The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

BridgeShape: Latent Diffusion Schrödinger Bridge for 3D Shape Completion

Dequan Kong[1], **Honghua Chen**[2*], **Zhe Zhu**[1], **Mingqiang Wei**[1]
[1]Nanjing University of Aeronautics and Astronautics
[2]School of Data Science, Lingnan University, Hong Kong SAR
{dqkong, zhuzhe0619, mqwei}@nuaa.edu.cn, honghuachen@LN.edu.hk

Abstract

Existing diffusion-based 3D shape completion methods typically use a conditional paradigm, injecting incomplete shape information into the denoising network via deep feature interactions (e.g., concatenation, cross-attention) to guide sampling toward complete shapes, often represented by voxel-based distance functions. However, these approaches fail to explicitly model the optimal global transport path, leading to suboptimal completions. Moreover, performing diffusion directly in voxel space imposes resolution constraints, limiting the generation of fine-grained geometric details. To address these challenges, we propose BridgeShape, a novel framework for 3D shape completion via latent diffusion Schrödinger bridge. The key innovations lie in two aspects: (i) BridgeShape formulates shape completion as an optimal transport problem, explicitly modeling the transition between incomplete and complete shapes to ensure a globally coherent transformation. (ii) We introduce a Depth-Enhanced Vector Quantized Variational Autoencoder (VQ-VAE) to encode 3D shapes into a compact latent space, leveraging self-projected multi-view depth information enriched with strong DINOv2 features to enhance geometric structural perception. By operating in a compact yet structurally informative latent space, BridgeShape effectively mitigates resolution constraints and enables more efficient and high-fidelity 3D shape completion. BridgeShape achieves state-of-the-art performance on 3D shape completion benchmarks, demonstrating superior fidelity at higher resolutions and for unseen object classes.

Code — https://github.com/kizzyk/BridgeShape

Introduction

With the rapid advancement of 3D acquisition technologies, 3D sensors have become increasingly accessible and affordable, including various LiDAR scanners deployed across different platforms and RGB-D cameras such as Microsoft Kinect and Intel RealSense. This democratization has significantly expanded their applications across content creation, mixed reality, and machine vision domains (Zollhöfer et al. 2018; Chu et al. 2023). Despite this progress, these sensors are still inherently limited by occlusions, restricted angular fields of view, and surface reflectance issues, often resulting in incomplete and fragmented 3D scans. Such deficiencies hinder downstream tasks that demand complete, high-fidelity 3D representations, thereby necessitating effective shape completion techniques to reconstruct the missing geometric structures in a plausible and accurate manner.

Early approaches to 3D shape completion primarily employ convolutional neural networks (Dai, Ruizhongtai Qi, and Nießner 2017; Han et al. 2017) or transformer-based architectures (Rao, Nie, and Dai 2022) to directly infer complete shapes from partial observations. Building on this, generative models such as variational autoencoders (VAEs) (Stutz and Geiger 2020; Mittal et al. 2022; Yan et al. 2022) and adversarial networks (Zhang et al. 2021; Wu et al. 2020; Smith and Meger 2017), were explored to better capture the underlying shape distribution and improve the realism of generated results. More recently, diffusion probabilistic models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) have emerged as a powerful generative framework, achieving state-of-the-art results across various domains (Ramesh et al. 2022; Rombach et al. 2022; Nichol and Dhariwal 2021). Inspired by these successes, recent works (Müller et al. 2023; Chou, Bahat, and Heide 2023; Chu et al. 2023; Cheng et al. 2023) have adapted diffusion models for 3D shape completion, adopting a conditional generation paradigm where incomplete shapes are injected into denoising networks via deep feature interactions (e.g., concatenation and cross-attention) to guide sampling toward complete shapes. However, most of these methods operate diffusion directly in voxel space, inherently limiting the resolution of generated shapes and restricting the ability to capture fine-grained geometric details.

Beyond the resolution bottleneck, a more fundamental limitation of diffusion-based methods is that their reverse processes always begin from Gaussian noise, which carries minimal information about the target distribution. As a result, the substantial discrepancy between the Gaussian prior and the target shape distribution necessitates additional conditioning mechanisms and more sampling steps to converge to the target distribution. However, such implicit conditioning strategies do not explicitly model the global optimal transport path from the incomplete to the complete shape distribution, leading to suboptimal reconstructions and loss of geometric fidelity (see Fig. 1 (c) and (e)).

---

[*] Corresponding author

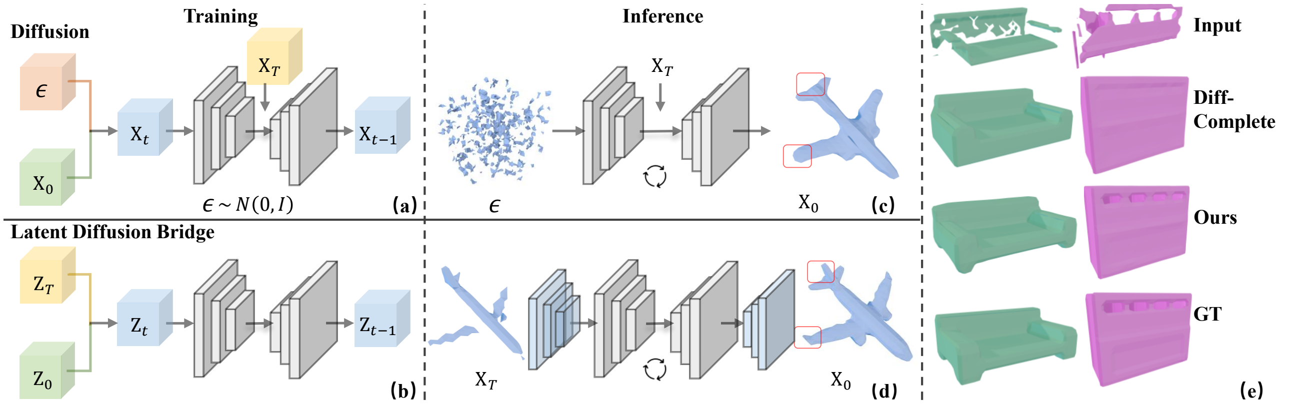

Figure 1: Comparison between existing diffusion-based shape completion paradigms and our proposed latent-diffusion-bridge-based approach. (a) Existing diffusion models incorporate an additional branch to inject deep features into the denoising process, transmitting incomplete shape information without explicitly modeling the transformation between the incomplete shape X_T and the complete shape X_0. (b) The proposed latent diffusion bridge explicitly models the optimal transport path between the latent distributions of incomplete and complete shapes (Z_T and Z_0, respectively). (c) Existing diffusion frameworks often produce less coherent completions with missing details, whereas (d) our latent diffusion bridge generates more structurally consistent and detailed 3D shapes. (e) Qualitative comparison of our BridgeShape with DiffComplete (Chu et al. 2023).

