Streaming3D: Sequential 3D Generation via Evidential Memory

Anonymous Author(s)
Affiliation
Address
email

Abstract

View-conditioned 3D generators such as SAM 3D, TRELLIS and Hunyuan3D produce high-quality object reconstructions from a single view, but real-world capture often arrives as long monocular streams. Naively applying above generators to streaming inputs will lead to severely inconsistent generation. To address this problem, we propose Streaming3D, the first training-free streaming mechanism that turns a frozen view-conditioned 3D generator into a streaming generator with constant cross-chunk memory. Specifically, Streaming3D achieves this by maintaining a compact evidential memory, which selectively caches the most informative historical frames based on their confidence scores. 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 long realistic and synthetic streams, Streaming3D outperforms latent-transport baselines, including KV-cache reuse and flow-based feature editing, across both photometric and geometric metrics. It maintains a constant memory footprint and stable reconstruction quality as sequence length increases. More details could be found at the project page: `Link`.

1 Introduction

Object-centric 3D generation is becoming a practical building block for vision and robotics [28, 23, 40]. Recent systems such as SAM 3D Objects [5] and TRELLIS.2 [50] can take an image, isolate an object, and reconstruct a Gaussian splat [13] or mesh within seconds. Yet their input model remains fundamentally non-streaming. They are designed for a single view, or a small set of views, while real capture devices produce long monocular streams: a phone circling an object, a robot observing while moving, or a wearable camera recording continuously. Naively applying a single-view 3D generator frame by frame yields temporally inconsistent reconstructions; feeding all frames via multi-diffusion method [3] or multiview fusion [21] is computationally expensive or even infeasible; and processing fixed-size chunks via flow matching-based method [17] loses the history needed for global consistency.

An alternative approach is to adapt techniques from streaming video generation or 3D reconstruction [18, 57]. For instance, state-transport mechanisms like KV banks [18] or FlowEdit-style velocity edits [17] could be introduced to propagate information across chunks. However, transporting latent states in this manner is fundamentally 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. Furthermore, as these methods transport and accumulate state over time, their memory footprint naturally grows with sequence length. Ultimately,

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

the transported state itself can become a source of error: over long streams, the very mechanism intended to preserve consistency may paradoxically degrade it.

To solve this problem, we propose a novel training-free method: **Streaming3D**. Instead of transporting latent states, Streaming3D incrementally retains the observed frames that provide the most reliable conditioning evidence for 3D generation as new streaming inputs arrive. Our key observation is that a frozen view-conditioned generator already exposes such conditioning evidence through its cross-attention maps. During a cheap one-step warmup pass, if a query token in a 3D volume attends to a frame both strongly and selectively, that frame provides confident evidence for the corresponding part of the 3D volume. We count this signal as an accumulative *evidence score* for this view with respect to the query token. The stream history is then treated as a pool of candidate views: for each query token, we maintain the historical frames with the highest evidence observed so far. At each chunk, we aggregate these token-level preferences, select a small bundle of frames that collectively cover the most query tokens, and run the original generator on this bounded view set.

Specifically, we instantiate this idea as *Token-Vote View Memory*, a compact streaming memory that retains, for each query token, the highest-evidence frames observed so far. At each chunk, token-level evidence is aggregated into frame-level ownership scores; the top-K frames are selected as a conditioning bundle and passed to the pre-trained 3D generator via *Evidence-Based Multi-Generation*. Because the memory stores only a fixed number of evidential frame indices per token, its footprint remains constant with stream length. The generator itself is left completely unchanged: no retraining, architectural modification, or auxiliary loss is required, since the memory module only controls which observed frames the generator sees.

Notably, Streaming3D provides two properties that latent transport does not offer. First, the memory will not linearly scale with the stream length. Second, evidence accumulation is monotonic: for each query token, the retained evidence score can only remain unchanged or improve as new frames arrive. Consequently, in terms of the evidence score, the selected conditioning bundle is never worse than one selected earlier. As a comparison, the latent-transport schemes cannot provide such a non-degradation guarantee as they require hand-engineered rules to align streaming frames in latent space, thus accumulating inconsistencies over time. In a nutshell, *view memory transports evidential frames rather than latent*, thereby sidestepping the latent-alignment problem entirely.

