Novelty Assessment Report

Paper: Streaming Visual Geometry Transformer
PDF URL: https://openreview.net/pdf?id=5APgTKsnx8
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Perceiving and reconstructing 3D geometry from videos is a fundamental yet challenging computer vision task. To facilitate interactive and low-latency applications, we propose a streaming visual geometry transformer that shares a similar philosophy with autoregressive large language models. We explore a simple and efficient design and employ a causal transformer architecture to process the input sequence in an online manner. We use temporal causal attention and cache the historical keys and values as implicit memory to enable efficient streaming long-term 3D reconstruction. This design can handle low-latency 3D reconstruction by incrementally integrating historical information while maintaining high-quality spatial consistency. For efficient training, we propose to distill knowledge from the dense bidirectional visual geometry grounded transformer (VGGT) to our causal model. For inference, our model supports the migration of optimized efficient attention operators (e.g., FlashAttention) from large language models. Extensive experiments on various 3D geometry perception benchmarks demonstrate that our model enhances inference speed in online scenarios while maintaining competitive performance, thereby facilitating scalable and interactive 3D vision systems.

Core Task Landscape

This paper addresses: **Streaming 3D Geometry Reconstruction from Video**
A total of **50 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Monocular Video Reconstruction**
- **RGB-D and Depth-Based Reconstruction**
- **Stereo and Multi-View Reconstruction**
- **Streaming and Sequential Reconstruction Architectures**
- **Semantic and Open-Vocabulary 3D Reconstruction**
- **Urban and Large-Scale Scene Reconstruction**
- **Specialized Application Domains**
- **Facial and Human Body Reconstruction**
- **Foundational Techniques and System Surveys**

Complete Taxonomy Tree

- Streaming 3D Geometry Reconstruction from Video Survey Taxonomy
- Monocular Video Reconstruction
  - Real-Time Monocular Dense Reconstruction (6 papers)
  - [7] Neuralrecon: Real-time coherent 3d reconstruction from monocular video (Jiaming Sun, 2021) View paper
  - [17] Mobile3DRecon: Real-time monocular 3D reconstruction on a mobile phone (Xingbin Yang, 2020) View paper
  - [32] PanoRecon: Real-Time Panoptic 3D Reconstruction from Monocular Video (Dong Wu, 2024) View paper
  - [34] SLAM3R: Real-Time Dense Scene Reconstruction from Monocular RGB Videos (Yuzheng Liu, 2024) View paper
  - [37] EPRecon: An Efficient Framework for Real-Time Panoptic 3D Reconstruction from Monocular Video (Zhen Zhou, 2025) View paper
  - [44] MonoFusion: Real-time 3D reconstruction of small scenes with a single web camera (Vivek Pradeep, 2013) View paper
  - Dynamic and Deformable Object Reconstruction (5 papers)
  - [1] Shape of motion: 4d reconstruction from a single video (Qianqian Wang, 2025) View paper
  - [8] Lasr: Learning articulated shape reconstruction from a monocular video (Gengshan Yang, 2021) View paper
  - [13] Recovering dynamic 3d sketches from videos (Jaeah Lee, 2025) View paper
  - [36] Morphable 3D models from video (W. Brand, 2001) View paper
  - [47] Safari from visual signals: Recovering volumetric 3D shapes (Antonio Agudo, 2022) View paper
  - Specialized Domain Monocular Reconstruction (7 papers)
  - [9] Endo3r: Unified online reconstruction from dynamic monocular endoscopic video (Guo Jiaxin, 2025) View paper
  - [10] Video based reconstruction of 3d people models (Thiemo Alldieck, 2018) View paper
  - [19] Towards Robust 3D Reconstruction from Colonoscopy Video (Shu-xian, 2025) View paper
  - [26] Dynamic 3D avatar creation from hand-held video input (Alexandru Eugen Ichim, 2015) View paper
  - [28] Real-time 3D reconstruction of colonoscopic surfaces for determining missing regions (Ruibin Ma, 2019) View paper
  - [40] Reconstruction of personalized 3D face rigs from monocular video (Pablo Garrido, 2016) View paper
  - [42] 3D shape reconstruction from a humanoid generated video sequence (Pedro A. Martinez, 2014) View paper
- RGB-D and Depth-Based Reconstruction
  - Real-Time RGB-D Dense Reconstruction (3 papers)

