ScaleBlind: POINT CLOUD COMPLETION UNDER UNKNOWN SCALE

Anonymous authors
Paper under double-blind review

ABSTRACT

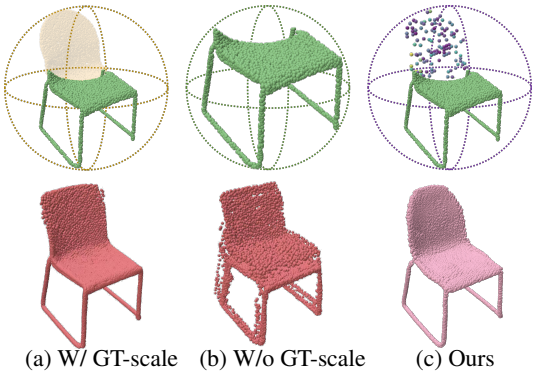
Point cloud completion aims to infer a complete 3D shape from a partial point cloud and serves as a fundamental building block for downstream tasks such as reconstruction, editing, and simulation. Despite the recent progress, existing learning-based methods often implicitly rely on access to the ground-truth shape scale (GT-scale) during both training- and testing-time normalization, assuming privileged information that is unavailable in real-world inference. This hidden assumption limits practical deployment and can lead to severe completion artifacts, *e.g.*, over- or under-completion and nested shells, once the oracle GT-scale cue is removed (see examples at Fig. B1). We observe that the recent foundation image generation models exhibit a strong capability of understanding objects and geometries, and producing multi-view consistent renderings, making them promising priors for GT-scale-free 3D completion. Motivated by this insight, we propose *ScaleBlind*, a novel framework that leverages foundation-model-based image completion to recover global scale directly from partial inputs and then faithfully produces the 3D completion. Specifically, *ScaleBlind* dreams out complete multi-view appearances from rendered partial views, lifts the inferred missing regions back into 3D to obtain a geometry-aware coarse completion, and further refines it via a powerful cross-modal fusion network with the original partial point cloud. By harnessing 2D foundation priors, our method eliminates the need for accessing GT-scale information at inference. Moreover, it provides a principled bridge between 2D generative priors and 3D point cloud completion. Extensive experiments demonstrate the superiority of our framework, making *ScaleBlind* the new state-of-the-art for the point cloud completion task.

1 INTRODUCTION

Point cloud completion, *i.e.*, recovering a complete 3D shape from a partial point cloud, is a long-standing problem with broad impact in graphics and vision. It is a critical component for downstream applications such as reconstruction pipelines Yang et al. (2025); Huang et al. (2023), robot manipulation Varley et al. (2017), and AR/VR content creation Guo et al. (2022), where partial observations are the norm. However, real-world point clouds acquired from commodity depth sensors or LiDAR are frequently incomplete due to occlusion, limited viewpoints, and sparse sampling, making completion both practically necessary and intrinsically ambiguous.

A large body of learning-based methods tackles completion by learning a direct mapping from partial inputs to full shapes. They extract global and local features and decode missing geometry in a coarse-to-fine manner. Beyond the ambiguity caused by severe occlusions, we observe a protocol issue that can simplify the task in synthetic benchmarks. That is, in many pipelines Yuan et al. (2018); Chen et al. (2023); Xiang et al. (2021); Zhou et al. (2022); Wen et al. (2022), shapes are canonicalized (*e.g.*, , centered and scaled into a unit sphere) *before* partial observations are generated. This practice implicitly ties the partial input to the ground-truth global scale and placement (see Fig. 1(a)), introducing a privileged cue that is not accessible at test time in practical scenarios. When this oracle scale cue is removed (*i.e.*, GT-scale-independent preprocessing, Fig. 1(b)), the performance of existing completion models can deteriorate sharply, often manifesting as global-scale distortions and over-completion artifacts.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) W/ GT-scale (b) W/o GT-scale (c) Ours

Figure 1: **GT shape scale and completion.** Given a partial point cloud, (a) existing learning-based methods (*e.g.*, GeoFormer Yu et al. (2024)) often assume access to the ground-truth shape scale during testing-time normalization (*e.g.*, the missing yellow back part), which provides privileged oracle information and leads to seemingly satisfactory completions. (b) In realistic settings where such information is unavailable in the normalization step, the resulting completions exhibit clear artifacts, such as over-completion and nested shells. (c) Motivated by this observation, we propose a novel solution to *'Dream'* the missing scale information at inference (*i.e.*, the back part), enabling faithful and scale-consistent completion without GT access.

In parallel, recent advances in 3D generative modeling enable high-quality shape synthesis from image or text prompts Xiang et al. (2025); Wu et al. (2026). A natural idea is to render a partial point cloud into a set of views, complete the partial renderings using strong pretrained image models Li et al. (2024); Liu et al. (2024), and then reconstruct a full 3D shape. However, these generative systems are not designed for completion: their outputs are not explicitly constrained to remain consistent with the observed partial geometry. As illustrated in Fig. B1(d), the wing region of the airplane contains irrational connection structures and suffers from the loss of partial geometric details. Moreover, the generated model exhibits a certain degree of scale deviation. For instance, the overall thickness of the chair (Fig. B1(d)) differs from that of the original model. Such deviations are particularly problematic in the unknown-scale setting, where scale and placement must be recovered rather than assumed.