To address these limitations, we introduce a new paradigm for 3D shape completion. Given that the task inherently involves transforming an incomplete shape into a plausible complete one, it is intuitive to start the generative process from the given partial input—rather than from noise—which provides a far more informative and structured prior. This insight motivates the use of diffusion bridge models (Li et al. 2023; Liu et al. 2023), which condition the diffusion process on both the starting (incomplete) and target (complete) distributions, thereby learning an explicit optimal transport path. While recent works such as I^2SB (Liu et al. 2023) have demonstrated the effectiveness of this idea in image domains, their potential for 3D shape completion remains largely unexplored.

We propose BridgeShape, a novel framework for 3D shape completion via the latent diffusion Schrödinger bridge (DSB). First, to address the computational challenges inherent in 3D representations, we introduce a Depth-Enhanced VQ-VAE (van den Oord, Vinyals, and Kavukcuoglu 2017) that compresses 3D shapes into a compact yet structurally informative latent space. This representation integrates self-projected multi-view depth information enriched with strong DINOv2 (Oquab et al. 2023) features, enhancing its geometric structural perception. Then, within the latent space, we formulate the DSB to explicitly model the optimal transport between the distributions of partial and complete shapes, enabling a globally consistent and detail-preserving transformation. By leveraging the structurally informative latent encoding and the optimality of the Schrödinger bridge formulation, BridgeShape achieves fine-grained, high-fidelity completions even under challenging conditions. Extensive experiments on large-scale 3D shape completion benchmarks demonstrate state-of-the-art performance, even when generalizing to unseen object categories and higher resolutions. Our main contributions are summarized as follows:

- We propose BridgeShape, casting 3D shape completion as an optimal-transport problem in latent space.
- We introduce a Depth-Enhanced VQ-VAE that encodes 3D shapes into a latent space, leveraging self-projected multi-view depth to enhance geometric perception.
- On large-scale 3D completion benchmarks, BridgeShape achieves impressive performance even at higher resolutions and on unseen object classes.

Related Work

3D Shape Completion

3D shape completion recovers missing regions in scans, crucial for scene understanding. Early learning-based methods (Dai, Ruizhongtai Qi, and Nießner 2017; Han et al. 2017) rely on convolutional neural networks, while more recent approaches leverage transformer-based architectures. For instance, 3D-EPN (Dai, Ruizhongtai Qi, and Nießner 2017) employs a 3D encoder-decoder framework to infer complete shapes, and PatchComplete (Rao, Nie, and Dai 2022) leverages multi-resolution patch priors for generalization. A series of point-cloud-based approaches (Wen et al. 2023; Li, Zhu, and Wei 2025; Yu et al. 2023; Zhu et al. 2023, 2024, 2025) have also addressed this task. Generative models—GANs (Zhang et al. 2021; Wu et al. 2020; Smith and Meger 2017) and autoencoders (Mittal et al. 2022; Yan et al. 2022)—model completion uncertainty. cGAN (Wu et al. 2020) distills the ambiguity by conditioning the completion on a learned multimodal distribution, while ShapeFormer (Yan et al. 2022) generates complete sequences. In contrast, we propose a novel framework that explicitly transports incomplete to complete shapes in a compact, depth-enhanced latent space, ensuring coherent and precise results.

Diffusion models for 3D generation

Diffusion models (Song et al. 2020; Rombach et al. 2022; Dhariwal and Nichol 2021; Sohl-Dickstein et al. 2015; Shim, Kang, and Joo 2023; Wang et al. 2024) have proven powerful for 3D shape generation and have recently been adapted to point cloud synthesis (Zhou, Du, and Wu 2021; Luo and Hu 2021; Vahdat et al. 2022). For conditional shape completion, SDFusion (Cheng et al. 2023) and Diffusion-SDF (Chou, Bahat, and Heide 2023) use diffusion to fill missing regions on synthetically cropped shapes. In contrast, our approach makes no strict assumptions on partial inputs, handling diverse noise and incompleteness. Another approach, DiffComplete (Chu et al. 2023), adopts a multi-level aggregation strategy to improve completion quality. However, performing diffusion directly in voxel space requires substantial memory at higher resolutions. While scaling techniques such as gradient accumulation can mitigate this, they incur extra training complexity. Moreover, a common limitation of these methods is their inability to explicitly model the optimal transport path between incomplete and complete shapes, often leading to suboptimal completions. BridgeShape overcomes this by directly modeling the optimal transport process within a compact latent space, thereby enabling higher-fidelity completions while alleviating resolution constraints.

Diffusion Bridge Models

Stochastic bridge models, which capture the evolution of stochastic processes constrained by fixed endpoints, have become an essential tool in probability theory (Aguilar et al. 2022; Chen, Georgiou, and Pavon 2021; Chen and Georgiou 2015). Integrated with diffusion models, they offer a novel approach for translating between distributions and modeling conditional probabilities without relying on prior information. This data-to-data generation paradigm has attracted significant interest in various generative tasks, including image-to-image translation (De Bortoli et al. 2021; Chen, Liu, and Theodorou 2021; Li et al. 2023; Liu et al. 2023), protein matching (Somnath et al. 2023), point cloud denoising (Vogel et al. 2024), and text-to-speech synthesis (Chen et al. 2023). However, the application of diffusion bridge models to 3D shape completion remains unexplored.

Method

Overview

The architecture of our approach is shown in Fig. 2. Given a partial scan and its corresponding ground truth complete shape, following previous practice (Chu et al. 2023), we represent the partial scan as a truncated signed distance field (TSDF) and the complete shape as a truncated unsigned distance field (TUDF) in a volumetric grid. To accelerate the diffusion-based completion framework, we first compress the complete shape into a low-resolution latent space using a VQ-VAE (Razavi, van den Oord, and Vinyals 2019; van den Oord, Vinyals, and Kavukcuoglu 2017; Cheng et al. 2023) enhanced by multi-view depth information. Built upon the pretrained latent space, the DSB (Liu et al. 2023) is utilized to model the diffusion process from the partial to the complete shape.

3D Shape Compression

VQ-VAE Compression. The VQ-VAE has an encoder $\mathcal{E}_c$ mapping a 3D shape into a latent space and a decoder $\mathcal{D}$ reconstructing it from the latent code. This architecture enables the application of diffusion models by providing a compact, lower-dimensional representation of the shape.