    - [30] Real-time 3D reconstruction and 6-DoF tracking with an event camera (Hanme Kim, 2016) View paper
    - [33] Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera (Shahram Izadi, 2011) View paper
    - [49] Mobile3dscanner: An online 3d scanner for high-quality object reconstruction with a mobile device (Xiaojun Xiang, 2021) View paper
    - Dynamic RGB-D Reconstruction (2 papers)
    - [5] Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera (Kaiwen Guo, 2017) View paper
    - [6] Real-time non-rigid reconstruction using an RGB-D camera (Michael ZollhÃ¶fer, 2014) View paper
- Stereo and Multi-View Reconstruction
    - Real-Time Stereo Reconstruction (2 papers)
    - [12] Stereoscan: Dense 3d reconstruction in real-time (Andreas Geiger, 2011) View paper
    - [18] Real-time dense reconstruction of tissue surface from stereo optical video (Haoyin Zhou, 2019) View paper
    - Multi-Camera Rig Reconstruction (4 papers)
    - [22] On-the-fly Large-scale 3D Reconstruction from Multi-Camera Rigs (Yijia Guo, 2025) View paper
    - [27] Parallel processing for real-time 3D reconstruction from video streams (Tobias Duckworth, 2014) View paper
    - [43] Real-time active 3d shape reconstruction for 3d video (X. Wu, 2003) View paper
    - [45] Real-time dynamic 3-D object shape reconstruction and high-fidelity texture mapping for 3-D video (T. Matsuyama, 2004) View paper
    - Offline Multi-View Reconstruction (1 papers)
    - [20] Markerless shape and motion capture from multiview video sequences (Kun Li, 2011) View paper
- Streaming and Sequential Reconstruction Architectures ★ (6 papers)
    - [0] Streaming Visual Geometry Transformer (Anon et al., 2026) View paper
    - [3] Streaming 4d visual geometry transformer (Zhuo Dong, 2025) View paper
    - [25] StreamSplat: Towards Online Dynamic 3D Reconstruction from Uncalibrated Video Streams (Wu, 2025) View paper
    - [35] Streaming surface reconstruction from real time 3D measurements (Tim BodenmÃ¼ller, 2009) View paper
    - [41] STream3R: Scalable Sequential 3D Reconstruction with Causal Transformer (Lan, 2025) View paper
    - [46] ReCon-GS: Continuum-Preserved Gaussian Streaming for Fast and Compact Reconstruction of Dynamic Scenes (Gao Qiankun, 2025) View paper
- Semantic and Open-Vocabulary 3D Reconstruction (2 papers)
    - [4] EA3D: Online Open-World 3D Object Extraction from Streaming Videos (Zhou XiaoYu, 2025) View paper
    - [14] Ov3R: Open-Vocabulary Semantic 3D Reconstruction from RGB Videos (Li, 2025) View paper
- Urban and Large-Scale Scene Reconstruction (4 papers)
    - [2] Detailed real-time urban 3d reconstruction from video (Marc Pollefeys, 2008) View paper
    - [15] Real time localization and 3d reconstruction (E. Mouragnon, 2006) View paper
    - [23] Towards urban 3d reconstruction from video (Amir Akbarzadeh, 2006) View paper
    - [39] Efficient 3d reconstruction, streaming and visualization of static and dynamic scene parts for multi-client live-telepresence in large-scale environments (Leif Van Holland, 2023) View paper
- Specialized Application Domains (4 papers)
    - [11] Udr-gs: Enhancing underwater dynamic scene reconstruction with depth regularization (Yu Du, 2024) View paper
    - [16] Real-time 3-D video reconstruction for guidance of transventricular neurosurgery (Prasad Vagdargi, 2023) View paper
    - [31] Binocular video-based 3D reconstruction and length quantification of cracks in concrete structures (Lu Deng, 2023) View paper
    - [50] Self-Supervised Feature Detection and 3D Reconstruction for Real-Time Neuroendoscopic Guidance (Prasad Vagdargi, 2025) View paper
- Facial and Human Body Reconstruction (2 papers)
    - [24] Dense 3D face alignment from 2D video for real-time use (LÃ¡szlÃ³ A. Jeni, 2017) View paper
    - [29] Deep learning-based 3D face reconstruction method for video stream (Wenjun Yang, 2024) View paper
- Foundational Techniques and System Surveys (3 papers)
    - [21] Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video (Rui Yu, 2015) View paper
    - [38] Real-time 3D reconstruction techniques applied in dynamic scenes: A systematic literature review (Anupama K. Ingale, 2021) View paper
    - [48] Real-time cloud visual simultaneous localization and mapping for indoor service robots (Yali Zheng, 2020) View paper

Narrative

Core task: streaming 3D geometry reconstruction from video. The field addresses how to recover three-dimensional structure incrementally as video frames arrive, rather than waiting for complete sequences. The taxonomy reflects a rich landscape organized around input modality, architectural strategy, and application domain. Monocular Video Reconstruction tackles the challenge of depth ambiguity from single cameras, while RGB-D and Depth-Based Reconstruction exploits sensor data for denser, more immediate geometry (e.g., KinectFusion[33]). Stereo and Multi-View Reconstruction leverages multiple viewpoints to triangulate structure, and Urban and Large-Scale Scene Reconstruction focuses on city-scale environments where efficiency and scalability are paramount (Urban 3D Reconstruction[2]). Semantic and Open-Vocabulary 3D Reconstruction integrates language or category labels into geometry, and Specialized Application Domains span medical endoscopy (Endo3R[9], Colonoscopy Reconstruction[19]) and other niche settings. Facial and Human Body Reconstruction targets articulated or deformable subjects (Video People Models[10], Morphable Models Video[36]), while Foundational Techniques and System Surveys provide overarching methodological reviews (Dynamic Scenes Review[38]).

Within Streaming and Sequential Reconstruction Architectures, a central tension emerges between real-time responsiveness and reconstruction fidelity. Works such as StreamSplat[25] and STream3R[41] emphasize low-latency processing pipelines that update geometry on-the-fly, often trading off global consistency for immediate feedback. Streaming Visual Geometry[0] sits squarely in this branch, prioritizing incremental updates and efficient data structures that accommodate continuous frame arrival. Compared to Streaming 4D Geometry[3], which extends the problem to dynamic scenes with temporal deformation, Streaming Visual Geometry[0] focuses on static or slowly changing environments where sequential integration is the main challenge. Meanwhile, ReCon-GS[46] explores Gaussian splatting representations for streaming contexts, highlighting how representation choice—voxel grids, surfels, or splats—shapes both speed and quality. These architectural decisions remain an active area of exploration, balancing the need for real-time operation against the desire for high-fidelity, globally consistent reconstructions.

