Junlin Han, Filippos Kokkinos, Philip Torr

GenAI, Meta and TVG, University of Oxford

Code

Demo

TL;DR

VFusion3D is a large, feed-forward 3D reconstruction/generative model trained with a small amount of 3D data and a large volume of synthetic multi-view data.

Motivation

Problem: The primary obstacle in developing foundation 3D generative models is the limited availability of 3D data, which are not readily accessible and are difficult to acquire.

Solution: Can we leverage other generative models to provide synthetic 3D data?

We propose using a video diffusion model, trained with extensive volumes of text, images, and videos, as a generator for multi-view data.
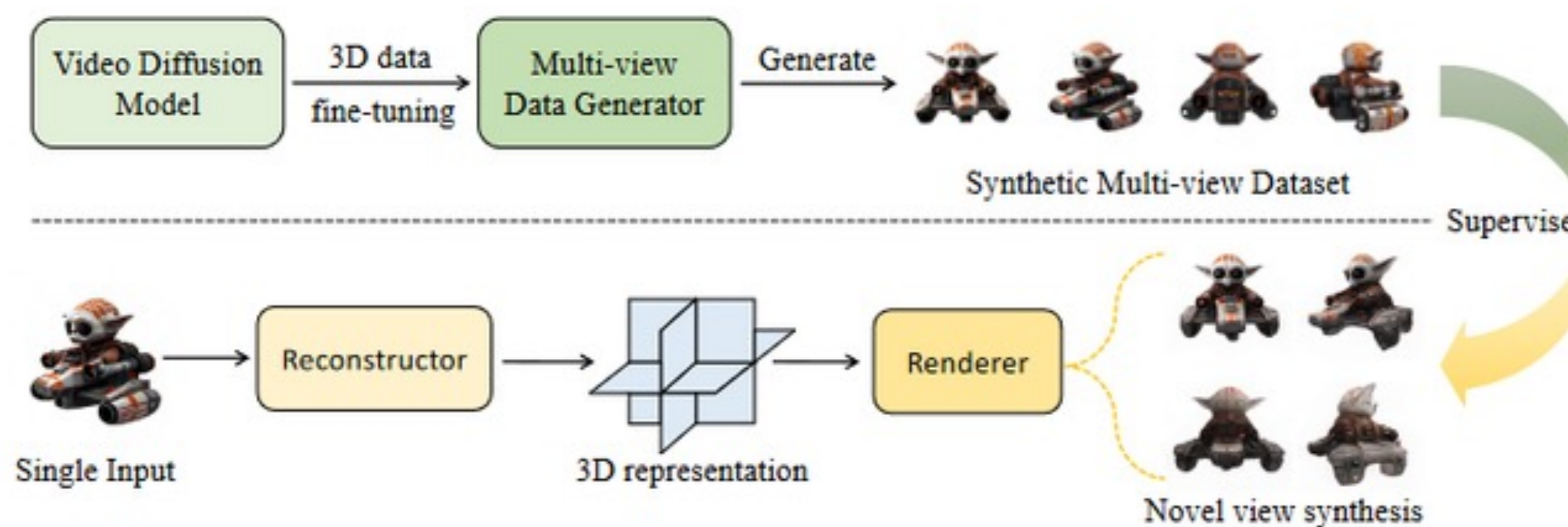
(1): Fine-tune a video diffusion model

We first fine-tune a video diffusion model using 100K rendered multi-view data, and then obtain metadata and classify high-quality data.

This fine-tuned video diffusion model can generate diverse and high-quality multi-view sequences for training 3D generative models.

(2): Train a large reconstruction model

We then train a large reconstruction model using several specially designed training strategies and the pre-generated ~3M synthetic data. We further fine-tune this model with a small amount (100K) rendered multi-view sequences to achieve optimal performance.