Let $\mathbf{X} \in \mathbb{R}^{D\times D\times D}$ represent a complete 3D shape in the form of a TUDF. The encoder $\mathcal{E}_c$ maps this shape to a latent vector $\mathbf{z} \in \mathbb{R}^{d\times d\times d}$, where $d < D$:

$$\mathbf{z} = \mathcal{E}_c(\mathbf{X}), \tag{1}$$

the latent vector $\mathbf{z}$ is then quantized by selecting the closest codebook element Z:

$$\mathcal{VQ}(\mathbf{z}) = \arg\min_{\mathbf{z}_i \in Z} \|\mathbf{z} - \mathbf{z}_i\|_2, \tag{2}$$

where $\mathbf{z}_i \in Z$ represents an element from the codebook, and $\|\cdot\|_2$ denotes the L2 norm. The decoder $\mathcal{D}$ reconstructs the shape from $\mathbf{z}$ defined as:

$$\mathbf{X}' = \mathcal{D}(\mathcal{VQ}(\mathbf{z})). \tag{3}$$

The encoder, decoder, and codebook are jointly optimized with the following total loss function:

$$\mathcal{L}_{\text{total}} = -\log p(\mathbf{X} \mid \mathbf{z}) + \|\hat{\mathbf{z}} - \text{sg}[\mathbf{z}]\|^2 + \|\,\text{sg}[\hat{\mathbf{z}}] - \mathbf{z}\|^2, \tag{4}$$

where sg[·] is the stop-gradient operation. The three terms in Equation 4 are: the reconstruction loss, the commitment loss, and the VQ objective.

Enhancement with Multi-View Depth Features. To strengthen the latent space representation, we incorporate multi-view depth features into the VQ-VAE. Specifically, we project the 3D shape into depth maps from N viewpoints, where each depth map is denoted as $\mathbf{D}_i$ for the i-th view, and its corresponding feature $\mathbf{F}_i$ is extracted using a pre-trained DINOv2 model (Oquab et al. 2023). To obtain a unified representation, we aggregate these features by averaging along the view dimension as a simple yet effective fusion strategy:

$$\mathbf{F}_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{F}_i. \tag{5}$$

This operation produces an aggregated feature map $\mathbf{F}_{\text{avg}}$ that fuses complementary information across all viewpoints while minimizing redundant details, resulting in a robust representation with minimal computational overhead. $\mathbf{F}_{\text{avg}}$ is then fused with the 3D shape's latent feature via a cross-attention mechanism at the end of the encoder $\mathcal{E}_c$. In this fusion process, the query matrix is derived from the 3D shape's latent feature, while the key and value matrix are derived from $\mathbf{F}_{\text{avg}}$. This fusion is performed at a critical stage—the end of the encoder—where the shape information has already been compressed into a high-level representation. This approach strikes a balance between model performance and resource utilization, avoiding unnecessary overhead while enriching the latent space representation.

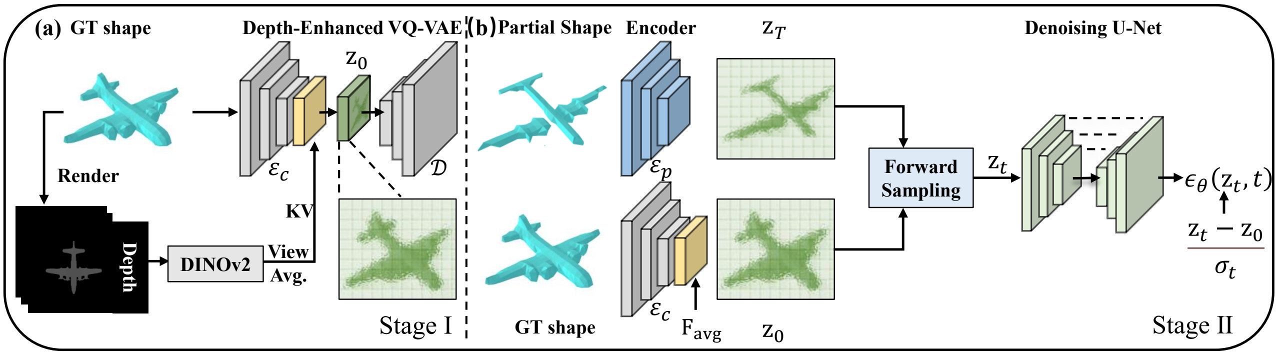

Figure 2: Overview of our training pipeline, which operates within the latent space based on the DSB. **Stage I**: Pre-training a Depth-Enhanced VQ-VAE on complete shapes to establish the latent space. **Stage II**: A co-trained encoder maps partial TSDF inputs into this latent space, where the diffusion bridge is applied to learn a structured diffusion trajectory between incomplete and complete shapes. This approach significantly enhances both efficiency and fidelity in shape completion.

Latent DSB

After compressing complete 3D shapes into a high-dimensional latent space, we construct a DSB to model the optimal transport between incomplete and complete shapes. As illustrated in Figure 2(b), we freeze the pre-trained VQ-VAE parameters and employ an additional trainable encoder, $\mathcal{E}_p$, to map partial inputs into the same latent space, where the latent representations of a given incomplete shape and its corresponding complete shape are denoted by $\mathbf{z}_T$ and $\mathbf{z}_0$, respectively. Then, the diffusion bridge enables efficient processing within this compact yet structurally informative space. In the following, we detail how 3D shape completion is formulated as an optimal transport problem via the diffusion bridge and provide an overview of the associated training and inference procedures.

Optimal Transport via DSB. DSB is a specialized diffusion process that progressively transforms $\mathbf{z}_0 \sim p_\mathrm{A}$ into $\mathbf{z}_T \sim p_\mathrm{B}$ via a sequence of intermediate states $\{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$ over T timesteps.