Related Works in Same Category

The following **5 sibling papers** share the same taxonomy leaf node with the original paper:

1. Streaming 4d visual geometry transformer

Authors: Zhuo Dong, Zheng, Wenzhao, Dong Zhuo, Guo Jiahe, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Perceiving and reconstructing 4D spatial-temporal geometry from videos is a fundamental yet challenging computer vision task. To facilitate interactive and real-time applications, we propose a streaming 4D visual geometry transformer that shares a similar philosophy with autoregressive large language models. We explore a simple and efficient design and employ a causal transformer architecture to process the input sequence in an online manner. We use temporal causal attention and cache the histor...

⚠ Similarity Notice

These papers share nearly identical titles ('Streaming Visual Geometry Transformer' vs. 'Streaming 4d visual geometry transformer'), identical architectural designs (temporal causal attention with cached memory tokens), identical methodology (distillation from VGGT teacher), and identical experimental setups. The candidate paper appears to be a published or revised version of the original submission, with minor differences in terminology (3D vs. 4D reconstruction) but essentially the same technical contribution and system.

---

2. StreamSplat: Towards Online Dynamic 3D Reconstruction from Uncalibrated Video Streams

Authors: Wu, Zike, Yan Qi, Zike Wu, Yi, et al. (13 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Real-time reconstruction of dynamic 3D scenes from uncalibrated video streams is crucial for numerous real-world applications. However, existing methods struggle to jointly address three key challenges: 1) processing uncalibrated inputs in real time, 2) accurately modeling dynamic scene evolution, and 3) maintaining long-term stability and computational efficiency. To this end, we introduce StreamSplat, the first fully feed-forward framework that transforms uncalibrated video streams of arbitrar...

Relationship Analysis

Both papers belong to the streaming and sequential reconstruction architectures category, employing causal or online processing for incremental 3D reconstruction from video. They overlap in their use of transformer-based architectures with temporal attention mechanisms and cached memory to enable low-latency streaming reconstruction. However, the original paper (StreamVGGT) focuses on distilling knowledge from a bidirectional teacher model to predict camera poses, depth, and point maps using temporal causal attention with implicit token memory, while the candidate paper (StreamSplat) targets dynamic 3D Gaussian Splatting reconstruction with a bidirectional deformation field to explicitly model scene dynamics and motion from uncalibrated video streams.

---

3. Streaming surface reconstruction from real time 3D measurements

Authors: Tim BodenmÃ¼ller | **Year/Venue**: 2009 | **URL**: View paper

Abstract

In dieser Dissertation wird eine robuste Methode zur schnellen, inkrementellen Erstellung von 3D OberflÃ¤chenmodellen aus Echtzeit-DatenstrÃ¶men vorgestellt. Das Ziel ist, dem Benutzer eines manuellen Scannersystems das entstehende 3D Modell wÃ¤hrend des Scannens als visuelle RÃ¼ckkopplung sofort anzuzeigen. Indem kontinuierlich Messpunkte aus dem Echtzeit-Datenstrom eingefÃ¼gt werden und damit das bestehende Netz lokal verfeinert wird, kann mit der prÃ¤sentierten Methode iterativ ein dichtes un...

Relationship Analysis

Both papers belong to the streaming and sequential reconstruction architectures category, focusing on online incremental 3D reconstruction from video streams. They overlap in their core objective of processing streaming 3D data incrementally without requiring full sequence reprocessing. However, the original paper (StreamVGGT) employs a modern transformer-based architecture with temporal causal attention and cached memory tokens inspired by large language models, while the candidate paper presents a classical geometric approach using iterative triangle mesh refinement with spatial data structures for real-time manual scanner systems.

---

4. STream3R: Scalable Sequential 3D Reconstruction with Causal Transformer

Authors: Lan, Yushi, Luo, Yihang, Yushi Lan, et al. (28 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

We present STream3R, a novel approach to 3D reconstruction that reformulates pointmap prediction as a decoder-only Transformer problem. Existing state-of-the-art methods for multi-view reconstruction either depend on expensive global optimization or rely on simplistic memory mechanisms that scale poorly with sequence length. In contrast, STream3R introduces an streaming framework that processes image sequences efficiently using causal attention, inspired by advances in modern language modeling. ...

Relationship Analysis

Both papers belong to the Streaming and Sequential Reconstruction Architectures category, employing causal transformer designs for online 3D reconstruction from video streams. They share overlapping approaches in using temporal causal attention mechanisms and cached memory tokens to enable incremental processing without reprocessing entire sequences. The key differences are: StreamVGGT uses knowledge distillation from a bidirectional teacher model (VGGT) to train its causal student architecture, while STream3R adopts a decoder-only transformer design inspired by LLMs with KVCache, training end-to-end on large-scale 3D datasets without distillation.

---

5. ReCon-GS: Continuum-Preserved Gaussian Streaming for Fast and Compact Reconstruction of Dynamic Scenes

Authors: Gao Qiankun, Jiaye Fu, Qiankun Gao, Wu Yanmin, Chengxiang Wen, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Online free-viewpoint video (FVV) reconstruction is challenged by slow per-frame optimization, inconsistent motion estimation, and unsustainable storage demands. To address these challenges, we propose the Reconfigurable Continuum Gaussian Stream, dubbed ReCon-GS, a novel storage-aware framework that enables high fidelity online dynamic scene reconstruction and real-time rendering. Specifically, we dynamically allocate multi-level Anchor Gaussians in a density-adaptive fashion to capture inter-f...

