Evaluating Latent Generative Paradigms For High-Fidelity 3D Shape Completion From A Single Depth Image

Evaluating Latent Generative Paradigms For High-Fidelity 3D Shape Completion From A Single Depth Image

Matthias Humt · Ulrich Hillenbrand · Rudolph Triebel

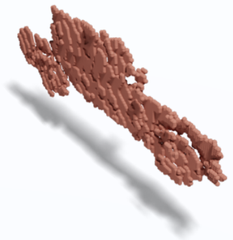
German Aerospace Center (DLR) · Technical University of Munich · Karlsruhe Institute of Technology

3DV 2026 — Vancouver

The Problem

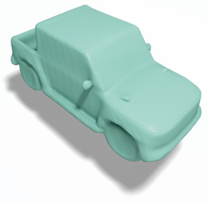
The Problem

Partial Depth Input

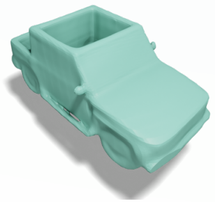


What does the
complete shape look like?

Multiple Plausible Completions



Sample 1



Sample 3

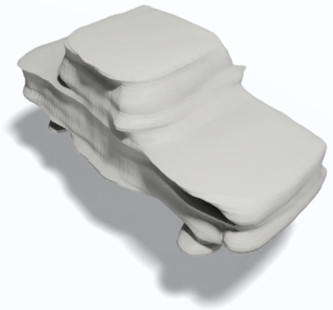
Ground Truth

Two Paradigms

Two Paradigms

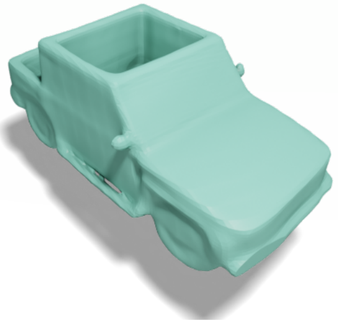
Discriminative

Predicts the *mean* over all plausible completions —
safe but blurry, misses fine detail



Generative

Samples *diverse* plausible completions —
sharp, high-fidelity, multimodal



Background

Background

3D Shape Completion

Well-studied with **discriminative** methods — but these predict a single mean shape, losing fine detail.

Generative Shape Completion

Diffusion and **AR** show promise for 3D generation, but no systematic study with a direct comparison to discriminative models exists for shape completion conditioned on partial input.

Contribution

First controlled comparison — same architecture (3DShape2VecSet), different latent spaces, evaluated on synthetic and real sensor data.

Research Questions

Research Questions

Which Generative Paradigm Is Better?

Diffusion (continuous VAE latents) vs. **Autoregressive** (discrete VQ-VAE latents)

Can Generative Beat Discriminative?