Given a reference path measure $\pi(\mathbf{z}_{0:T})$ that characterizes the ideal diffusion trajectory, our goal is to learn a process $p^*(\mathbf{z}_{0:T})$ that satisfies $p^*(\mathbf{z}_0) = p_\mathrm{A}$ and $p^*(\mathbf{z}_T) = p_\mathrm{B}$, while minimizing the Kullback-Leibler divergence between $\pi(\mathbf{z}_{0:T})$ and $p^*(\mathbf{z}_{0:T})$. This formulation is equivalent to the Schrödinger Bridge problem (Léonard 2013; Chen et al. 2023; Vogel et al. 2024), whose dynamics are characterized by the following stochastic differential equations (SDEs):

$$\mathrm{d}\mathbf{z}_t = [\mathbf{f}(\mathbf{z}_t, t) + g^2(t) \nabla \log \Psi_t(\mathbf{z}_t)]\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w}_t, \quad (6)$$

$$\mathrm{d}\mathbf{z}_t = [\mathbf{f}(\mathbf{z}_t, t) - g^2(t) \nabla \log \hat{\Psi}_t(\mathbf{z}_t)]\,\mathrm{d}t + g(t)\,\mathrm{d}\bar{\mathbf{w}}_t, \quad (7)$$

where $\mathbf{f}(\mathbf{z}_t, t)$ is the drift term, $g(t)$ is the diffusion coefficient and $\mathbf{w}_t$ is a Wiener process. The terms $\nabla \log \Psi_t(\mathbf{z}_t)$ and $\nabla \log \hat{\Psi}_t(\mathbf{z}_t)$ represent the extra nonlinear drift terms that solve the coupled partial differential equations (PDEs):

$$\begin{cases} \frac{\partial \Psi}{\partial t} = -\nabla \Psi^{\mathrm{T}} \mathbf{f} - \frac{1}{2} \mathrm{Tr}(g^2 \nabla^2 \Psi) \\ \frac{\partial \hat{\Psi}}{\partial t} = -\nabla (\hat{\Psi} \mathbf{f}) + \frac{1}{2} \mathrm{Tr}(g^2 \nabla^2 \hat{\Psi}) \end{cases} \quad (8)$$

such that

$$\Psi_0 \hat{\Psi}_0 = p_\mathrm{A}, \quad \Psi_T \hat{\Psi}_T = p_\mathrm{B}. \quad (9)$$

However, directly solving the Schrödinger bridge formulation is computationally prohibitive. To mitigate it, we leverage a framework from recent works (Liu et al. 2023; Chen et al. 2023; Vogel et al. 2024) under the assumption that paired training data is available, i.e.,

$$p(\mathbf{z}_0, \mathbf{z}_T) = p_A(\mathbf{z}_0)\, p_B(\mathbf{z}_T \mid \mathbf{z}_0). \quad (10)$$

In the context of 3D shape completion, the latent distribution over incomplete shapes is modeled by a joint distribution: $p_A(\mathbf{z}_0)$ represents the latent distribution of complete shapes, and $p_B(\mathbf{z}_T \mid \mathbf{z}_0)$ characterizes the latent distribution of missing components conditioned on the complete shapes.

Given a boundary pair $\mathbf{z}_0$ (complete shape) and $\mathbf{z}_T$ (incomplete shape), and assuming $\mathbf{f} := 0$, the posterior distribution at each timestep t is defined as:

$$q(\mathbf{z}_t \mid \mathbf{z}_0, \mathbf{z}_T) = \mathcal{N}\left(\mathbf{z}_t; \mu_t(\mathbf{z}_0, \mathbf{z}_T), \Sigma_t\right), \quad (11)$$

where the mean μ_t and covariance Σ_t are computed as:

$$\mu_t(\mathbf{z}_0, \mathbf{z}_T) = \frac{\sigma_{b,t}^2}{\sigma_{b,t}^2 + \sigma_t^2} \mathbf{z}_0 + \frac{\sigma_t^2}{\sigma_{b,t}^2 + \sigma_t^2} \mathbf{z}_T, \quad (12)$$

$$\Sigma_t = \frac{\sigma_t^2\, \sigma_{b,t}^2}{\sigma_t^2 + \sigma_{b,t}^2}. \quad (13)$$

Here, σ_t^2 and $\sigma_{b,t}^2$ represent the accumulated variance from $\mathbf{z}_T$ and $\mathbf{z}_0$, respectively. This framework enables efficient computation of the coupled PDEs (see Equation 8), thereby facilitating the latent diffusion process. Moreover, the sampling mechanism in Equation 11 is both tractable and sufficiently expressive to cover the generative trajectory, resulting in an effective and efficient shape completion pipeline.

Considering that extremely sparse incomplete shapes can introduce significant uncertainty in the missing regions, establishing a robust optimal transport path may become particularly challenging. To address this, we inject stochasticity into the latent distribution of incomplete shapes before constructing the optimal transport process—following the strategy employed in (Liu et al. 2023).

l_1-err. ↓	Chair	Table	Sofa	Lamp	Plane	Car	Dresser	Watercraft	Avg.
3D-EPN (Dai, Ruizhongtai Qi, and Nießner 2017)	0.418	0.377	0.392	0.388	0.421	0.259	0.381	0.356	0.374
ConvONet (Peng et al. 2020)	0.210	0.247	0.254	0.234	0.185	0.195	0.250	0.184	0.220
SDF-StyleGAN (Zheng et al. 2022)	0.321	0.256	0.289	0.280	0.295	0.224	0.273	0.282	0.278
cGCA (Zhang et al. 2022)	0.174	0.212	0.179	0.239	0.170	0.161	0.204	0.143	0.185
AutoSDF (Mittal et al. 2022)	0.201	0.258	0.226	0.275	0.184	0.187	0.248	0.157	0.217
ShapeFormer (Yan et al. 2022)	0.104	0.175	0.133	0.176	0.136	0.127	0.157	0.119	0.141
PatchComplete (Rao, Nie, and Dai 2022)	0.134	0.095	0.084	0.087	0.061	0.053	0.134	0.058	0.088
Diffcomplete (Chu et al. 2023)	0.070	0.073	0.061	0.059	0.015	0.025	0.086	0.031	0.053
BrigdeShape (Ours)	**0.055**	**0.059**	**0.047**	**0.038**	**0.012**	**0.023**	**0.055**	**0.022**	**0.039**

Table 1: Quantitative comparison for shape completion on 3D-EPN (Dai, Ruizhongtai Qi, and Nießner 2017).