We evaluate Streaming3D as a lightweight streaming wrapper around pre-trained SAM 3D and TRELLIS.2 on long monocular streams. Across photometric and geometric metrics, Streaming3D improves over latent-transport baselines while maintaining a constant memory footprint and avoiding the long-horizon degradation observed in KV-bank and FlowEdit-style alternatives. In summary, our contributions are threefold:

- **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.
- **Token-Vote View Memory.** We introduce a compact view-memory mechanism that transports evidence rather than latent states. The memory is interpretable, coordinate-free, and training-free; its footprint is constant in stream length.
- **Superior Streaming Performance.** We show that the proposed mechanism matches or improves latent-transport baselines while avoiding their long-horizon degradation.

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 [11, 41], surface meshes [53, 47, 52], 3D Gaussian splats [12], and native structured latent volumes [55, 48, 25, 24, 5, 51]. Along the second axis, existing approaches cover closed-form regressors [11, 41, 53, 47, 12, 51], 3D latent diffusion models [55, 48, 25, 24, 5], and SDS-based optimization methods [37, 26, 46, 43]. Some layout-aware variants [5, 22, 1] 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, 8, 31, 32, 39, 15] relax this setting by first hallucinating additional views before reconstructing 3D geometry. Video-conditioned reconstruction models [45, 20, 54, 56] and 4D-aware generators [2, 6, 38] extend the input format further to short temporal clips. More recently, MV-SAM3D [21] 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 [5] 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 [9, 19, 7, 14, 36, 10, 34, 35], which incrementally estimates camera motion and scene structure from video. Recent methods have extended modern feed-forward reconstruction models to online settings. Notably, Spann3R [42] augments a DUSt3R-style encoder [45] with a token-addressable spatial memory, enabling online pointmap fusion over long image streams. Similarly, SLAM3R [30], also built upon DUSt3R, introduces a real-time end-to-end dense reconstruction system that directly predicts 3D pointmaps from RGB videos. Point3R [49] 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 [33] addresses this limitation by adopting the more powerful VGGT transformer, which supports image sets of arbitrary length. CUT3R [44] 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 [57] and Stream3R [18] employ causal transformers to implicitly cache historical visual tokens. In addition, because CUT3R suffers from severe drift on long streaming inputs, TTT3R [4] proposes a simple empirical state-update rule to improve sequence-length generalization. Streaming3D 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.

3 Method

We extend view-conditioned 3D generators — SAM 3D Objects [5] and TRELLIS.2 [50] — to handel long streaming inputs without retraining, architectural change, or auxiliary loss. The *Token-Vote View Memory* (Sec. 3.2) enables efficient long-range memory by retaining compact, 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 [13], mesh, or latent volume) $\hat{y} = f_\theta(v; z_0)$. Recent 3D generation models, i.e., SAM 3D [5], TRELLIS [52], TRELLIS.2 [51], Hunyuan3D 2.0 [55] and CraftsMan3D [24], 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 — with at least one cross-attention layer of the form:

$$\mathbf{A}_v \;=\; \mathrm{softmax}\!\left(\frac{\mathbb{Q}\,\mathbb{K}^\top}{\sqrt{d}}\right) \;\in\; [0,1]^{Q\times P}, \tag{1}$$