Text/single-image to 3D results

Input   OpenLRM   LGM   VFusion3D   Input   OpenLRM   LGM   VFusion3D

Scaling trends
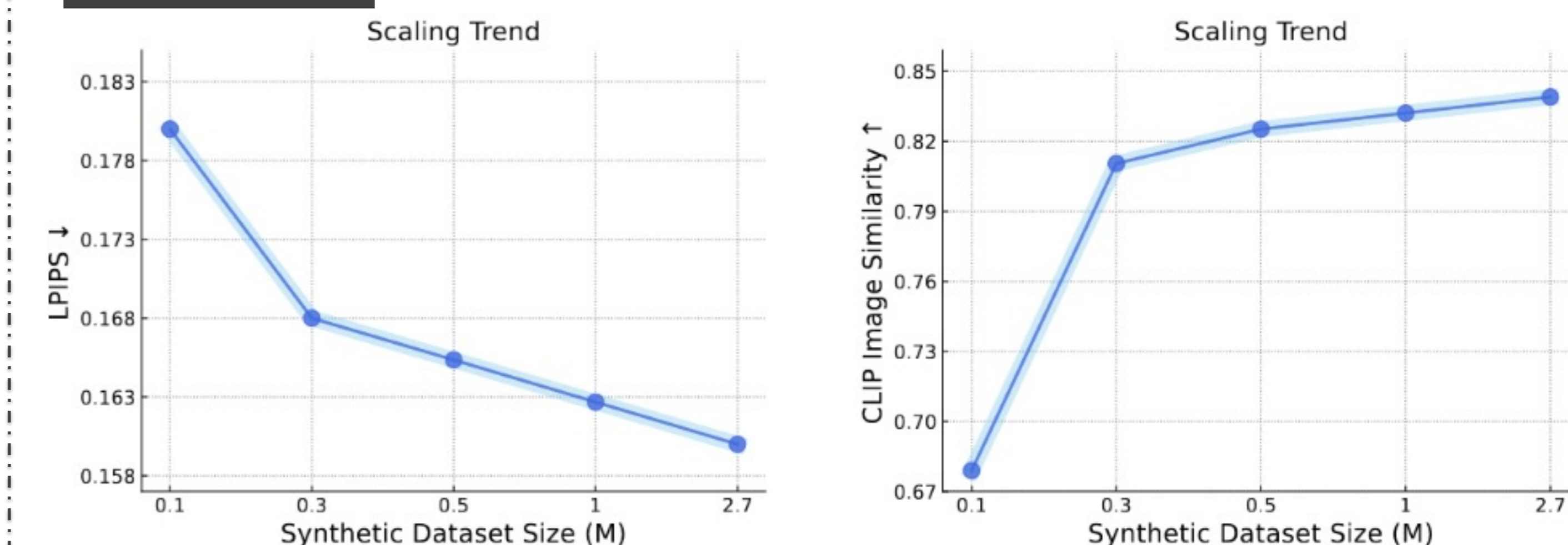
The generation quality consistently improves as the synthetic dataset size increases. VFusion3D also scales with several other factors, like stronger video diffusion models and the size of 3D data used for fine-tuning.

New follow-up work

FlexiView3D: Feed-Forward 3D Generation with Flexible Reconstruction Model and Input View Curation
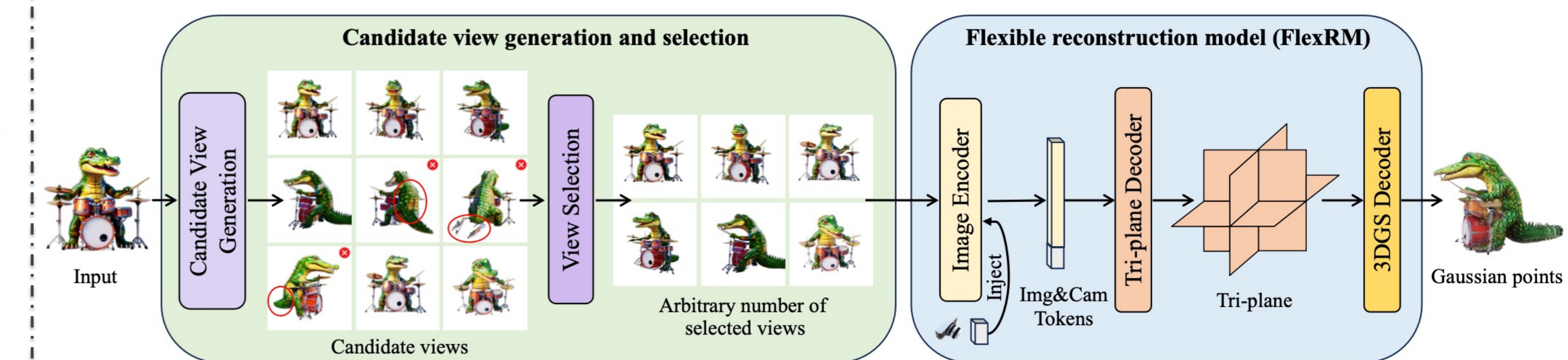
FlexiView3D first generates a pool of candidate views and then filters out those that are not good enough. The selected views are fed into a flexible reconstruction model that can take an arbitrary amount (1-32) of input views to directly output 1M Gaussian points in less than a second, with SOTA performance in various 3D generation and reconstruction tasks.