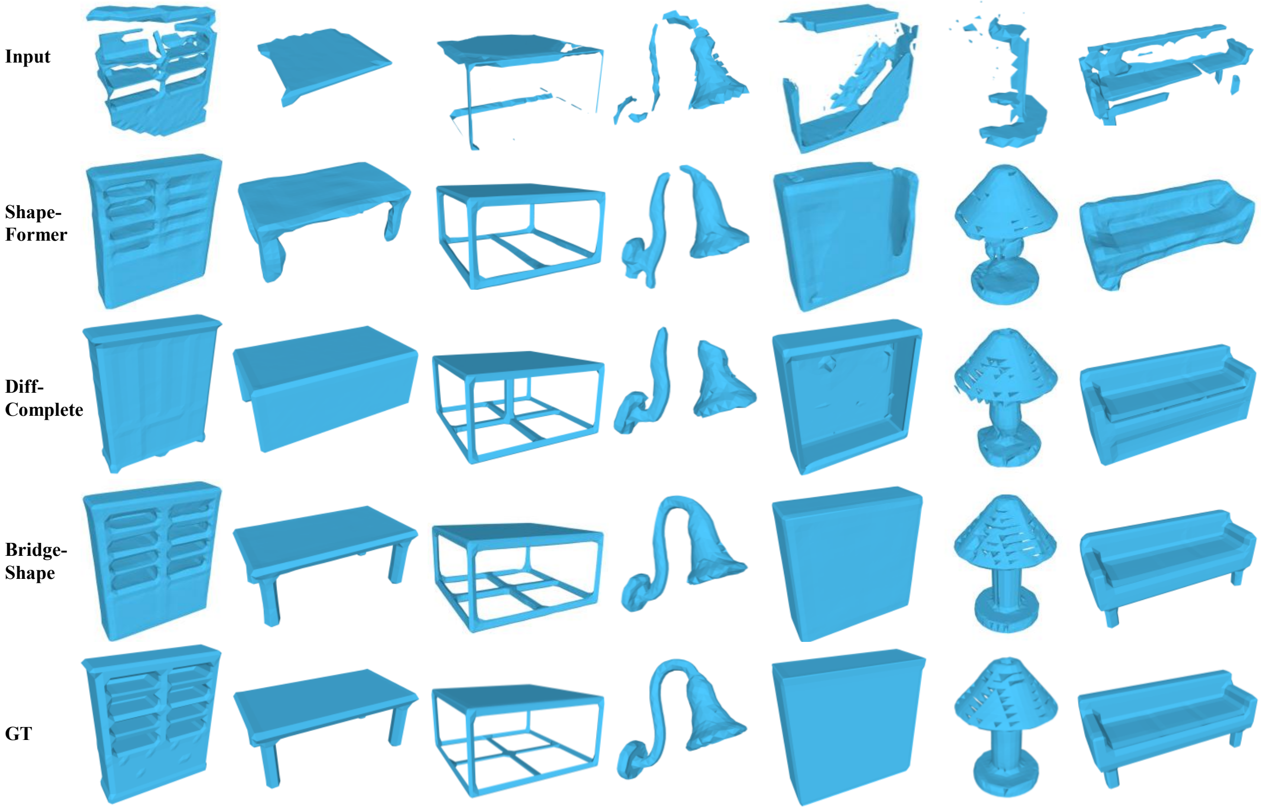

Figure 3: Qualitative comparison of shape completion on 3D-EPN (Dai, Ruizhongtai Qi, and Nießner 2017).

Noise Prediction and Inference. Our training objective is to accurately predict the noise injected at each timestep during the forward diffusion process. This is accomplished using a neural network ϵ_θ that estimates the noise $\epsilon_\theta(\mathbf{z}_t, t)$ at every timestep t. The training loss is defined as:

$$\mathcal{L} = \|\epsilon_\theta(\mathbf{z}_t, t) - \frac{\mathbf{z}_t - \mathbf{z}_0}{\sigma_t}\|_2^2. \tag{14}$$

During inference, we can run standard DDPM (Ho, Jain, and Abbeel 2020) to iteratively refine the complete shape's latent representation, starting from $\mathbf{z}_T$(the latent code of the incomplete shape) and using reverse dynamics governed by the mean $\mu_t(\hat{\mathbf{z}}_0, \mathbf{z}_T)$ and covariance Σ_t as:

$$p(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \hat{\mathbf{z}}_\mathbf{0}) = \mathcal{N}(\mathbf{z}_t; \mu_t(\hat{\mathbf{z}}_\mathbf{0}, \mathbf{z}_T), \Sigma_t). \tag{15}$$

This induces the same marginal density of Schrödinger bridge paths when $\hat{\mathbf{z}}_0$ is closed to $\mathbf{z}_0$ (Liu et al. 2023).

Experiments

Experimental Settings

Datasets. We evaluate on two large-scale shape completion benchmarks. 3D-EPN (Dai, Ruizhongtai Qi, and Nießner 2017) comprises 25,590 training and 5,384 testing instances across eight ShapeNet categories, each with six partial scans (32^3 TSDF) and corresponding complete shapes (32^3, 64^3, or 128^3 TUDF). PatchComplete (Rao, Nie, and Dai 2022) includes synthetic ShapeNet data and real-world scan data from ScanNet (Dai et al. 2017). For

	Bag		Lamp		Bathtub		Bed		Basket		Printer		Laptop		Bench		Avg.	
	CD ↓	IoU ↑	CD	IoU	CD	IoU	CD	IoU	CD	IoU	CD	IoU	CD	IoU	CD	IoU	CD	IoU
3D-EPN	5.01	73.8	8.07	47.2	4.21	57.9	5.84	58.4	7.90	54.0	5.15	73.6	3.90	62.0	4.54	48.3	5.58	59.4
Few-Shot	8.00	56.1	15.1	25.4	7.05	45.7	10.0	39.6	8.72	40.6	9.26	56.7	10.4	31.3	8.11	27.2	9.58	40.3
IF-Nets	4.77	69.8	5.70	50.8	4.72	55.0	5.34	60.7	4.44	50.2	5.83	70.5	6.47	58.3	5.03	49.7	5.29	58.1
Auto-SDF	5.81	56.3	6.57	39.1	5.17	41.0	6.01	44.6	6.70	39.8	7.52	49.9	4.81	51.1	4.31	39.5	5.86	45.2
ConvONet	5.10	70.8	5.42	52.6	4.96	60.4	5.42	63.2	6.16	54.6	5.56	72.1	4.78	57.3	4.69	49.6	5.26	60.1
PatchComplete	3.94	77.6	**4.68**	56.4	3.78	66.3	4.49	66.8	5.15	61.0	4.63	77.6	3.77	63.8	3.70	53.9	4.27	65.4
DiffComplete	3.86	78.3	4.80	57.9	3.52	68.9	**4.16**	67.1	**4.94**	**65.5**	4.40	76.8	3.52	67.4	3.56	58.2	4.10	67.5
BridgeShape (Ours)	**3.70**	**80.0**	4.94	**62.7**	**3.41**	**69.8**	4.32	**69.6**	5.50	65.1	**4.09**	**81.4**	**3.13**	**72.7**	**3.38**	**59.1**	**4.06**	**70.1**

