2 Introduction & Motivation

The Core Problem

Existing foundation models require:

- ❌ **VGGT:** Redundant multi-task targets (point maps + depth + pose) + multi-stage architecture
- ❌ **DUSt3R:** Point maps insufficient for metric consistency
- ❌ **Traditional SfM/MVS:** Brittle under textureless regions, specularities

Central Question:

Can a **SINGLE PLAIN ViT** with **MINIMAL TARGETS** (depth + rays only) suffice for unified 3D reconstruction?

YES → DepthAnything3

Key Achievement: DA3 reaches SOTA performance using **only public academic datasets**, without proprietary data.

Limitations of Prior Work

Traditional Pipelines (SfM / MVS / SLAM)

Modular systems are brittle under textureless regions, specularities, or large baselines.

Early Deep Learning Approaches

Fixed input cardinality, complex architectures, and limited scalability.

Current Foundation Models (e.g., VGGT)

Rely on multi-head task bundles, redundant targets, and heavy design overhead.

DA3 Solution Highlights

Unified Representation

Depth + Ray formulation handles arbitrary inputs (single image, stereo, video) in one model.

Minimal Architecture

Single plain ViT backbone—no complex multi-stage pipelines or redundant heads.

SOTA Performance

+35.7% pose accuracy improvement, trained only on public datasets.