11 Strengths & Limitations

Critical Analysis

Core Strengths

Minimal Modeling Strategy SCALABLE

Single plain ViT + depth-ray targets avoid complex bespoke architectures. Inherits scaling laws directly from DINOv2 pretraining.

Unified "Any-View" Framework

Seamlessly handles monocular, multi-view, and video inputs. Pose-optional design bridges the gap between uncalibrated images and metric 3D.

SOTA Geometry & Pose

Outperforms VGGT by huge margins (+35.7% Pose AUC). Provides robust foundation for feed-forward novel view synthesis (3DGS).

Limitations & Challenges

Computational Cost at Inference OPTIMIZATION

Extracting pose from ray maps via optimization can be slow. Pose head (DC) mitigates this but adds dependency on tokens.

Static Scene Limitation

The depth-ray formulation does not explicitly model motion or deformation fields, suggesting **POTENTIAL limitations for dynamic scenes**. *Note: The paper does not evaluate dynamic scene performance, so this remains an open question for future work.*

Memory Scaling

Trained on 2-18 views; beyond this requires memory optimization. Token budgeting strategies needed for large-scale multi-view processing.