Table 2: Quantitative comparison with state-of-the-art methods (Dai, Ruizhongtai Qi, and Nießner 2017; Wallace and Hariharan 2019; Chibane, Alldieck, and Pons-Moll 2020; Mittal et al. 2022; Peng et al. 2020; Rao, Nie, and Dai 2022; Chu et al. 2023) on synthetic objects (Chang et al. 2015) of unseen categories. (CD $\times 10^2$ and IoU $\times 10^2$)

	Bag		Lamp		Bathtub		Bed		Basket		Printer		Avg.	
	CD ↓	IoU ↑	CD	IoU	CD	IoU	CD	IoU	CD	IoU	CD	IoU	CD	IoU
3D-EPN	8.83	53.7	14.3	20.7	7.56	41.0	7.76	47.8	7.74	36.5	8.36	63.0	9.09	44.0
Few-Shot	9.10	44.9	11.9	19.6	7.77	38.2	9.07	34.9	8.02	34.3	8.30	62.2	9.02	38.6
IF-Nets	8.96	44.2	10.2	24.9	7.19	39.5	8.24	44.9	6.74	42.7	8.28	60.7	8.26	42.6
Auto-SDF	9.30	48.7	11.2	24.4	7.84	36.6	7.91	38.0	7.54	36.1	9.66	49.9	8.90	38.9
ConvONet	9.12	52.5	9.83	20.3	7.93	41.2	8.14	41.6	7.39	37.0	7.62	64.9	8.34	42.9
PatchComplete	8.23	58.3	9.42	28.4	6.77	48.0	7.24	48.4	6.60	45.5	6.84	70.5	7.52	49.8
DiffComplete	**7.05**	48.5	**6.84**	30.5	8.22	48.5	7.20	46.6	7.42	**59.2**	**6.36**	**74.5**	7.18	51.3
BridgeShape (Ours)	7.67	**61.2**	8.07	**36.3**	**6.28**	**50.9**	**6.87**	**51.3**	**6.20**	50.2	6.83	71.0	**6.99**	**53.5**

Table 3: Quantitative comparison with state-of-the-art methods (Dai, Ruizhongtai Qi, and Nießner 2017; Wallace and Hariharan 2019; Chibane, Alldieck, and Pons-Moll 2020; Mittal et al. 2022; Peng et al. 2020; Rao, Nie, and Dai 2022; Chu et al. 2023) on real-world objects (Dai et al. 2017) of unseen categories. (CD $\times 10^2$ and IoU $\times 10^2$)

ShapeNet, 18 categories are used for training and 8 for testing, with 3,202 training models and 1,325 test models, each having four partial scans. The ScanNet data consists of objects extracted from bounding boxes, with complete shapes provided by Scan2CAD (Avetisyan et al. 2019). All objects are represented as 32^3 TSDF units. These datasets allow us to evaluate our method on both synthetic and real-world data, testing its ability to handle unseen categories.

Evaluation. We evaluate our method using standard metrics for shape completion. On the 3D-EPN dataset (Dai, Ruizhongtai Qi, and Nießner 2017), we report the mean l_1 error across all voxels of the TUDF predictions. For PatchComplete benchmark (Rao, Nie, and Dai 2022), we use the l_1 Chamfer Distance (CD) and Intersection over Union (IoU) to evaluate the geometry of predicted shapes. 10K points are sampled on surfaces for CD calculation.

Evaluation on known object categories

The quantitative and qualitative comparison with state-of-the-art methods (Dai, Ruizhongtai Qi, and Nießner 2017; Zhang et al. 2022; Mittal et al. 2022; Zheng et al. 2022; Peng et al. 2020; Rao, Nie, and Dai 2022; Chu et al. 2023; Yan et al. 2022) on the 3D-EPN dataset (Dai, Ruizhongtai Qi, and Nießner 2017) is presented in Table 1 and Figure 3. For the probabilistic methods, we report the average results across five inferences with different initializations. In comparison to the second-ranked method DiffComplete (Chu et al. 2023), BridgeShape reduces the l1 error by approximately 26% (from 0.053 to 0.039), while producing high-fidelity, realistic shapes with fewer surface artifacts. This is because DiffComplete reverses from unstructured Gaussian noise and relies on deep feature interactions to inject incomplete shape information, while our approach directly starts the generation process from the input incomplete shapes, providing a far more informative prior. Additionally, compared to GAN-based SDF-StyleGAN (Zheng et al. 2022), autoregressive methods AutoSDF (Mittal et al. 2022) and ShapeFormer (Yan et al. 2022), our diffusion bridge-based generative model exhibits enhanced mode coverage and superior sampling quality. Moreover, while deterministic approaches (Dai, Ruizhongtai Qi, and Nießner 2017; Peng et al. 2020; Rao, Nie, and Dai 2022) perform one-step mappings, BridgeShape refines shapes iteratively during the diffusion bridge process, significantly enhancing accuracy.

Evaluation on Unseen Object Categories

We evaluate BridgeShape's ability to generalize by comparing against state-of-the-art methods—3D-EPN (Dai, Ruizhongtai Qi, and Nießner 2017), Few-Shot (Wallace and

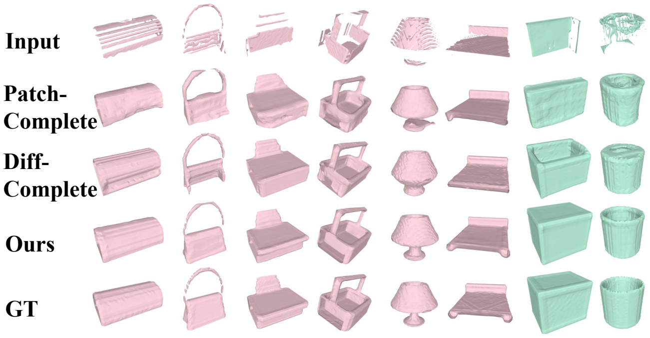

Figure 4: Qualitative comparison on the synthetic (pink) ShapeNet (Chang et al. 2015) dataset and real-world (green) ScanNet (Dai et al. 2017) dataset.

Mechanisms	l_1-err. ↓
Conventional diffusion	0.047
DSB	0.045
DSB + Stochasticity (Ours)	**0.039**

Table 4: Ablation on different diffusion mechanisms.