Relationship Analysis

Both papers belong to the streaming and sequential reconstruction architectures category, employing causal or online frameworks for incremental 3D reconstruction from video. They overlap in addressing low-latency streaming reconstruction with temporal consistency, but differ fundamentally in their approach: the original paper (StreamVGGT) uses a causal transformer with cached token memory for multi-view geometry reconstruction from RGB video, while the candidate paper (ReCon-GS) focuses on 3D Gaussian Splatting with hierarchical motion representation and dynamic reconfiguration for free-viewpoint video synthesis. The original emphasizes transformer-based dense correspondence and multi-task prediction (depth, pose, tracking), whereas the candidate prioritizes storage-efficient deformation modeling and real-time rendering through anchor-driven motion decomposition.

Contributions Analysis

Overall novelty summary. The paper proposes a streaming visual geometry transformer that processes video frames incrementally using temporal causal attention and cached historical keys/values. It resides in the 'Streaming and Sequential Reconstruction Architectures' leaf, which contains six papers total (including this one). This leaf represents a relatively focused research direction within the broader taxonomy of fifty papers spanning nine major categories. The concentration of only six papers in this leaf suggests a moderately sparse area compared to more crowded domains like Monocular Video Reconstruction, which encompasses multiple subtopics and over fifteen papers.

The taxonomy tree reveals that neighboring research directions include Monocular Video Reconstruction (with real-time dense reconstruction and dynamic object modeling), RGB-D and Depth-Based Reconstruction (leveraging sensor data for immediate geometry), and Stereo and Multi-View Reconstruction (exploiting multiple viewpoints). The paper's streaming architecture distinguishes it from these sensor-modality-focused branches by emphasizing causal processing and autoregressive design principles borrowed from large language models. While sibling papers like StreamSplat and STream3R also target low-latency incremental updates, the paper's explicit use of transformer-based causal attention and knowledge distillation from bidirectional models positions it at the intersection of streaming reconstruction and modern deep learning architectures.

Among twenty candidates examined across three contributions, five refutable pairs were identified. The streaming transformer with temporal causal attention examined ten candidates and found two potentially overlapping works. The cached historical token memory mechanism examined nine candidates with two refutable matches. The knowledge distillation training strategy examined only one candidate, which appeared to provide prior work. These statistics indicate that within the limited search scope, each contribution encounters some degree of prior overlap, though the majority of examined candidates (fifteen out of twenty) did not clearly refute the contributions. The relatively small candidate pool means the analysis captures top semantic matches rather than exhaustive coverage.

Given the limited search scope of twenty candidates, the paper appears to operate in a moderately explored area where streaming architectures for 3D reconstruction are gaining traction but remain less saturated than traditional SLAM or RGB-D fusion methods. The contribution-level statistics suggest incremental novelty rather than entirely uncharted territory, with each technical component finding at least one overlapping prior work among the examined candidates. A more exhaustive literature search would be necessary to assess whether the specific combination of causal transformers, cached memory, and distillation training constitutes a genuinely novel synthesis or a natural extension of existing streaming reconstruction paradigms.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Streaming visual geometry transformer with temporal causal attention

Description: The authors introduce StreamVGGT, a causal transformer architecture that replaces global self-attention with temporal causal attention to enable efficient, low-latency streaming 3D reconstruction. This design allows incremental processing of video frames without reprocessing entire sequences.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Evict3R: Training-Free Token Eviction for Memory-Bounded Streaming Visual Geometry Transformers

URL: View paper

Brief Assessment

Evict3R[56] focuses on memory management through token eviction for an existing streaming architecture (StreamVGGT), not on proposing the causal transformer architecture itself. The candidate paper explicitly builds upon StreamVGGT's temporal causal attention design rather than claiming to introduce it.

---

2. Causal Motion Tokenizer for Streaming Motion Generation

URL: View paper

Brief Assessment

Causal Motion Tokenizer[51] focuses on human motion generation from text using causal convolution in a motion tokenizer, not on streaming 3D reconstruction from video. The domains and tasks are fundamentally different.

---

3. Stargen: A spatiotemporal autoregression framework with video diffusion model for scalable and controllable scene generation

URL: View paper

Brief Assessment

StarGen[54] focuses on scene generation using video diffusion models with spatiotemporal autoregression, not on streaming 3D reconstruction from video with causal transformers for geometry perception.

---

4. Occworld: Learning a 3d occupancy world model for autonomous driving

URL: View paper

Brief Assessment

OccWorld[52] focuses on 3D occupancy prediction for autonomous driving using spatial-temporal transformers, not on streaming 3D reconstruction from video with causal attention for geometry tasks.

---

5. Ar4d: Autoregressive 4d generation from monocular videos

URL: View paper

Brief Assessment

AR4D[53] focuses on 4D generation from monocular videos using autoregressive generation of 3D representations frame-by-frame, not on streaming 3D reconstruction with causal transformers for real-time geometry perception.

---

6. Occsora: 4d occupancy generation models as world simulators for autonomous driving

URL: View paper

Brief Assessment

OccSora[55] focuses on 4D occupancy generation for autonomous driving using diffusion models, not streaming 3D reconstruction with causal transformers. The candidate addresses a different problem domain (world simulation vs. incremental geometry reconstruction) and employs fundamentally different architectures (diffusion-based generation vs. causal attention for streaming).