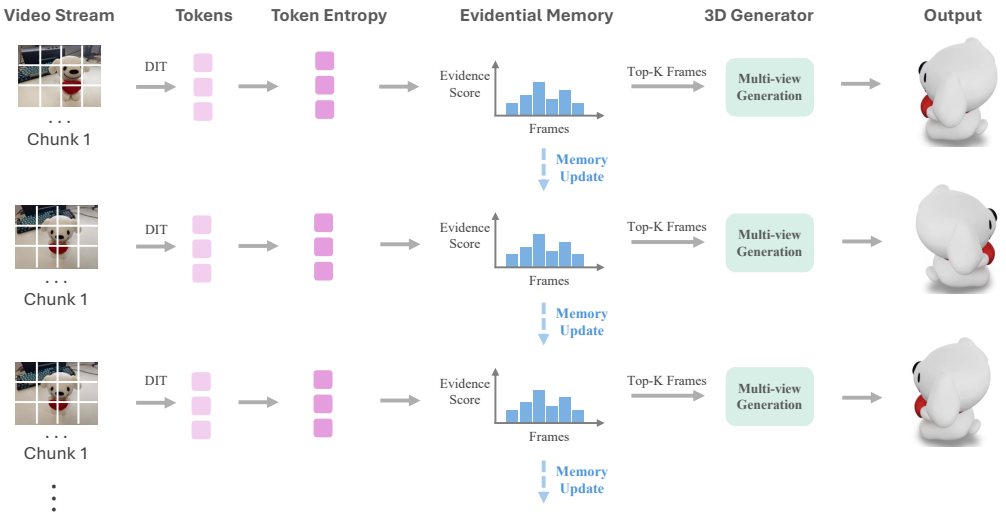

Figure 1: **Framework of Streaming3D.** During training, a video DiT predicts the velocity of the noised video latent, while its intermediate features guide a geometry DiT to predict geometry velocity. This coupled training enforces geometry-consistent generation. During inference, only the video branch is used for efficient generation.

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 preceding chunks, while maintaining a memory footprint that does not scale linearly with stream length T.

3.2 Token-Vote View Memory

To maintain an efficient memory footprint that does not scale linearly with the total sequence length T, we introduce a *Token-Vote View Memory* mechanism. 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 an **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 a certain chunk, we run a single denoising step of the structure-stage generator on the chunk's condition set, with a *frozen* prior z_0 that is sampled once on chunk 0 and reused thereafter. We extract the cross-attention map equation 1 at one fixed layer L, for a query token q that excludes special tokens (CLS, register, etc.), and compute the per-token cross-attention induced by view v (with P patch tokens) as below:

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

$$\mathbf{M}_v[q] = \left(\sum_{i=1}^{P} \mathbf{A}_v[q,i] \right) \cdot \Big(1 - \mathbf{H}_v(\tilde{A}_v[q]) \Big), \tag{3}$$

Algorithm 1 Streaming inference with the Token-Vote View Memory.

Require: generator f_θ, chunk C, chunk frame v, chunk frame index k, layer L, token range $[a, b]$, memory depth D, bundle size K.

Sample $z_0 \sim \mathcal{N}(0, I)$ ▷ frozen prior
$\mathbf{M} \leftarrow \mathbf{0}^{Q\times D}$, $\mathbf{F} \leftarrow \mathbf{0}^{Q\times D}$ ▷ evidence and frame-index memory
for $k = 0, 1, 2, \ldots$ **do**
  Run one warmup denoising step of f_θ on $\mathcal{C}$ with prior z_0
  Extract $\mathbf{A}_v$ at layer L via Eq. 1 for each $v \in \mathcal{C}$
  Compute per-view evidence score $\mathbf{M}_v \in [0, 1]^Q$ on token range $[a, b]$
  Update memory: row-wise top-D merge for each query token q
  Compute ownership counts $\{n_{f_r}\}$ via Eq. 4
  Select bundle $\mathcal{V}^\star$ via selecting the top-K frames
  Fuse per-view scores via Eq. 5 during the full forward
  output $\hat{y} \leftarrow f_\theta(\mathcal{V}^\star; z_k)$ with fresh z_k
end for

which yields a per-view evidence score $\mathbf{M}_v[q] \in [0, 1]$ for query token q. Notably, there are three intentional designs in this probe as explained below: **Frozen prior**: reusing z_0 across chunks makes evidence score $\mathbf{M}_v[q]$ comparable across chunks; any difference between two chunks' entropy maps is attributable to the views, not to a fresh draw of noise. **Single denoising step**: the first step's attention is dominated by the input views' content; later steps are progressively shaped by the model's own prediction, which would be circular for a view-selection objective.

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 K 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. The total footprint is $2 \times Q \times D$ scalars, *constant in stream length* T — about 50 KB for SAM 3D (Q=4096, D=4).

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 $\mathbf{F}$:

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

and select the K frames with the largest counts for $\mathcal{V}$. 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

Table 1: Quantitative comparison on appearance and geometry metrics.