Hariharan 2019), IF-Nets (Chibane, Alldieck, and Pons-Moll 2020), Auto-SDF (Mittal et al. 2022), ConvONet (Peng et al. 2020), PatchComplete (Rao, Nie, and Dai 2022), Diff-Complete (Chu et al. 2023)—on synthetic ShapeNet (Chang et al. 2015) and real-world ScanNet data (Dai et al. 2017). As shown in Table 2 and Table 3, BridgeShape excels in handling unseen categories, outperforming them in terms of CD and IoU on average. This can be attributed to the fact that BridgeShape enforces globally coherent transport via the Schrödinger bridge formulation, effectively modeling shared structures among different categories. This leads to superior performance on unseen categories, including real-world scans that are typically cluttered and noisy. Figure 4 provides compelling visual evidence that BridgeShape robustly generalizes to unseen categories, producing high-fidelity completions that preserve fine geometric details on both synthetic and real-world datasets.

Ablation Studies

We perform component ablations on 3D-EPN (Dai, Ruizhongtai Qi, and Nießner 2017), reporting average performance across all categories unless otherwise noted.

Effects of DSB. One of the key innovations in BridgeShape is the DSB, which explicitly models optimal transport between the latent distributions of incomplete and complete shapes. To assess its impact, we replace the DSB with a conventional diffusion paradigm with an additional conditional branch, similar to DiffComplete (Chu et al. 2023). However, as shown in Table 4, this approach yields suboptimal results, underscoring the importance of explicitly modeling the transport process to achieve superior shape completions. Moreover, injecting stochasticity into the latent distribution of incomplete shapes can further boost performance.

Variants	Rec. l_1 ↓	Comp. l_1 ↓	Comp. IoU ↑
W/o. Depth	0.0042	0.0413	92.36%
W. Depth (Ours)	**0.0021**	**0.0389**	**94.41%**

Table 5: Ablation on the effect of Depth.

Categories	l_1-err. (32^3)	l_1-err. (64^3)	l_1-err. (128^3)
Chair	0.055	0.045	**0.044**
Table	0.059	**0.053**	0.054
Sofa	**0.047**	0.051	0.050
Lamp	0.038	0.034	**0.026**
Plane	0.012	**0.009**	**0.009**
Car	**0.023**	0.024	0.024
Dresser	**0.055**	0.062	0.058
Watercraft	0.022	**0.019**	**0.019**
Avg.	0.039	0.037	**0.036**

Table 6: Quantitative results on 3D-EPN. Low is better.

Effects of Depth-Enhanced VQ-VAE. The Depth-Enhanced VQ-VAE serves as the foundation for the DSB. Unlike a standard VQ-VAE (Razavi, van den Oord, and Vinyals 2019; van den Oord, Vinyals, and Kavukcuoglu 2017; Cheng et al. 2023), which captures a compact latent representation but struggles to retain structural details, our approach integrates self-projected multi-view depth information enriched with DINOv2 (Oquab et al. 2023) features to enhance geometric perception. To assess its effectiveness, we compare it against a standard VQ-VAE without depth features. As shown in Table 5, integrating multi-view depth into our VQ-VAE halves its reconstruction l_1 error ($0.0042 \rightarrow 0.0021$), yields a 0.002 reduction in completion l_1 error, and boosts IoU by $\sim$2% ($92.36\% \rightarrow 94.41\%$). Removing depth reverses these gains, underscoring the critical role of multi-view depth. Additional ablation studies and qualitative results are provided in the supplementary material.

Results on Higher Voxel Resolution

While our primary experiments use a voxel resolution of 32^3, we further evaluate on 3D-EPN (Dai, Ruizhongtai Qi, and Nießner 2017) at 64^3 and 128^3 to assess scalability (Table 6). Higher resolutions enable finer-grained shape completion and yield increasingly higher average accuracy. These results confirm that BridgeShape effectively leverages increased resolution to enhance completion quality. Qualitative results are provided in the supplementary material.

Conclusion

We present BridgeShape, a latent diffusion Schrödinger bridge framework for 3D shape completion. By enriching a compact latent space with multi-view depth features, BridgeShape efficiently transports incomplete shapes to complete ones, achieving high-fidelity and structurally consistent results. Experiments on large-scale benchmarks show superior performance, fine geometric preservation, and strong generalization to unseen categories.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. T2322012, No. 62572240).

References

Aguilar, J.; Baron, J. W.; Galla, T.; and Toral, R. 2022. Sampling rare trajectories using stochastic bridges. *Physical Review E*, 105(6): 064138.

Avetisyan, A.; Dahnert, M.; Dai, A.; Savva, M.; Chang, A. X.; and Nießner, M. 2019. Scan2cad: Learning cad model alignment in rgb-d scans. In *IEEE/CVF International Conference on Computer Vision*, 2614–2623.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chen, T.; Liu, G.-H.; and Theodorou, E. A. 2021. Likelihood training of schr\" odinger bridge using forward-backward sdes theory. *arXiv preprint arXiv:2110.11291*.

Chen, Y.; and Georgiou, T. 2015. Stochastic bridges of linear systems. *IEEE Transactions on Automatic Control*, 61(2): 526–531.

Chen, Y.; Georgiou, T. T.; and Pavon, M. 2021. Stochastic control liaisons: Richard sinkhorn meets gaspard monge on a schrodinger bridge. *Siam Review*, 63(2): 249–313.

Chen, Z.; He, G.; Zheng, K.; Tan, X.; and Zhu, J. 2023. Schrodinger bridges beat diffusion models on text-to-speech synthesis. *arXiv preprint arXiv:2312.03491*.

Cheng, Y.-C.; Lee, H.-Y.; Tulyakov, S.; Schwing, A. G.; and Gui, L.-Y. 2023. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *IEEE/CVF International Conference on Computer Vision*, 4456–4465.

Chibane, J.; Alldieck, T.; and Pons-Moll, G. 2020. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE/CVF International Conference on Computer Vision*, 6970–6981.

Chou, G.; Bahat, Y.; and Heide, F. 2023. Diffusion-sdf: Conditional generative modeling of signed distance functions. In *IEEE/CVF International Conference on Computer Vision*, 2262–2272.

Chu, R.; Xie, E.; Mo, S.; Li, Z.; Nießner, M.; Fu, C.-W.; and Jia, J. 2023. Diffcomplete: Diffusion-based generative 3d shape completion. *Advances in neural information processing systems*, 36: 75951–75966.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5828–5839.

Dai, A.; Ruizhongtai Qi, C.; and Nießner, M. 2017. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *IEEE/CVF International Conference on Computer Vision*, 5868–5877.

De Bortoli, V.; Thornton, J.; Heng, J.; and Doucet, A. 2021. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34: 17695–17709.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Han, X.; Li, Z.; Huang, H.; Kalogerakis, E.; and Yu, Y. 2017. High-resolution shape completion using deep neural networks for global structure and local geometry inference. In *IEEE/CVF International Conference on Computer Vision*, 85–93.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Léonard, C. 2013. A survey of the schr\" odinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*.

