cheng, H., Yeung, S.K.: Photo-slam: Real-time simultaneous localization and photorealistic mapping for monocular, stereo, and rgb-d cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21584–21593 (2024) 2, 4
22. Huang, Z., Boss, M., Vasishta, A., Rehg, J.M., Jampani, V.: Spar3d: Stable point-aware reconstruction of 3d objects from single images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16860–16870 (2025) 3
23. Jampani, V., Maninis, K.K., Engelhardt, A., Karpur, A., Truong, K., Sargent, K., Popov, S., Araujo, A., Martin Brualla, R., Patel, K., et al.: Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. Advances in Neural Information Processing Systems **36**, 76061–76084 (2023) 9
24. Jiang, Y., Zhang, L., Gao, J., Hu, W., Yao, Y.: Consistent4d: Consistent 360 dynamic object generation from monocular video. In: International Conference on Learning Representations (2024) 4
25. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023) 3
26. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023) 2, 3, 5
27. Kim, J., Kang, J., Choi, J., Han, B.: Fifo-diffusion: Generating infinite videos from text without training. In: Advances in Neural Information Processing Systems (2024) 2
28. Klein, G., Murray, D.: Parallel tracking and mapping for small ar workspaces. In: 2007 6th IEEE and ACM international symposium on mixed and augmented reality. pp. 225–234. IEEE (2007) 4
29. Kong, X., Liu, S., Lyu, X., Taher, M., Qi, X., Davison, A.J.: Eschernet: A generative model for scalable view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9503–9513 (2024) 3
30. Kong, X., Liu, S., Lyu, X., Taher, M., Qi, X., Davison, A.J.: Eschernet: A generative model for scalable view synthesis. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9503–9513. IEEE (2024) 10
31. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: Flowedit: Inversion-free text-based editing using pre-trained flow models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19721–19730 (2025) 2

32. Lan, Y., Luo, Y., Hong, F., Zhou, S., Chen, H., Lyu, Z., Yang, S., Dai, B., Loy, C.C., Pan, X.: Stream3r: Scalable sequential 3d reconstruction with causal transformer. arXiv preprint arXiv:2508.10893 (2025) 2, 4
33. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European conference on computer vision. pp. 71–91. Springer (2024) 4
34. Li, B., Wu, D., Li, J., Zhou, S., Zeng, Z., Li, L., Zha, H.: Mv-sam3d: Adaptive multi-view fusion for layout-aware 3d generation. arXiv preprint arXiv:2603.11633 (2026) 2, 3, 4
35. Li, J., Tan, H., Zhang, K., Xu, Z., Luan, F., Xu, Y., Hong, Y., Sunkavalli, K., Shakhnarovich, G., Bi, S.: Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. arXiv preprint arXiv:2311.06214 (2023) 2, 3
36. Li, P., Liu, Y., Long, X., Zhang, F., Lin, C., Li, M., Qi, X., Zhang, S., Luo, W., Tan, P., Wang, W., Liu, Q., Guo, Y.: Era3d: High-resolution multiview diffusion using efficient row-wise attention. arXiv preprint arXiv:2405.11616 (2024) 3
37. Li, W., Liu, J., Yan, H., Chen, R., Liang, Y., Chen, X., Tan, P., Long, X.: Craftsman3d: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. arXiv preprint arXiv:2405.14979 (2024) 3, 5
38. Li, Y., Zou, Z.X., Liu, Z., Wang, D., Liang, Y., Yu, Z., Liu, X., Guo, Y.C., Liang, D., Ouyang, W., et al.: Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 3
39. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 300–309 (2023) 3
40. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025) 9
41. Liu, M., Xu, C., Jin, H., Chen, L., Varma T, M., Xu, Z., Su, H.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. Advances in Neural Information Processing Systems **36**, 22226–22246 (2023) 2, 3
42. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9298–9309 (2023) 3
43. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023) 3
44. Liu, Y., Dong, S., Wang, S., Yin, Y., Yang, Y., Fan, Q., Chen, B.: Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16651–16662 (2025) 4
45. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9970–9980 (2024) 3
46. Maggio, D., Lim, H., Carlone, L.: Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. arXiv preprint arXiv:2505.12549 (2025) 4
47. Matsuki, H., Murai, R., Kelly, P.H.J., Davison, A.J.: Gaussian splatting slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18039–18048 (2024) 4

48. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12663–12673 (2023) 3
49. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision. pp. 405–421 (2020) 2, 3
50. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: Orb-slam: A versatile and accurate monocular slam system. IEEE transactions on robotics **31**(5), 1147–1163 (2015) 4
51. Mur-Artal, R., Tardós, J.D.: Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE transactions on robotics **33**(5), 1255–1262 (2017) 4
52. Newcombe, R.A., Lovegrove, S.J., Davison, A.J.: Dtam: Dense tracking and mapping in real-time. In: 2011 international conference on computer vision. pp. 2320–2327. IEEE (2011) 4
53. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022) 3
54. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022) 3
55. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., Ghanem, B.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. arXiv preprint arXiv:2306.17843 (2023) 3
56. Qiu, L., Chen, G., Gu, X., Zuo, Q., Xu, M., Wu, Y., Yuan, W., Dong, Z., Bo, L., Han, X.: Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9914–9925 (2024) 3
57. Ren, J., Xie, K., Mirzaei, A., Liang, H., Zeng, X., Kreis, K., Liu, Z., Torralba, A., Fidler, S., Kim, S.W., et al.: L4gm: Large 4d gaussian reconstruction model. Advances in Neural Information Processing Systems **37**, 56828–56858 (2024) 4
58. Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., Su, H.: Zero123++: A single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023) 3
59. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023) 3
60. Singer, U., Sheynin, S., Polyak, A., Ashual, O., Makarov, I., Kokkinos, F., Goyal, N., Vedaldi, A., Parikh, D., Johnson, J., Taigman, Y.: Text-to-4d dynamic scene generation. In: International Conference on Machine Learning. pp. 31915–31929 (2023) 4
61. Sun, J., Xie, Y., Chen, L., Zhou, X., Bao, H.: Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15598–15607 (2021) 4
62. Sun, J., Zhang, B., Shao, R., Wang, L., Liu, W., Xie, Z., Liu, Y.: Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. arXiv preprint arXiv:2310.16818 (2023) 3
63. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In: European Conference on Computer Vision. pp. 1–18. Springer (2024) 2, 3
64. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023) 3

65. Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22819–22829 (2023) 3
66. Tochilkin, D., Pankratz, D., Liu, Z., Huang, Z., Letts, A., Li, Y., Liang, D., Laforte, C., Jampani, V., Cao, Y.P.: Triposr: Fast 3d object reconstruction from a single image. arXiv preprint arXiv:2403.02151 (2024) 3
67. Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. arXiv preprint arXiv:2403.12008 (2024) 3
68. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12619–12629 (2023) 3
69. Wang, H., Agapito, L.: 3d reconstruction with spatial memory. In: 2025 International Conference on 3D Vision (3DV). pp. 78–89. IEEE (2025) 4
70. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. arXiv preprint arXiv:2503.11651 (2025) 4
71. Wang, J., Zhang, F., Li, X., Tan, V.Y., Pang, T., Du, C., Sun, A., Yang, Z.: Error analyses of auto-regressive video diffusion models: A unified framework. arXiv preprint arXiv:2503.10704 (2025) 2
72. Wang, P., Tan, H., Bi, S., Xu, Y., Luan, F., Sunkavalli, K., Wang, W., Xu, Z., Zhang, K.: Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. arXiv preprint arXiv:2311.12024 (2023) 3
73. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 4
74. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20697–20709 (2024) 4
75. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in neural information processing systems **36**, 8406–8441 (2023) 3
76. Wang, Z., Wang, Y., Chen, Y., Xiang, C., Chen, S., Yu, D., Li, C., Su, H., Zhu, J.: Crm: Single image to 3d textured mesh with convolutional reconstruction model. arXiv preprint arXiv:2403.05034 (2024) 3
77. Wei, X., Zhang, K., Bi, S., Tan, H., Luan, F., Deschaintre, V., Sunkavalli, K., Su, H., Xu, Z.: Meshlrm: Large reconstruction model for high-quality meshes. arXiv preprint arXiv:2404.12385 (2024) 3
78. Wu, K., Liu, F., Cai, Z., Yan, R., Wang, H., Hu, Y., Duan, Y., Ma, K.: Unique3d: High-quality and efficient 3d mesh generation from a single image. Advances in Neural Information Processing Systems **37**, 125116–125141 (2024) 3
79. Wu, S., Lin, Y., Zhang, F., Zeng, Y., Xu, J., Torr, P., Cao, X., Yao, Y.: Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. Advances in Neural Information Processing Systems **37**, 121859–121881 (2024) 3
80. Wu, Y., Zheng, W., Zhou, J., Lu, J.: Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. arXiv preprint arXiv:2507.02863 (2025) 4
81. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., et al.: Native and compact structured latents for 3d generation. arXiv preprint arXiv:2512.14692 (2025) 2, 3, 5, 10

82. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21469–21480 (2025) 3, 5, 10
83. Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y.: Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191 (2024) 3
84. Xu, Y., Shi, Z., Wang, Y., Chen, H., Yang, C., Peng, S., Shen, Y., Wetzstein, G.: Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. arXiv preprint arXiv:2403.14621 (2024) 3
85. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21924–21935 (2025) 4
86. Yi, T., Fang, J., Wang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., Wang, X.: Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6796–6807 (2024) 3
87. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021) 3
88. Zhao, Y., Yan, Z., Xie, E., Hong, L., Li, Z., Lee, G.H.: Animate124: Animating one image to 4d dynamic scene. arXiv preprint arXiv:2311.14603 (2023) 4
89. Zhao, Z., Lai, Z., Lin, Q., Zhao, Y., Liu, H., Yang, S., Feng, Y., Yang, M., Zhang, S., Yang, X., et al.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. arXiv preprint arXiv:2501.12202 (2025) 3, 5
90. Zheng, Y., Li, X., Nagano, K., Liu, S., Kreis, K., Hilliges, O., De Mello, S.: A unified approach for text- and image-guided 4d scene generation. arXiv preprint arXiv:2311.16854 (2023) 4
91. Zhou, K., Bian, J.W., Zheng, J.Q., Zhong, J., Xie, Q., Trigoni, N., Markham, A.: Manydepth2: Motion-aware self-supervised monocular depth estimation in dynamic scenes. IEEE Robotics and Automation Letters (2025) 4
92. Zhou, K., Chen, Y., Zhan, F., Hua, H., Chen, G., Chang, X., Qu, A., Du, Y., Liu, Z., Liang, P.P., et al.: Gem-4d: Geometry-enhanced video world models for robot manipulation. arXiv preprint arXiv:2605.22882 (2026) 4
93. Zhou, K., Wang, Y., Chen, G., Beaudouin, G., Zhan, F., Liang, P., Wang, M.: Page-4d: Disentangled pose and geometry estimation for vggt-4d perception. In: International Conference on Learning Representations. vol. 2026, pp. 36401–36414 (2026) 4
94. Zhuo, D., Zheng, W., Guo, J., Wu, Y., Zhou, J., Lu, J.: Streaming 4d visual geometry transformer. arXiv preprint arXiv:2507.11539 (2025) 2, 4
95. Zhuo, D., Zheng, W., Guo, J., Wu, Y., Zhou, J., Lu, J.: Streaming visual geometry transformer. In: International Conference on Learning Representations. vol. 2026, pp. 88055–88072 (2026) 10

6 Additional Analysis

6.1 Efficiency and Memory Footprint

Stream3D adds one warmup probe per chunk to extract cross-attention evidence. Tab. 6 shows that generation time scales similarly with bundle size K for Stream3D and MV-SAM3D. This confirms that increasing K mainly affects the shared multi-view generation pass, while the evidential-memory mechanism adds little extra scaling cost. Stream3D-Fast uses the same memory, selection, and fusion mechanism as Stream3D, but replaces the default stage-1 sampler with the official few-step distilled sampler. It therefore represents a latency–quality trade-off within the same framework rather than a separate method.

Table 6: Generation-time scaling with bundle size K.

Model	$K = 4$ gen total (s)	$K = 8$ gen total (s)	$K = 8$ / $K = 4$
Stream3D	649	955	1.47×
MV-SAM3D	638	945	1.48×

6.2 Object-Centric Alignment and Evaluation Protocol

Outputs from object-centric 3D generation methods often live in method-specific canonical coordinate frames. Directly comparing them in their native coordinates would conflate reconstruction quality with arbitrary choices of global pose, scale, and axis convention. We therefore align all method outputs to the same benchmark coordinate system before computing metrics.

The alignment uses the generated source geometry, such as mesh vertices or Gaussian centers, and estimates a global Sim(3) transform to match the ground-truth object geometry. We initialize from multiple candidate orientations to account for common canonical-frame differences and then refine with ICP. This alignment is restricted to a global similarity transform: it does not deform geometry, add missing parts, modify texture, or correct local artifacts.

After automatic alignment, we perform a rendered-view audit in the fixed evaluation cameras. This audit catches gross semantic pose or scale errors that can occur when several candidates have similar Chamfer distances but correspond to different object orientations. If such an error is detected, the remaining ranked global Sim(3) candidates are reviewed, and an alternative is selected only when it visibly resolves the alignment failure. Final metrics are recomputed after the accepted alignment. This makes comparisons across heterogeneous methods more consistent while preserving local reconstruction errors for evaluation.

Table 7: Ablation on evidence-score fusion variants.