---

7. Streaming 4d visual geometry transformer

URL: View paper

Prior Art Analysis

Streaming 4D Geometry[3] demonstrates that the concept of using a causal transformer architecture with temporal causal attention for streaming 3D/4D reconstruction was already proposed. Both papers employ nearly identical architectural designs: replacing global self-attention with temporal causal attention to enable efficient, incremental processing of video frames. The candidate paper explicitly states it uses 'temporal causal attention' and processes inputs 'in an online manner' with a 'causal transformer architecture,' which directly matches the original paper's claimed novelty. Large portions of the methodology descriptions are nearly identical, suggesting the original paper was not the first to propose this contribution.

Evidence

Evidence 1 - **Rationale**: These passages are nearly identical, demonstrating that Streaming 4D Geometry[3] already proposed the same causal transformer architecture with temporal causal attention for streaming reconstruction, refuting the original paper's novelty claim. - **Original**: we propose StreamVGGT, a causal transformer architecture specifically designed for efficient, low-latency streaming 3D visual geometry reconstruction, as shown in figure 1. Unlike conventional offline frameworks that necessitate reprocessing the entire sequence with the arrival of each new frame, we... - **Candidate**: we propose streamvggt, a causal transformer architecture specifically designed for efficient, real-time streaming 4d visual geometry reconstruction, as shown in figure 1. unlike conventional offline frameworks that necessitate reprocessing the entire sequence with the arrival of each new frame, we i...

Evidence 2 - **Rationale**: The mathematical formulation and description of temporal causal attention are identical in both papers, confirming that Streaming 4D Geometry[3] already implemented this exact mechanism. - **Original**: we utilize a causal transformer architecture to explicitly model the causal structure intrinsic for streaming data. Specifically, for our streamvggt: {gt}t t=1 = decoder(temporal selfattn({ft}t t=1)). (5) the temporal causal attention restricts each frame to attend only to itself and its predecessor... - **Candidate**: we utilize a causal transformer architecture to explicitly model the causal structure intrinsic for streaming data. specifically, for our streamvggt: {gt}t t=1 = decoder(temporal selfattn({ft}t t=1)). (5) the temporal causal attention restricts each frame to attend only to itself and its predecessor...

---

8. MOGO: Residual Quantized Hierarchical Causal Transformer for High-Quality and Real-Time 3D Human Motion Generation

URL: View paper

Brief Assessment

MOGO[58] focuses on 3D human motion generation from text using hierarchical causal transformers for motion token sequences, not streaming 3D reconstruction from video frames. The causal attention mechanism serves a different purpose (motion token generation) in a different domain (text-to-motion) compared to the original paper's video-based 3D geometry reconstruction.

---

9. Motionstream: Real-time video generation with interactive motion controls

URL: View paper

Brief Assessment

MotionStream[57] focuses on real-time video generation with motion controls using causal attention for streaming synthesis, not 3D reconstruction from video. The architectural goals and application domains are fundamentally different.

---

10. STream3R: Scalable Sequential 3D Reconstruction with Causal Transformer

URL: View paper

Prior Art Analysis

STream3R[41] demonstrates that a causal transformer architecture for streaming 3D reconstruction was proposed prior to the original paper. Both papers introduce causal attention mechanisms that replace global self-attention to enable efficient streaming 3D reconstruction from video sequences. STream3R[41] explicitly describes using 'causal attention' where 'each token is restricted to attend only to the current and previous frames in the sequence' and implements a 'cached memory token' mechanism. The original paper's StreamVGGT uses 'temporal causal attention' with 'cached token memory' for the same purpose. The architectural designs, problem formulations, and technical implementations are nearly identical, indicating that STream3R[41] proposed this approach before the original submission.

Evidence

Evidence 1 - **Rationale**: Both papers introduce causal transformer architectures for streaming 3D reconstruction. STream3R[41] explicitly states it 'processes image sequences efficiently using causal attention' for streaming 3D reconstruction, which directly parallels the original paper's 'temporal causal attention mechanism' for 'streaming 3d visual geometry reconstruction.' - **Original**: we propose streamvggt, a causal transformer architecture specifically designed for efficient, low-latency streaming 3d visual geometry reconstruction, as shown in figure 1. unlike conventional offline frameworks that necessitate reprocessing the entire sequence with the arrival of each new frame, we... - **Candidate**: we present stream 3r, a novel approach to 3d reconstruction that reformulates pointmap prediction as a decoder-only transformer problem. existing state-ofthe-art methods for multi-view reconstruction either depend on expensive global optimization or rely on simplistic memory mechanisms that scale po...

Evidence 2 - **Rationale**: Both papers describe replacing global self-attention with causal attention that restricts tokens to attend only to current and previous frames. The technical implementation and motivation are identical - enabling efficient streaming 3D reconstruction by respecting temporal causality. - **Original**: we introduce the spatio-temporal decoder by replacing all the global self-attention layers with temporal attention layers. in the standard global self-attention mechanism, each image token ft attends to all other tokens in the sequence, which can result in high computational costs when handling long... - **Candidate**: as mentioned in sec. 3, given the streaming inputs, for each current image, it, our method first tokenizes it into the features ft = encoder(it). the main difference lies in the decoder side: rather than performing bi-directional attention over the whole sequence [12] or interacting with a learnable...