Data	Method	Appearance			Geometry			
		PSNR↑	SSIM↑	LPIPS↓	Depth MAE↓	Depth RMSE↓	Acc@5cm↑	RelAcc@5↑
Data1	Zero123	–	–	–	–	–	–	–
	SAM3D	–	–	–	–	–	–	–
	TRELLIS.2	–	–	–	–	–	–	–
	EscherNet [16]	–	–	–	–	–	–	–
	TRELLIS+M.D.	–	–	–	–	–	–	–
	Ours	–	–	–	–	–	–	–
Data2	Zero123	–	–	–	–	–	–	–
	SAM3D	–	–	–	–	–	–	–
	TRELLIS.2	–	–	–	–	–	–	–
	EscherNet [16]	–	–	–	–	–	–	–
	TRELLIS+M.D.	–	–	–	–	–	–	–
	Ours	–	–	–	–	–	–	–

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 Token-Vote View Memory admits two structural properties that distinguish it from any latent-transport scheme. First, the cross-chunk footprint is at most $2 \times Q \times D$ scalars and is therefore bounded independently of stream length T. Second, for every query token q and rank j, the cached evidence values $\mathbf{M}^{(k)}[q, j]$ are *non-decreasing* in k: only retains the M largest entries of the union of the previous row and the new candidates, so the j-th largest can only increase. The conditioning bundle supplied to f_θ at chunk k is therefore never worse, in evidence-score terms, at chunk $k-1$, while the footprint stays constant in T. KV banks, prev-chunk query banks, and FlowEdit-style velocity edits admit no analogous non-degradation guarantee, and their state grows linearly in k. The mechanism is not a new generative model, not a fine-tuning or distillation scheme, and not a view-acquisition policy: it selects among views already captured, leaves f_θ's weights and architecture untouched, and involves no learning.

4 Experiment

We evaluate Streaming3D on long-stream 3D generation, where the model receives a sequence of continuously arriving posed images and must maintain a coherent 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 Streaming3D 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 xxx GPU. We use SAM3D with its standard settings as the underlying generation backbone. We adopt Depth Anything 3 [27] to estimate the camera pose and depth of input frames to get point map, which are fed into SAM 3D for 3D generation.

and all default hyperparameters, memory depth, bundle size, attention layer, stream length, ...

Dataset: We evaluate on two complementary benchmarks. The first is the GSO benchmark, 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 geometry 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 DL3DV, which contains realistic scene-level captures with complex geometry, natural textures, and diverse camera trajectories. Compared with GSO, DL3DV

Table 2: Ablation studies with different streaming method

Data	Method	Appearance			Geometry			
		PSNR↑	SSIM↑	LPIPS↓	Depth MAE↓	Depth RMSE↓	Acc@5cm↑	RelAcc@5↑
Data1	MD with K random views	–	–	–	–	–	–	–
	MD with K farthest-point-sampled views	–	–	–	–	–	–	–
	KV-Cache.	–	–	–	–	–	–	–
	Flowedit.	–	–	–	–	–	–	–
	MV-SAM3D	–	–	–	–	–	–	–
	Ours	–	–	–	–	–	–	–

introduces more challenging long-range consistency requirements because the stream may contain repeated structures, large viewpoint changes, partial observations, and accumulated occlusions.

Baseline: We compare Streaming3D with representative single-view, multi-view, and streaming baselines. Single-view baselines, including Zero123, SAM3D, and TRELLIS, generate 3D content from individual observations and therefore provide a lower bound on streaming consistency. EscherNet serves as a strong multi-view baseline that benefits from multiple posed observations but is not designed for unbounded streaming inputs. We also compare against TRELLIS+M.D. [52] and TRELLIS.2+M.D. [51] which aggregates 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 PSNR, SSIM, and LPIPS 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 Depth MAE, Depth RMSE, Acc@5cm, and RelAcc@5%. These metrics measure absolute depth quality, relative geometric consistency, and fine-grained 3D accuracy.

