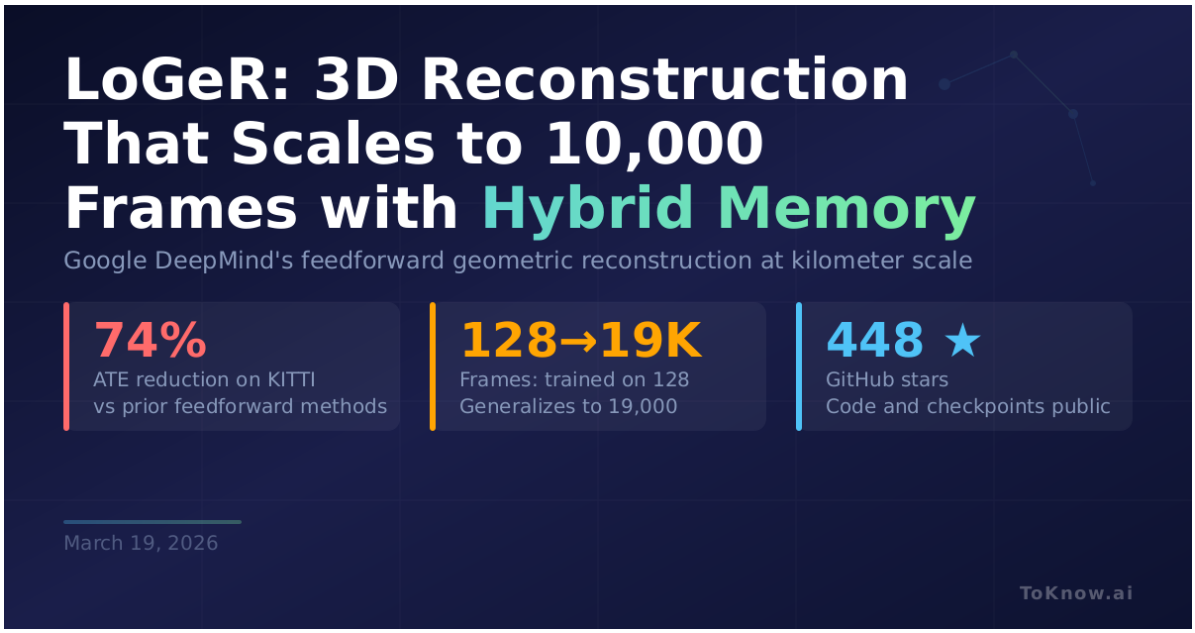
LoGeR: DeepMind's 3D Reconstruction That Scales to 10,000 Frames with Hybrid Memory

Kabui, Charles

2026-03-19



Google DeepMind's LoGeR (Long-context Geometric Reconstruction) scales dense 3D reconstruction from video to over 10,000 frames and kilometer-long sequences, all without any post-processing optimization. Current 3D reconstruction models face a hard tradeoff: bidirectional attention models like VGGT produce excellent local geometry but choke on long videos

due to quadratic compute costs, while recurrent models scale linearly but lose geometric coherence over time. LoGeR breaks this tradeoff with a hybrid memory module that combines two strategies. A parametric memory (Test-Time Training) adapts model weights on the fly to anchor the global coordinate frame and prevent scale drift. A non-parametric memory (Sliding Window Attention) keeps recent frames in full detail for precise alignment between consecutive chunks. Trained on just 128-frame sequences, LoGeR generalizes to thousands of frames at inference. On KITTI, it cuts Absolute Trajectory Error by over 74% compared to prior feedforward methods, reaching an ATE of 18.65. On the VBR dataset (up to 19,000 frames), it delivers a 30.8% improvement over previous best results.

For anyone working in autonomous driving, robotics, or AR/VR, this is a practical step forward. Reconstructing a full driving route from dashcam footage, or building a persistent spatial map of an office from a walkthrough video, previously required expensive optimization pipelines. LoGeR does it in a single feedforward pass with linear-time scaling. The code and checkpoints are public.

Where the Seoul World Model generates video of real cities, LoGeR reconstructs their 3D geometry from video. Together, they represent two sides of the same shift: vision models that operate at city scale, not clip scale.

Sources:

- LoGeR Paper (arXiv)
- LoGeR Project Page
- LoGeR GitHub Repository
- HuggingFace Daily Papers (March 10, 2026)