Evidence 3 - **Rationale**: Both papers implement memory caching mechanisms for streaming inference. The original paper's 'implicit memory mechanism that caches historical token' directly corresponds to STream3R[41]'s 'kv cache during inference' - both enable efficient incremental processing without recomputing previous frames. - **Original**: cached memory token. unlike the training phase, where all frames are input simultaneously and processed with temporal attention, streaming 3d reconstruction requires the model to handle frame-by-frame input and perform incremental 3d reconstruction during inference. to address this, we introduce an ... - **Candidate**: causal attention for long-context 3d reasoning. as mentioned in sec. 3, given the streaming inputs, for each current image, it, our method first tokenizes it into the features ft = encoder(it). the main difference lies in the decoder side: rather than performing bi-directional attention over the who...

Contribution 2: Cached historical token memory mechanism

Description: The authors propose caching historical keys and values as implicit memory tokens, allowing the model to maintain long-term context while supporting efficient incremental reconstruction. This mechanism enables the model to replicate temporal causal attention behavior during streaming inference.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. KVzip: Query-Agnostic KV Cache Compression with Context Reconstruction

URL: View paper

Brief Assessment

KVzip[59] focuses on compressing KV caches for LLMs to reduce memory overhead during inference, not on caching historical keys and values for incremental 3D reconstruction in visual geometry transformers. The technical domains and applications are fundamentally different.

---

2. Zsmerge: Zero-shot kv cache compression for memory-efficient long-context llms

URL: View paper

Brief Assessment

ZSMerge[61] focuses on KV cache compression for LLMs through token merging and eviction strategies, not on streaming visual geometry reconstruction with temporal causal attention for 3D tasks.

---

3. Streammem: Query-agnostic kv cache memory for streaming video understanding

URL: View paper

Prior Art Analysis

StreamMem[62] demonstrates that caching historical keys and values as implicit memory tokens for efficient incremental reconstruction was already proposed and implemented prior to the original paper. Both papers employ the same fundamental mechanism: caching key-value pairs from previously processed frames to maintain long-term context while supporting streaming inference. StreamMem[62] explicitly describes caching the kv cache using attention scores and maintaining a fixed-size memory, which directly parallels the original paper's claimed contribution of 'caching historical keys and values as implicit memory tokens' for 'efficient incremental reconstruction.'

Evidence

Evidence 1 - **Rationale**: Both papers propose caching key-value pairs as memory for streaming processing. StreamMem[62] explicitly describes maintaining a 'fixed-size kv memory' for 'streaming video understanding,' which directly corresponds to the original paper's claim of caching 'historical keys and values as implicit memory' for 'efficient streaming long-term 3d reconstruction.' - **Original**: we use temporal causal attention and cache the historical keys and values as implicit memory to enable efficient streaming long-term 3d reconstruction. this design can handle low-latency 3d reconstruction by incrementally integrating historical information while maintaining high-quality spatial cons... - **Candidate**: in this work, we proposestreammem, aquery-agnostic kv cache memory mechanism for streaming video understanding. specifically, streammem encodes new video frames in a streaming manner, compressing the kv cache using attention scores between visual tokens and generic query tokens, while maintaining a ...

Evidence 2 - **Rationale**: Both papers describe the same incremental processing mechanism where current frame tokens attend to cached memory from previous frames. StreamMem[62]'s approach of merging 'newly computed kvs' with 'compressed kv memory from the previous time step' demonstrates the same cached memory mechanism for incremental reconstruction claimed by the original paper. - **Original**: during inference, streamvggt performs cross attention between the cached memory tokens and the image tokens derived from the current frame: gt = decoder(crossattn(ft , {mt}t-1 t=1)), mt = tokencachedmemory(gt). this design enables the model to replicate the temporal causal attention behavior obser... - **Candidate**: at each time stept, a new segment of frames is received from the video stream. these frames first undergo an input filtering step to remove temporal redundancy. the filtered frames are then encoded by the vision encoder and processed by the mllm to produce key-value (kv) representations {ki t, vi t ...

---

4. Longlive: Real-time interactive long video generation

URL: View paper

Brief Assessment

LongLive[60] focuses on video generation with KV-cache for frame-level autoregressive models, while the original paper addresses 3D geometry reconstruction from streaming visual inputs. The technical domains and applications are fundamentally different.

---

5. FAEDKV: Infinite-Window Fourier Transform for Unbiased KV Cache Compression

URL: View paper

Brief Assessment

FAEDKV[64] focuses on frequency-domain KV cache compression using Fourier transforms for LLMs, not on streaming 3D visual geometry reconstruction with temporal causal attention for incremental scene updates.

---

6. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft

URL: View paper

Brief Assessment

Memory Forcing[63] focuses on spatial memory for Minecraft scene generation using 3D reconstruction and point-to-frame retrieval, not on caching keys/values as implicit memory tokens for temporal causal attention in streaming 3D reconstruction tasks.

---

7. Streaming 4d visual geometry transformer

URL: View paper

Prior Art Analysis

Streaming 4D Geometry[3] demonstrates that the cached historical token memory mechanism was already proposed. Both papers describe an identical approach where historical keys and values are cached as implicit memory tokens to enable efficient incremental reconstruction. The candidate paper explicitly implements the same 'implicit memory mechanism that caches historical token' and uses it to 'replicate the temporal causal attention behavior' during streaming inference. The descriptions and implementations are nearly word-for-word identical, indicating the original paper was not the first to propose this mechanism.