4.2 Main Results

Table 1 reports quantitative results on both GSO and DL3DV. Across the two datasets, Streaming3D achieves the strongest overall performance on both appearance and geometry metrics. The gains are consistent across all metrics, indicating that the improvement is not limited to image-level rendering quality but also reflects better 3D structure. We could conclude: **(1)** Compared with single-view baselines such as Zero123, SAM3D, TRELLIS, and TRELLIS.2, Streaming3D substantially improves both appearance and geometry. This confirms that streaming observations provide important information that cannot be recovered from a single image alone. Single-view methods often generate plausible observed geometry but hallucinate unobserved regions inconsistently across time. **(2)** Compared with multi-view baselines such as EscherNet, TRELLIS+M.D. and TRELLIS.2+M.D., Streaming3D shows stronger long-stream behavior. EscherNet benefits from multiple posed views, but it is designed for bounded multi-view input rather than continuously arriving streams. TRELLIS+M.D. can improve consistency by fusing multiple views, but full-context diffusion becomes computationally expensive as the stream grows and does not provide an explicit mechanism for compact long-range memory. Streaming3D addresses this limitation by retaining token-level evidence rather than storing or reprocessing all frames. As a result, it maintains high reconstruction quality while keeping memory usage constant.

4.3 Streaming Baseline Comparison

Table 2 further compares Streaming3D with several streaming alternatives. **(1)** The SAM3D with multi-diffusion variants evaluate whether selecting a small set of representative views can approximate streaming memory. Random sampling is unstable because it may discard geometrically important observations. Farthest-point sampling improves geometric diversity by selecting views that are well separated in pose space. These results highlight the limitation of frame-level view selection. Streaming generation requires a more fine-grained memory mechanism: different spatial tokens may require evidence from different historical views. A single selected frame can be useful for one surface region but uninformative or even misleading for another. Streaming3D allows the generator to combine the most reliable historical evidence for each spatial region, improving both appearance fidelity and geometric consistency. **(2)** We also compare against cache- and transport-based streaming baselines. KV-cache reuse provides an efficient mechanism for carrying historical information forward, but cached tokens are not explicitly filtered according to geometric reliability. As a result,

Table 3: Ablation with different module: mainly ablate hyperparameter selction.

Data	Method	Appearance			Geometry			
		PSNR↑	SSIM↑	LPIPS↓	Depth MAE↓	Depth RMSE↓	Acc@5cm↑	RelAcc@5↑
Data1	K = 4	–	–	–	–	–	–	–
	K = 8	–	–	–	–	–	–	–
	K = 10	–	–	–	–	–	–	–
	K = 12	–	–	–	–	–	–	–
	K = 12/(m=80)	–	–	–	–	–	–	–
	K = 12/(m=100)	–	–	–	–	–	–	–
	K = 12/(m=120)	–	–	–	–	–	–	–

the memory can accumulate stale or ambiguous evidence, especially under long camera motion. FlowEdit performs local flow-based latent editing and improves short-range consistency, but it operates on fixed-size chunks and therefore loses long-range history. This limitation becomes visible in long streams, where global consistency depends on observations that may fall outside the current editing window. In contrast, Streaming3D maintains a compact but persistent evidence memory. The memory does not simply reuse all past latent states; it selects and aggregates historical information according to token-level evidence. textbf(3) We also compare Streaming3D with MV-SAM3D [21], which selects a compact set of informative input views using visibility-aware weighting. Although MV-SAM3D is effective in bounded multi-view settings, it is not designed for streaming generation: once the input stream grows, selecting a small global view subset can discard historical observations that are locally important for specific surfaces or tokens. In contrast, Streaming3D maintains a token-wise evidential memory, allowing different spatial regions to retrieve different historical views according to their accumulated evidence. This provides a more flexible and scalable mechanism for long-stream generation, preserving global consistency without requiring all past frames to be regenerated or jointly fused.