Variant	CD-L2 ↓	IoU ↑	P-FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Img-FID ↓
Product (default)	**0.0478**	**0.775**	**40.64**	**16.145**	**0.866**	**0.1396**	**66.71**
Sum $\alpha = 0.25$	0.0505	0.7681	44.21	15.82	0.8603	0.1481	75.88
Sum $\alpha = 0.50$	0.0527	0.7591	44.36	15.64	0.8586	0.1500	76.61
Sum $\alpha = 0.75$	0.0545	0.7536	44.42	15.55	0.8584	0.1521	77.67
Evi(N)	0.056	0.749	44.98	15.730	0.863	0.1486	74.08
Ent	0.053	0.760	47.26	15.892	0.864	0.1449	67.75

6.3 Evidence-Score Fusion

Tab. 7 analyzes different evidence-score fusion variants. Our default product form combines two complementary signals: attention scale and attention concentration. The scale term measures whether a view receives above-average attention mass for a query token, while the concentration term measures whether that attention is spatially localized over image patches.

The ablation shows that both terms are necessary. Attention-mass-only evidence can favor views with diffuse, non-specific attention, while entropy-only evidence can favor sharply peaked but weak responses. Weighted-sum variants improve over some single-component scores but remain less stable than the product form. The product score requires a view to be both strongly attended and spatially focused, making the retained evidence more discriminative for token-level 3D generation.

Overall, these results support three conclusions: Stream3D maintains bounded evidence memory independent of stream length, its runtime overhead is mainly a fixed warmup probe while generation cost is dominated by the frozen backbone, and its product-form evidence score better identifies useful historical views than single-component or weighted-sum alternatives.

6.4 More Visulization

Fig. 6 illustrates the key difference between independent single-view generation and streaming generation. SAM3D produces plausible outputs from individual frames, but its results do not accumulate evidence over time. Stream3D instead updates a compact evidential memory as new chunks arrive, allowing later generations to use information from earlier informative views. This leads to progressively more complete reconstructions over long streams.

Fig. 7 further compares Stream3D with single-view and multi-view generation baselines. Single-view methods are constrained by partial observation and must hallucinate large unobserved regions. Multi-view diffusion baselines improve consistency by fusing multiple views, but they operate on a bounded view set and lack an explicit long-range memory mechanism. Stream3D improves over these baselines by retaining token-level evidence from the stream and selecting the most informative historical observations for generation.

Fig. 6: Long-horizon comparison between SAM3D and Stream3D. We visualize the generation results of the single-view SAM3D baseline and Stream3D as more frames are observed from the input stream. SAM3D processes each observation independently and therefore lacks a mechanism to accumulate evidence across time. In contrast, Stream3D maintains an adaptive evidential memory and progressively incorporates informative historical views. As the stream grows, Stream3D produces more complete and stable 3D assets, while SAM3D remains limited by the information available from individual frames.

Fig. 8 compares different streaming strategies. Cache- and transport-based methods reuse latent or feature states, but these states can become stale or accumulate errors over long sequences. Frame-level MV-SAM3D variants are more stable, but selecting views globally can discard observations that are important for specific spatial regions. Stream3D addresses this by maintaining token-wise evidential memory, which enables different query tokens to rely on different historical views while keeping the conditioning bundle bounded.

7 Limitations

The method is built on top of the quality of the underlying pretrained generator. If the base model fails to reconstruct the object from individual views, the proposed evidential memory cannot fully recover the missing geometry or appearance.

Fig. 7: Qualitative comparison with single-view and multi-view 3D generation baselines. We compare the ground truth with SAM3D, TRELLIS+M.D., TRELLIS.2, TRELLIS.2+M.D., Stream3D(SAM3D), and Stream3D(TRELLIS.2). Single-view methods recover plausible visible regions but often hallucinate incomplete or inconsistent unseen geometry. Multi-view diffusion improves view consistency with a bounded input set, but can still miss fine geometry or produce artifacts under long-stream observations. Stream3D improves both completeness and consistency by selecting informative historical views through token-level evidential memory, and the gains are visible across both SAM3D and TRELLIS.2 backbones.

Fig. 8: Qualitative comparison with streaming baselines. We compare the ground truth with FlowEdit, KV-Cache, MV-SAM3D, MV-SAM3D with random view selection, and Stream3D(SAM3D). FlowEdit improves short-range consistency but is limited by local latent editing and can lose long-range evidence. KV-Cache carries historical states forward but does not explicitly filter them by geometric reliability. MV-SAM3D variants select a compact set of frames at the view level, which can miss local evidence needed by specific spatial regions. In contrast, Stream3D retains token-level evidence over time and selects a bounded conditioning bundle, leading to more complete geometry and more stable appearance under long-stream generation.