Evidence

Evidence 1 - **Rationale**: The abstract descriptions of the cached memory mechanism are nearly identical, showing that Streaming 4D Geometry[3] already proposed caching historical keys and values as implicit memory. - **Original**: we use temporal causal attention and cache the historical keys and values as implicit memory to enable efficient streaming long-term 3d reconstruction. this design can handle low-latency 3d reconstruction by incrementally integrating historical information while maintaining high-quality spatial cons... - **Candidate**: we use temporal causal attention and cache the historical keys and values as implicit memory to enable efficient streaming long-term 4d reconstruction. this design can handle realtime 4d reconstruction by incrementally integrating historical information while maintaining high-quality spatial consist...

Evidence 2 - **Rationale**: The detailed description of the cached memory token mechanism is identical in both papers, confirming prior work exists. - **Original**: unlike the training phase, where all frames are input simultaneously and processed with temporal attention, streaming 3d reconstruction requires the model to handle frame-by-frame input and perform incremental 3d reconstruction during inference. to address this, we introduce an implicit memory mecha... - **Candidate**: unlike the training phase, where all frames are input

simultaneously and processed with temporal attention, streaming 4d reconstruction requires the model to handle frame-by-frame input and perform incremental 4d reconstruction during inference. to address this, we introduce an implicit memory mecha...

Evidence 3 - **Rationale**: The validation and purpose of the cached memory mechanism are described identically, demonstrating that Streaming 4D Geometry[3] already implemented and validated this approach. - **Original**: this design enables the model to replicate the temporal causal attention behavior observed during training. through experimental validation, we demonstrate that the model with cached memory token achieves performance comparable to that of full-sequence input inference. this confirms the effectivenes... - **Candidate**: this design enables the model to replicate the temporal causal attention behavior observed during training. through experimental validation, we demonstrate that the model with cached memory token achieves performance comparable to that of full-sequence input inference. this confirms the effectivenes...

---

8. LaCache: Ladder-Shaped KV Caching for Efficient Long-Context Modeling of Large Language Models

URL: View paper

Brief Assessment

LaCache[66] focuses on optimizing KV cache storage patterns for LLMs through ladder-shaped caching and iterative compaction, while the original paper caches historical keys and values as implicit memory tokens for 3D visual geometry reconstruction with temporal causal attention in streaming video processing.

---

9. MEDA: Dynamic KV Cache Allocation for Efficient Multimodal Long-Context Inference

URL: View paper

Brief Assessment

MEDA[65] focuses on KV cache compression for multimodal LLMs through dynamic allocation and merging strategies, not on caching historical tokens as implicit memory for incremental 3D reconstruction with temporal causal attention.

Contribution 3: Knowledge distillation training strategy from bidirectional teacher

Description: The authors develop a distillation-based training approach that transfers geometric understanding from the bidirectional VGGT teacher to the causal student model. This strategy unifies multi-task supervision through teacher-generated pseudo-ground truth, reducing training costs while maintaining high accuracy.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Streaming 4d visual geometry transformer

URL: View paper

Prior Art Analysis

Streaming 4D Geometry[3] demonstrates that the knowledge distillation training strategy from a bidirectional VGGT teacher to a causal student model was already proposed. Both papers describe an identical distillation approach where the densely connected, bidirectional VGGT serves as the teacher model to transfer geometric understanding to the causal student. The candidate paper explicitly states it uses 'knowledge distillation from the densely connected, bidirectional visual geometry grounded transformer (vggt) as the teacher model' and that this strategy 'helps mitigate error accumulation.' The methodology descriptions are nearly identical, indicating the original paper was not the first to propose this training strategy.

Evidence

Evidence 1 - **Rationale**: Both papers describe using a bidirectional teacher model with global attention to guide a causal student model through knowledge distillation, demonstrating prior work exists. - **Original**: to reduce training cost while retaining high accuracy, we adopt a distillation-based training strategy in which the densely connected, bidirectional vggt (wang et al., 2025a) serves as the teacher. by transferring its global geometric understanding, our causal student model achieves performance comp... - **Candidate**: to address this, we introduce a knowledge distillation strategy, where a teacher model uses global attention to guide the causal student model. the teacher processes all frames in the sequence to captures richer context, while the student is limited to the current and previous frames. the use of kno...

Evidence 2 - **Rationale**: The experimental validation of the distillation strategy is described identically, demonstrating that Streaming 4D Geometry[3] already validated this approach. - **Original**: to assess the effectiveness of our knowledge-distillation strategy, we evaluated three variants on the 7-scenes (shotton et al., 2013), nrgbd (azinovi´c et al., 2022), and eth3d (schops et al., 2017) datasets: (i) the global self-attention teacher vggt (wang et al., 2025a), (ii) streamvggt without k... - **Candidate**: to assess the effectiveness of our knowledge-distillation strategy, we evaluate three variants on the 7-scenes, nrgbd, and eth3d datasets: (i) the global selfattention teacher vggt, (ii) streamvggt without knowledge distillation, and (iii) streamvggt with knowledge distillation.

Appendix: Text Similarity Detection

Textual similarity detection checked 22 papers and found 6 similarity segment(s) across 2 paper(s).