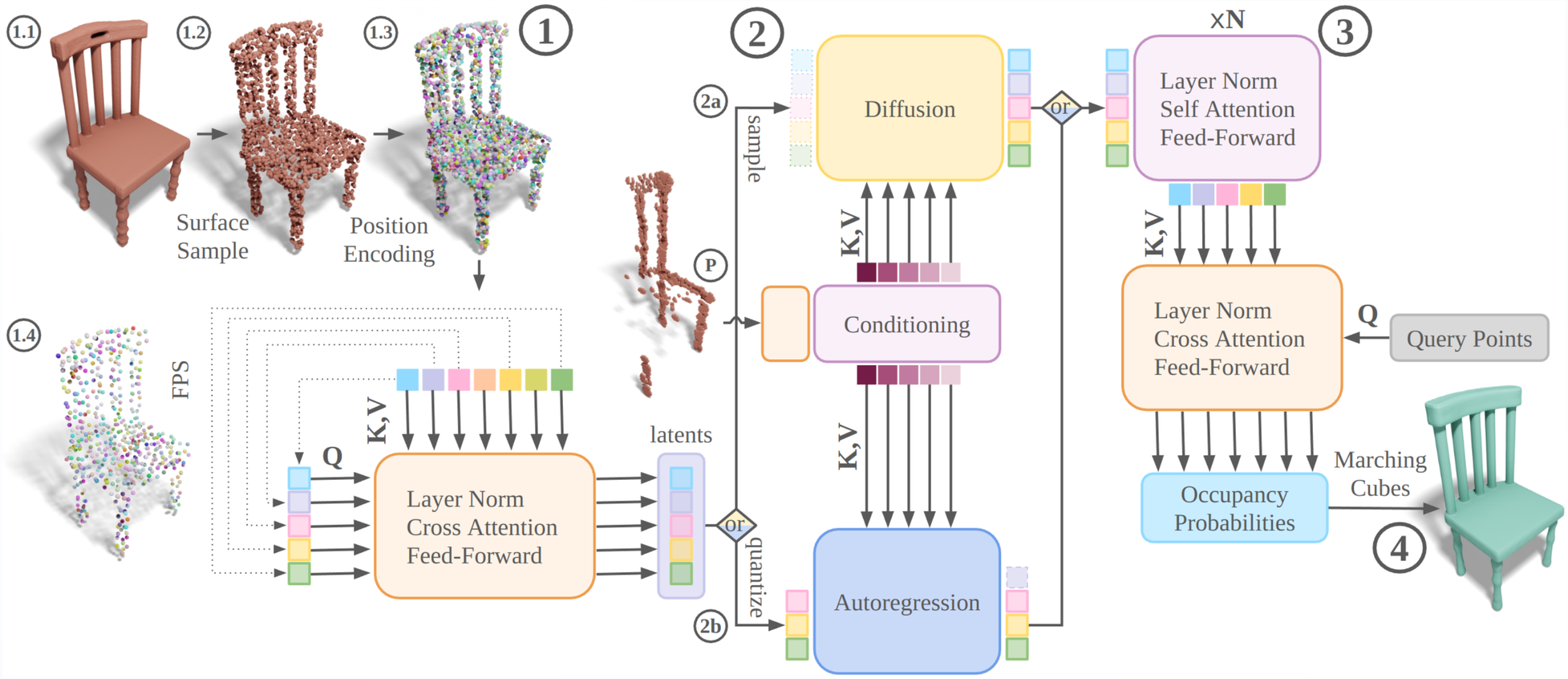
Best-of-N sampling leverages the model's true strength — multimodality

Do Results Transfer To Real Sensors?

Zero-shot evaluation on real Kinect depth scans — no fine-tuning

Architecture

Architecture



Shared encoder–decoder backbone (3DShape2VecSet). Only the **latent representation** differs: **VAE** (continuous) for Diffusion · **VQ-VAE** (discrete, codebook) for Autoregressive

Generation Process

Generation Process

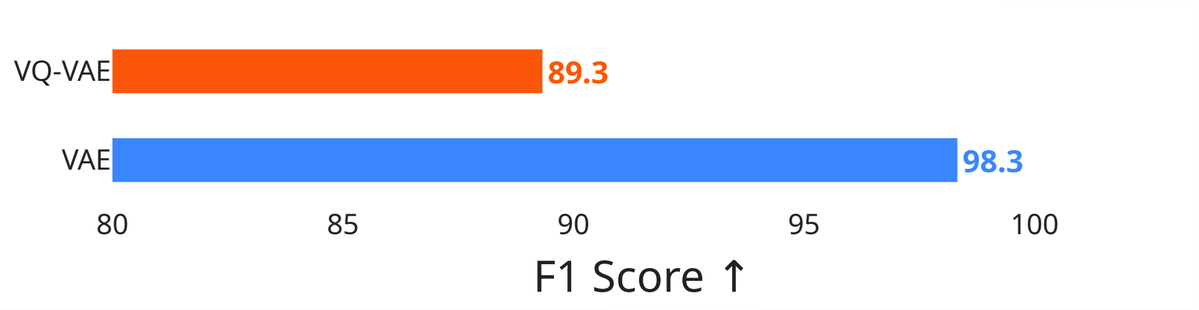
Diffusion (Denoising Steps)