4.4 Ablation Studies

5 Conclusion

We introduced Streaming3D, a training-free framework for extending frozen view-conditioned 3D generators to long monocular streams. Instead of transporting latent states across chunks, which can accumulate alignment errors and grow with sequence length, Streaming3D transports *evidence*: it uses cross-attention from the frozen generator to identify which historical views provide reliable conditioning signals for each 3D query token. This leads to a compact *Token-Vote View Memory* that maintains constant cross-chunk memory while preserving the most informative observations for future generation. Combined with *Evidence-Based Multi-Generation*, the selected evidential views enable the original generator to produce temporally consistent and geometrically coherent 3D outputs without retraining, architectural modification, or auxiliary losses. Our experiments on long synthetic and realistic streams show that Streaming3D improves both appearance and geometry over single-view generators, bounded multi-view baselines, and streaming alternatives such as KV-cache reuse and flow-based feature editing. The results highlight a key distinction between streaming reconstruction and streaming generation: while reconstruction primarily aggregates observed geometry, streaming 3D generation must also synthesize unobserved structure while remaining consistent with an expanding visual history. By retaining token-level evidence rather than all frames or unstable latent states, Streaming3D provides a scalable mechanism for long-horizon 3D generation with bounded memory. More broadly, Streaming3D reframes streaming 3D generation as an evidence selection problem rather than a latent transport problem. This perspective suggests a practical path for turning powerful fixed-input 3D generators into online systems: keep the generator frozen, expose the reliability signals already present in its attention maps, and use them to decide what the model should remember. We believe this provides a simple and general foundation for future streaming 3D and 4D generation systems that must operate over long, redundant, and continuously arriving visual observations.

References

[1] T. Anciukevičius, Z. Xu, M. Fisher, P. Henderson, H. Bilen, N. J. Mitra, and P. Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12608–12618, 2023.

[2] S. Bahmani, I. Skorokhodov, V. Rong, G. Wetzstein, L. Guibas, P. Wonka, S. Tulyakov, J. J. Park, A. Tagliasacchi, and D. B. Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024.

[3] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023.

[4] X. Chen, Y. Chen, Y. Xiu, A. Geiger, and A. Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025.

[5] X. Chen, F.-J. Chu, P. Gleize, K. J. Liang, A. Sax, H. Tang, W. Wang, M. Guo, T. Hardin, X. Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

[6] Z. Chen, Y. Wang, F. Wang, Z. Wang, and H. Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024.

[7] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1052–1067, 2007.

[8] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023.

[9] J. Engel, T. Schöps, and D. Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014.

[10] C. Forster, M. Pizzoli, and D. Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 15–22. IEEE, 2014.

[11] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

[12] Z. Huang, M. Boss, A. Vasishta, J. M. Rehg, and V. Jampani. Spar3d: Stable point-aware reconstruction of 3d objects from single images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16860–16870, 2025.

[13] B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[14] G. Klein and D. Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM international symposium on mixed and augmented reality*, pages 225–234. IEEE, 2007.

[15] X. Kong, S. Liu, X. Lyu, M. Taher, X. Qi, and A. J. Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9503–9513, 2024.

[16] X. Kong, S. Liu, X. Lyu, M. Taher, X. Qi, and A. J. Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9503–9513, 2024.

[17] V. Kulikov, M. Kleiner, I. Huberman-Spiegelglas, and T. Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19721–19730, 2025.

[18] Y. Lan, Y. Luo, F. Hong, S. Zhou, H. Chen, Z. Lyu, S. Yang, B. Dai, C. C. Loy, and X. Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer. *arXiv preprint arXiv:2508.10893*, 2025.

[19] D. D. Lee, P. Pham, Y. Largman, and A. Ng. Advances in neural information processing systems 22. *Tech Rep*, 1(1):1–11, 2009.

[20] V. Leroy, Y. Cabon, and J. Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pages 71–91. Springer, 2024.

[21] B. Li, D. Wu, J. Li, S. Zhou, Z. Zeng, L. Li, and H. Zha. Mv-sam3d: Adaptive multi-view fusion for layout-aware 3d generation. *arXiv preprint arXiv:2603.11633*, 2026.

[22] B. Li, D. Wu, J. Li, S. Zhou, Z. Zeng, L. Li, and H. Zha. Mv-sam3d: Adaptive multi-view fusion for layout-aware 3d generation. *arXiv preprint arXiv:2603.11633*, 2026.

[23] J. Li, H. Tan, K. Zhang, Z. Xu, F. Luan, Y. Xu, Y. Hong, K. Sunkavalli, G. Shakhnarovich, and S. Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023.

[24] W. Li, J. Liu, H. Yan, R. Chen, Y. Liang, X. Chen, P. Tan, and X. Long. Craftsman3d: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024.