The following **2 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. STream3R: Scalable Sequential 3D Reconstruction with Causal Transformer

Detected in: Core Task (sibling), Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

2. Streaming 4d visual geometry transformer

Detected in: Core Task (sibling), Contribution: contribution_1, Contribution: contribution_2, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Streaming Visual Geometry Transformer View paper
- [1] Shape of motion: 4d reconstruction from a single video View paper
- [2] Detailed real-time urban 3d reconstruction from video View paper
- [3] Streaming 4d visual geometry transformer View paper
- [4] EA3D: Online Open-World 3D Object Extraction from Streaming Videos View paper
- [5] Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera View paper

- [6] Real-time non-rigid reconstruction using an RGB-D camera View paper
- [7] Neuralrecon: Real-time coherent 3d reconstruction from monocular video View paper
- [8] Lasr: Learning articulated shape reconstruction from a monocular video View paper
- [9] Endo3r: Unified online reconstruction from dynamic monocular endoscopic video View paper
- [10] Video based reconstruction of 3d people models View paper
- [11] Udr-gs: Enhancing underwater dynamic scene reconstruction with depth regularization View paper
- [12] Stereoscan: Dense 3d reconstruction in real-time View paper
- [13] Recovering dynamic 3d sketches from videos View paper
- [14] Ov3R: Open-Vocabulary Semantic 3D Reconstruction from RGB Videos View paper
- [15] Real time localization and 3d reconstruction View paper
- [16] Real-time 3-D video reconstruction for guidance of transventricular neurosurgery View paper
- [17] Mobile3DRecon: Real-time monocular 3D reconstruction on a mobile phone View paper
- [18] Real-time dense reconstruction of tissue surface from stereo optical video View paper
- [19] Towards Robust 3D Reconstruction from Colonoscopy Video View paper
- [20] Markerless shape and motion capture from multiview video sequences View paper
- [21] Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video View paper
- [22] On-the-fly Large-scale 3D Reconstruction from Multi-Camera Rigs View paper
- [23] Towards urban 3d reconstruction from video View paper
- [24] Dense 3D face alignment from 2D video for real-time use View paper
- [25] StreamSplat: Towards Online Dynamic 3D Reconstruction from Uncalibrated Video Streams View paper
- [26] Dynamic 3D avatar creation from hand-held video input View paper
- [27] Parallel processing for real-time 3D reconstruction from video streams View paper
- [28] Real-time 3D reconstruction of colonoscopic surfaces for determining missing regions View paper
- [29] Deep learning-based 3D face reconstruction method for video stream View paper
- [30] Real-time 3D reconstruction and 6-DoF tracking with an event camera View paper
- [31] Binocular video-based 3D reconstruction and length quantification of cracks in concrete structures View paper
- [32] PanoRecon: Real-Time Panoptic 3D Reconstruction from Monocular Video View paper
- [33] Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera View paper
- [34] SLAM3R: Real-Time Dense Scene Reconstruction from Monocular RGB Videos View paper
- [35] Streaming surface reconstruction from real time 3D measurements View paper
- [36] Morphable 3D models from video View paper
- [37] EPRecon: An Efficient Framework for Real-Time Panoptic 3D Reconstruction from Monocular Video View paper
- [38] Real-time 3D reconstruction techniques applied in dynamic scenes: A systematic literature review View paper
- [39] Efficient 3d reconstruction, streaming and visualization of static and dynamic scene parts for multi-client live-telepresence in large-scale environments View paper
- [40] Reconstruction of personalized 3D face rigs from monocular video View paper
- [41] STream3R: Scalable Sequential 3D Reconstruction with Causal Transformer View paper
- [42] 3D shape reconstruction from a humanoid generated video sequence View paper
- [43] Real-time active 3d shape reconstruction for 3d video View paper
- [44] MonoFusion: Real-time 3D reconstruction of small scenes with a single web camera View paper
- [45] Real-time dynamic 3-D object shape reconstruction and high-fidelity texture mapping for 3-D video View paper
- [46] ReCon-GS: Continuum-Preserved Gaussian Streaming for Fast and Compact Reconstruction of Dynamic Scenes View paper
- [47] Safari from visual signals: Recovering volumetric 3D shapes View paper
- [48] Real-time cloud visual simultaneous localization and mapping for indoor service robots View paper
- [49] Mobile3dscanner: An online 3d scanner for high-quality object reconstruction with a mobile device View paper
- [50] Self-Supervised Feature Detection and 3D Reconstruction for Real-Time Neuroendoscopic Guidance View paper
- [51] Causal Motion Tokenizer for Streaming Motion Generation View paper
- [52] Occworld: Learning a 3d occupancy world model for autonomous driving View paper
- [53] Ar4d: Autoregressive 4d generation from monocular videos View paper
- [54] Stargen: A spatiotemporal autoregression framework with video diffusion model for scalable and controllable scene generation View paper
- [55] Occsora: 4d occupancy generation models as world simulators for autonomous driving View paper
- [56] Evict3R: Training-Free Token Eviction for Memory-Bounded Streaming Visual Geometry Transformers View paper
- [57] Motionstream: Real-time video generation with interactive motion controls View paper
- [58] MOGO: Residual Quantized Hierarchical Causal Transformer for High-Quality and Real-Time 3D Human Motion Generation View paper
- [59] KVzip: Query-Agnostic KV Cache Compression with Context Reconstruction View paper
- [60] Longlive: Real-time interactive long video generation View paper
- [61] Zsmerge: Zero-shot kv cache compression for memory-efficient long-context llms View paper
- [62] Streammem: Query-agnostic kv cache memory for streaming video understanding View paper
- [63] Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft View paper
- [64] FAEDKV: Infinite-Window Fourier Transform for Unbiased KV Cache Compression View paper
- [65] MEDA: Dynamic KV Cache Allocation for Efficient Multimodal Long-Context Inference View paper
- [66] LaCache: Ladder-Shaped KV Caching for Efficient Long-Context Modeling of Large Language Models View paper
- [67] Beyond homogeneous attention: Memory-efficient llms via fourier-approximated kv cache View paper