Autoregressive (Token Accumulation)

1. The Latent Space Is The Bottleneck

1. The Latent Space Is The Bottleneck

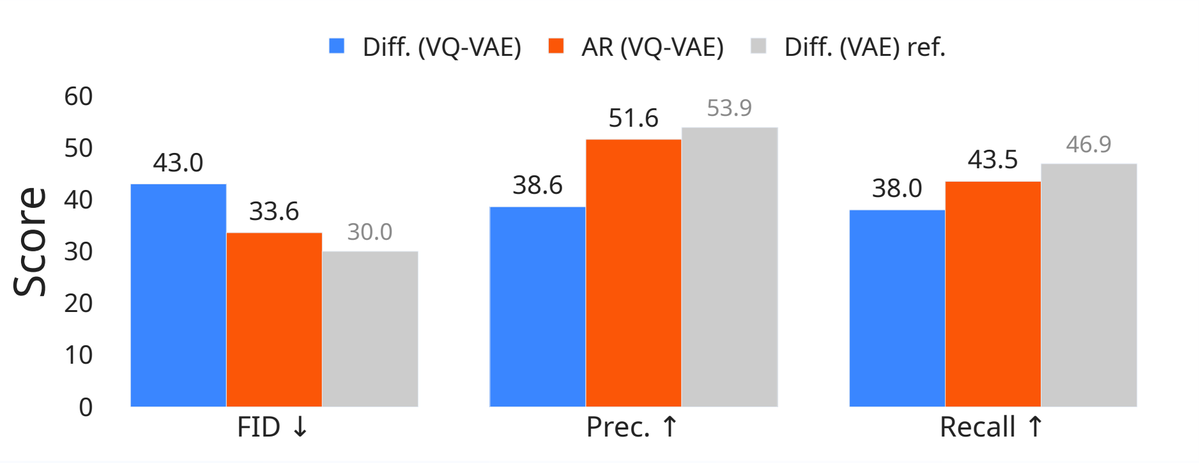
Reconstruction Quality (Upper Bound)



VAE **F1=98.3** vs. VQ-VAE **F1=89.3**

- **The naive comparison is unfair**
- Diffusion uses **VAE** (continuous, higher quality)
- AR uses **VQ-VAE** (discrete, codebook bottleneck)

Same Latent Space, Class-Conditional:

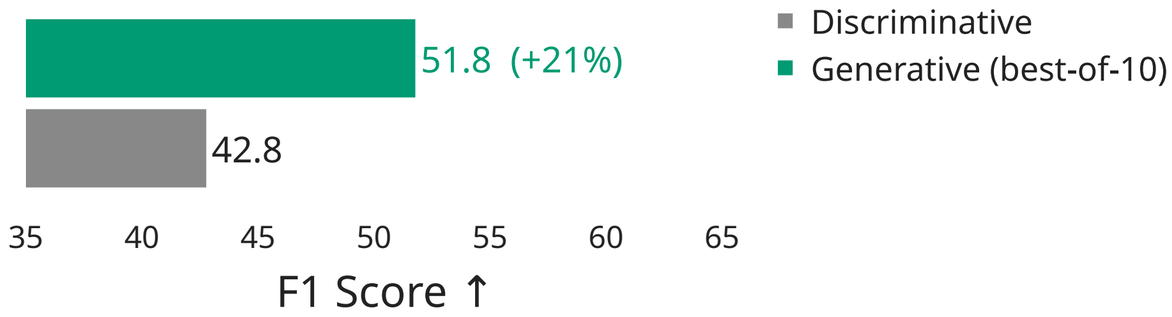


- On the *same* VQ-VAE space: **AR ≥ Diffusion**
- FID **33.6** vs. **43.0** · Prec. **51.6** vs. **38.6**
  - VAE ref FID=30.0 — gap is representation quality, not paradigm