Li, A.; Zhu, Z.; and Wei, M. 2025. GenPC: Zero-shot Point Cloud Completion via 3D Generative Priors. In *IEEE/CVF International Conference on Computer Vision*, 1308–1318.

Li, B.; Xue, K.; Liu, B.; and Lai, Y.-K. 2023. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 1952–1961.

Liu, G.-H.; Vahdat, A.; Huang, D.-A.; Theodorou, E. A.; Nie, W.; and Anandkumar, A. 2023. I ^ 2 SB: Image-to-Image Schr\" odinger Bridge. *arXiv preprint arXiv:2302.05872*.

Luo, S.; and Hu, W. 2021. Diffusion probabilistic models for 3d point cloud generation. In *IEEE/CVF International Conference on Computer Vision*, 2837–2845.

Mittal, P.; Cheng, Y.-C.; Singh, M.; and Tulsiani, S. 2022. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 306–315.

Müller, N.; Siddiqui, Y.; Porzi, L.; Bulo, S. R.; Kontschieder, P.; and Nießner, M. 2023. Diffrf: Rendering-guided 3d radiance field diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4328–4338.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 8162–8171.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Peng, S.; Niemeyer, M.; Mescheder, L.; Pollefeys, M.; and Geiger, A. 2020. Convolutional occupancy networks. In *European Conference on Computer Vision*, 523–540.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*.

Rao, Y.; Nie, Y.; and Dai, A. 2022. Patchcomplete: Learning multi-resolution patch priors for 3d shape completion on unseen categories. *Advances in Neural Information Processing Systems*, 35: 34436–34450.

Razavi, A.; van den Oord, A.; and Vinyals, O. 2019. Generating Diverse High-Fidelity Images with VQ-VAE-2. In

Advances in Neural Information Processing Systems, volume 32, 14837–14847.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Shim, J.; Kang, C.; and Joo, K. 2023. Diffusion-based signed distance fields for 3d shape generation. In *IEEE/CVF International Conference on Computer Vision*, 20887–20897.

Smith, E. J.; and Meger, D. 2017. Improved adversarial systems for 3d object generation and reconstruction. In *Conference on Robot Learning*, 87–96.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2256–2265.

Somnath, V. R.; Pariset, M.; Hsieh, Y.-P.; Martinez, M. R.; Krause, A.; and Bunne, C. 2023. Aligned diffusion Schrödinger bridges. In *Uncertainty in Artificial Intelligence*, 1985–1995.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Stutz, D.; and Geiger, A. 2020. Learning 3d shape completion under weak supervision. *International Journal of Computer Vision*, 128: 1162–1181.

Vahdat, A.; Williams, F.; Gojcic, Z.; Litany, O.; Fidler, S.; Kreis, K.; et al. 2022. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35: 10021–10039.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, volume 30, 6306–6315.

Vogel, M.; Tateno, K.; Pollefeys, M.; Tombari, F.; Rakotosaona, M.-J.; and Engelmann, F. 2024. P2P-Bridge: Diffusion Bridges for 3D Point Cloud Denoising. In *European Conference on Computer Vision*, 184–201.

Wallace, B.; and Hariharan, B. 2019. Few-shot generalization for single-image 3d reconstruction via priors. In *IEEE/CVF International Conference on Computer Vision*, 3818–3827.

Wang, Z.; Xu, Q.; Tan, F.; Chai, M.; Liu, S.; Pandey, R.; Fanello, S.; Kadambi, A.; and Zhang, Y. 2024. Mvdd: Multiview depth diffusion models. In *European Conference on Computer Vision*, 236–253.

Wen, X.; Xiang, P.; Han, Z.; Cao, Y.-P.; Wan, P.; Zheng, W.; and Liu, Y.-S. 2023. PMP-Net++: Point Cloud Completion by Transformer-Enhanced Multi-Step Point Moving Paths. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 852–867.

Wu, R.; Chen, X.; Zhuang, Y.; and Chen, B. 2020. Multimodal shape completion via conditional generative adversarial networks. In *European conference on computer vision*, 281–296.

Yan, X.; Lin, L.; Mitra, N. J.; Lischinski, D.; Cohen-Or, D.; and Huang, H. 2022. Shapeformer: Transformer-based shape completion via sparse representation. In *IEEE/CVF International Conference on Computer Vision*, 6239–6249.

Yu, X.; Rao, Y.; Wang, Z.; Lu, J.; and Zhou, J. 2023. AdaPoinTr: Diverse Point Cloud Completion With Adaptive Geometry-Aware Transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 14114–14130.

Zhang, D.; Choi, C.; Park, I.; and Kim, Y. M. 2022. Probabilistic implicit scene completion. *arXiv preprint arXiv:2204.01264*.

Zhang, J.; Chen, X.; Cai, Z.; Pan, L.; Zhao, H.; Yi, S.; Yeo, C. K.; Dai, B.; and Loy, C. C. 2021. Unsupervised 3d shape completion through gan inversion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1768–1777.

Zheng, X.; Liu, Y.; Wang, P.; and Tong, X. 2022. SDF-StyleGAN: Implicit SDF-Based StyleGAN for 3D Shape Generation. In *Computer Graphics Forum*, volume 41, 52–63.

Zhou, L.; Du, Y.; and Wu, J. 2021. 3d shape generation and completion through point-voxel diffusion. In *IEEE/CVF International Conference on Computer Vision*, 5826–5835.

Zhu, Z.; Chen, H.; He, X.; Wang, W.; Qin, J.; and Wei, M. 2023. SVDFormer: Complementing Point Cloud via Self-view Augmentation and Self-structure Dual-generator. In *IEEE/CVF International Conference on Computer Vision*, 14508–14518.

Zhu, Z.; Chen, H.; He, X.; and Wei, M. 2025. PointSea: Point Cloud Completion via Self-structure Augmentation. *International Journal of Computer Vision*, 133(7): 4770–4794.

Zhu, Z.; Nan, L.; Xie, H.; Chen, H.; Wang, J.; Wei, M.; and Qin, J. 2024. CSDN: Cross-Modal Shape-Transfer Dual-Refinement Network for Point Cloud Completion. *IEEE Transactions on Visualization and Computer Graphics*, 30(7): 3545–3563.

Zollhöfer, M.; Stotko, P.; Görlitz, A.; Theobalt, C.; Nießner, M.; Klein, R.; and Kolb, A. 2018. State of the art on 3D reconstruction with RGB-D cameras. In *Computer Graphics Forum*, volume 37, 625–652.