Evaluating Latent Generative Paradigms for High-Fidelity 3D Shape Completion from a Single Depth Image

Matthias Humt[1,2] Ulrich Hillenbrand[1] Rudolph Triebel[1,3]

[1] German Aerospace Center (DLR) [2] Technical University of Munich (TUM) [3] Karlsruhe Institute of Technology (KIT)

① The latent space is the bottleneck
AR matches or beats diffusion on the same discrete space — diffusion's apparent edge traces to VAE vs. VQ-VAE

② Generative ≫ Discriminative
Sampling diverse completions and selecting the best yields +21% F1 — even two samples suffice

③ Zero-shot to real sensors
On real Kinect scans, **generative** best-of-10 reaches F1 52.9 vs. 46.0 **discriminative** — no fine-tuning

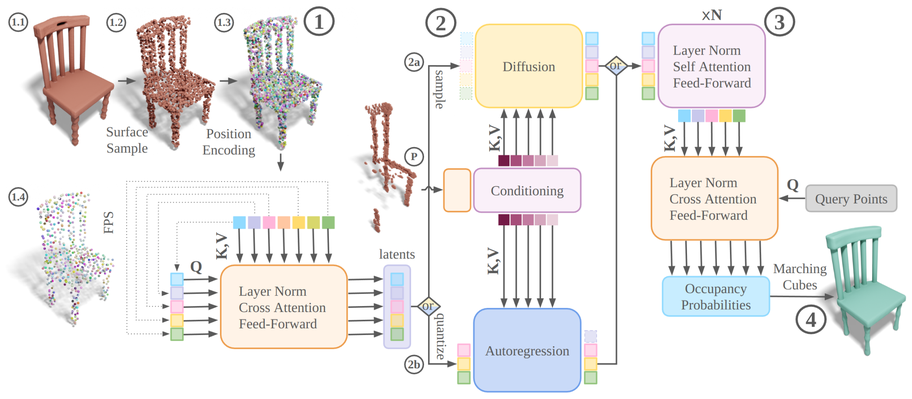
Motivation

Problem: A single **depth image** leaves most of an object's geometry unobserved. **Discriminative** models resolve this ambiguity by predicting the *mean* — producing blurry, unrealistic completions.

Insight: **Generative** models can instead sample diverse, high-fidelity completions. But which paradigm — diffusion or autoregressive — is better, and why?

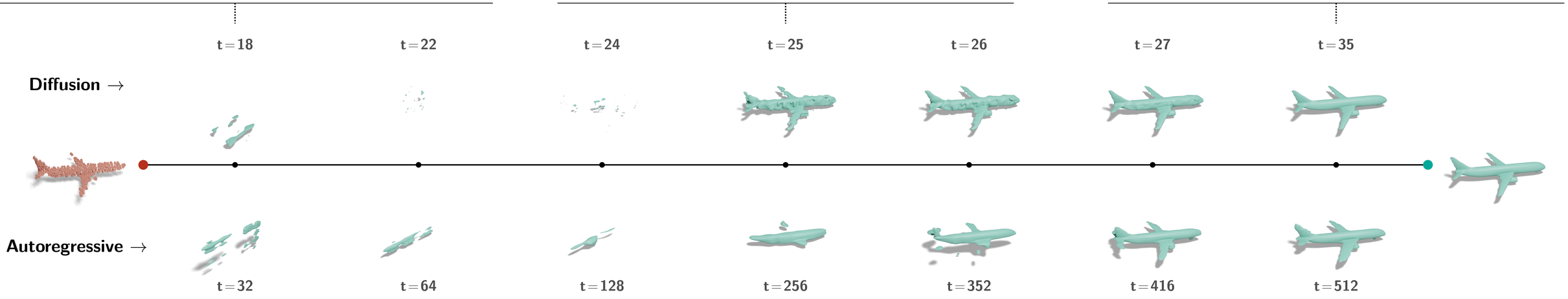
This work: Controlled comparison of both paradigms in a shared encoder-decoder pipeline, against a **discriminative** baseline, on noisy **depth input** including real sensor data.

Architecture

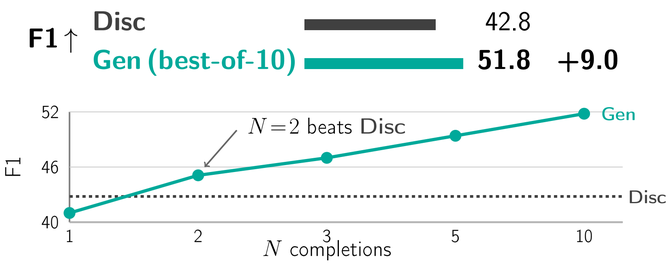


Reconstruction: VAE vs VQ-VAE

- Reconstruction quality sets an **upper bound** for any **generative** model built on top
- Diffusion uses VAE latents; AR uses VQ-VAE
- Diffusion's apparent edge is a **latent-quality effect**, not a paradigm difference



Generative ≫ Discriminative



- **Discriminative** models average over ambiguity
- **Generative** models sample diverse completions
- The best of even **two** already beats **discriminative**

AR ≥ Diffusion on Same Latent Space

Class-conditional generation on VQ-VAE latents

On the same discrete space, AR matches or exceeds diffusion.

VAE latents: FID 30.0, Prec. 53.9, Rec. 46.9 — the latent representation, not the paradigm, is the bottleneck.

Real-World Generalization

Automatica/YCB — real Kinect **depth**, zero-shot (trained on synthetic ShapeNet only)