2. Generative >> Discriminative

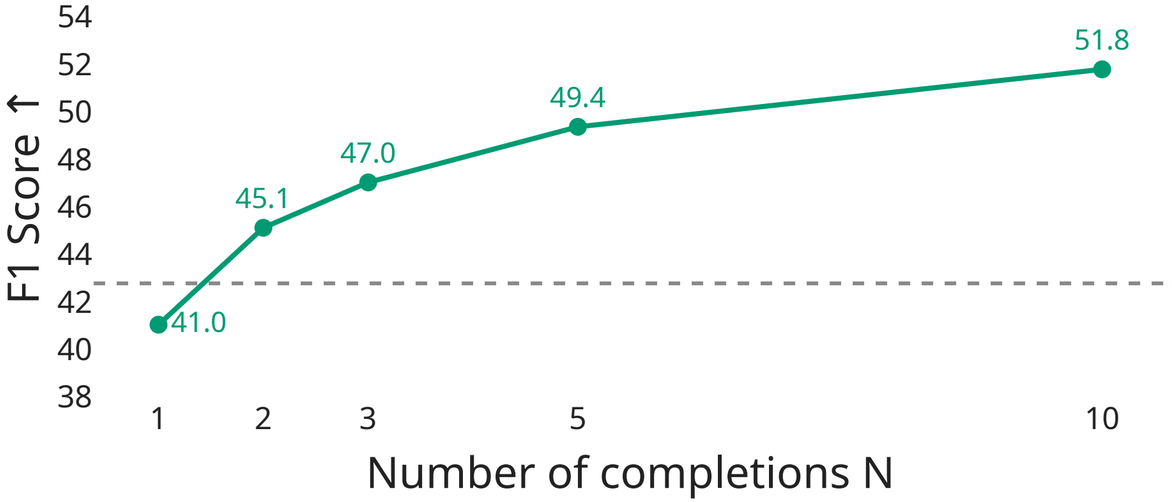
2. Generative >> Discriminative

Kinect Shape Completion — Mean F1



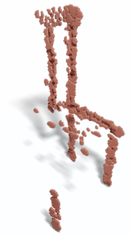
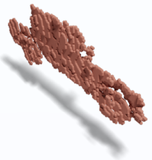
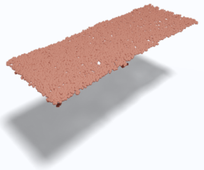
- Generate N completions, select the one closest to GT

Best-Of-N Sampling



- N=1 already competitive with discriminative
- **N=2 beats discriminative** · N=10: **+21% F1**

2. Qualitative Comparison

Input	Ground Truth	Gen (best)	Gen (sample)	Discriminative
				
				
				

Generative samples are sharper and capture fine detail · Discriminative output is blurry (conditional mean)

3. Zero-Shot To Real Sensors

3. Zero-Shot To Real Sensors

Domain Shift — F1 Score

Trained on ShapeNet · tested on YCB objects with simulated Kinect noise

- Gen **F1=52.9** vs. Disc F1=46.0 vs. Prior 43.4
- +15% improvement under domain shift

Real Kinect Scans (Zero-Shot)

Summary

Summary

1. Latent Space Is The Bottleneck

Diffusion appears superior only because **VAE** > **VQ-VAE** latent quality. On the *same* latent space: **AR ≥ Diffusion**.

2. Generative >> Discriminative

Best-of-N sampling consistently outperforms **discriminative** models. Even N=2 already beats the baseline.

3. Zero-Shot To Real Sensors

Trained on synthetic ShapeNet, tested on real Kinect — **generative** models generalize better than **discriminative**, no fine-tuning needed.

Limitations & Outlook

Oracle Selection

Best-of-N uses ground truth; practical deployment needs a learned ranking metric

VQ-VAE Gap

Discrete tokenization limits reconstruction quality; better codebooks could close this

Depth Only

Incorporating RGB could provide texture and semantic cues to reduce ambiguity

Questions?

Project Page