[25] Y. Li, Z.-X. Zou, Z. Liu, D. Wang, Y. Liang, Z. Yu, X. Liu, Y.-C. Guo, D. Liang, W. Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[26] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 300–309, 2023.

[27] H. Lin, S. Chen, J. Liew, D. Y. Chen, Z. Li, G. Shi, J. Feng, and B. Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[28] M. Liu, C. Xu, H. Jin, L. Chen, M. Varma T, Z. Xu, and H. Su. 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.

[29] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023.

[30] Y. Liu, S. Dong, S. Wang, Y. Yin, Y. Yang, Q. Fan, and B. Chen. Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16651–16662. IEEE, 2025.

[31] Y. Liu, C. Lin, Z. Zeng, X. Long, L. Liu, T. Komura, and W. Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023.

[32] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9970–9980, 2024.

[33] D. Maggio, H. Lim, and L. Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv preprint arXiv:2505.12549*, 2025.

[34] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[35] R. Mur-Artal and J. D. Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017.

[36] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. Dtam: Dense tracking and mapping in real-time. In *2011 international conference on computer vision*, pages 2320–2327. IEEE, 2011.

[37] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[38] J. Ren, K. Xie, A. Mirzaei, H. Liang, X. Zeng, K. Kreis, Z. Liu, A. Torralba, S. Fidler, S. W. Kim, et al. L4gm: Large 4d gaussian reconstruction model. *Advances in Neural Information Processing Systems*, 37:56828–56858, 2024.

[39] Y. Shi, P. Wang, J. Ye, M. Long, K. Li, and X. Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

[40] J. Tang, Z. Chen, X. Chen, T. Wang, G. Zeng, and Z. Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.

[41] D. Tochilkin, D. Pankratz, Z. Liu, Z. Huang, A. Letts, Y. Li, D. Liang, C. Laforte, V. Jampani, and Y.-P. Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.

[42] H. Wang and L. Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pages 78–89. IEEE, 2025.

[43] H. Wang, X. Du, J. Li, R. A. Yeh, and G. Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12619–12629, 2023.

[44] Q. Wang, Y. Zhang, A. Holynski, A. A. Efros, and A. Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2025.

[45] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20697–20709, 2024.

[46] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023.

[47] K. Wu, F. Liu, Z. Cai, R. Yan, H. Wang, Y. Hu, Y. Duan, and K. Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *Advances in Neural Information Processing Systems*, 37:125116–125141, 2024.

[48] S. Wu, Y. Lin, F. Zhang, Y. Zeng, J. Xu, P. Torr, X. Cao, and Y. Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *Advances in Neural Information Processing Systems*, 37:121859–121881, 2024.

[49] Y. Wu, W. Zheng, J. Zhou, and J. Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. *arXiv preprint arXiv:2507.02863*, 2025.

[50] J. Xiang, X. Chen, S. Xu, R. Wang, Z. Lv, Y. Deng, H. Zhu, Y. Dong, H. Zhao, N. J. Yuan, et al. Native and compact structured latents for 3d generation. *arXiv preprint arXiv:2512.14692*, 2025.

[51] J. Xiang, X. Chen, S. Xu, R. Wang, Z. Lv, Y. Deng, H. Zhu, Y. Dong, H. Zhao, N. J. Yuan, et al. Native and compact structured latents for 3d generation. *arXiv preprint arXiv:2512.14692*, 2025.

[52] J. Xiang, Z. Lv, S. Xu, Y. Deng, R. Wang, B. Zhang, D. Chen, X. Tong, and J. Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21469–21480, 2025.

[53] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

[54] J. Yang, A. Sax, K. J. Liang, M. Henaff, H. Tang, A. Cao, J. Chai, F. Meier, and M. Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21924–21935, 2025.

[55] Z. Zhao, Z. Lai, Q. Lin, Y. Zhao, H. Liu, S. Yang, Y. Feng, M. Yang, S. Zhang, X. Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

[56] K. Zhou, Y. Wang, G. Chen, X. Chang, G. Beaudouin, F. Zhan, P. P. Liang, and M. Wang. Page-4d: Disentangled pose and geometry estimation for 4d perception. *arXiv e-prints*, pages arXiv–2510, 2025.

[57] D. Zhuo, W. Zheng, J. Guo, Y. Wu, J. Zhou, and J. Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025.