Motivated by these observations, we propose *ScaleBlind*, a GT-scale-free point cloud completion framework designed for inference without access to ground-truth scale information. Our key insight is that recent foundational image generation models possess strong geometric understanding and can reason out missing object evidence with high multi-view consistency, providing a powerful prior for recovering the global extent of a shape from partial geometry. Specifically, *ScaleBlind* renders the partial point cloud from multiple viewpoints and completes the resulting multiple renderings using a pre-trained large VLM (*i.e.*, Nano-banana Team et al. (2023)), yielding plausible full-view appearances and missing regions. We then lift these multi-view completions into 3D via a visual-hull-alike voxel voting scheme, forming a probabilistic volumetric prior. This prior is further used to extract high-confidence missing point candidates and to estimate an oriented bounding box (OBB) for joint normalization together with the partial input. Finally, the normalized partial points and confidence-weighted candidates are fused and processed by a prior-guided coarse-to-fine completion network, where high-level multi-view features first guide the generation of a coarse geometric scaffold, followed by two refinement stages that progressively inject mid-/low-level image cues to recover detailed structures and geometry, producing the final completion while preserving the observed geometry.

Crucially, our framework excels on two parallel fronts: (i) establishing a state-of-the-art 3D completion architecture, and (ii) addressing the practical GT-scale-free challenge. This robust geometric capability stems from our novel cascaded refinement network. It first constructs a coarse geometric scaffold guided by global semantics, and then progressively carves out fine-grained details via a powerful dual-mode fusion strategy. By simultaneously measuring 3D geometric discrepancies and injecting hierarchical 2D image cues, the network effectively prevents wild hallucinations and faithfully recovers intricate structures. Consequently, extensive experiments demonstrate the dual superiority of our approach: under the conventional GT-scale-tied setting, *ScaleBlind* sets a new state-of-the-art, validating the representational power of our core architecture; simultaneously, under the realistic GT-scale-free setting, it exhibits a commanding advantage, producing well-proportioned completions free from severe geometric distortions.

We defer related work discussions to Appendix B. To summarize, our principal contributions are three-fold:

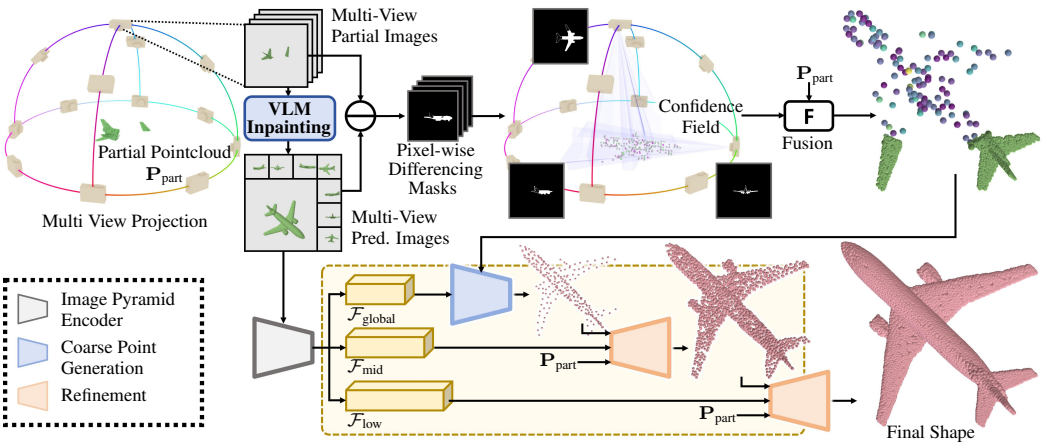


Figure 2: **Method Overview.** Our method consists of two core modules - prior acquisition (top, Sec. 2.2) and prior-guided point completion (bottom, Sec. 2.3). *Prior acquisition*: given a partial point cloud $\mathbf{P}_{\text{part}}$ (the gray silhouette reveals the large missing portion), our method first exploits a pretrained VLM to inpaint the partial renders and then obtains the complete shape prior via a visual-hull-like voxel voting. *Point completion*: guided by the full shape prior, our method first produces a complete coarse point cloud, and then refines and consolidates the shape in two refinement steps by injecting hierarchical image cues.

- We formulate *GT-scale-free* point cloud completion under unknown global scale, and reveal that GT-scale-tied canonicalization in existing pipelines introduces an unrealistic *scale prior leakage* into partial inputs.
- We propose to recover missing global extent by leveraging a pretrained VLM to complete multi-view renderings of the partial scan, and lift the multi-view evidence into a scale-consistent probabilistic 3D prior via voxel voting.
- We incorporate the recovered prior into a prior-guided coarse-to-fine completion network with joint normalization and confidence-aware fusion, yielding faithful completions that remain consistent with the observed geometry under GT-scale-independent preprocessing.

2 Method

2.1 Overview

Problem definition. Given the partial point cloud input $\mathbf{P}_{\text{part}}$, the goal of point cloud completion is to produce the full shape $\mathbf{P}_{\text{full}}$, respecting the partial observation. In the literature of supervised point cloud completion, a common preprocessing step is to normalize the partial input (*i.e.*, centering and scaling into a unit sphere). When the missing region is structural-preserving (*e.g.*, the middle of a chair leg), the on-the-fly normalization of the partial input produces consistent scale and placement as in the ground truth point cloud. However, as shown in Fig. 1(a), when the missing region is structure-disruptive (*e.g.*, the chair back part), GT scale information is still incorporated to match the training time operation, so as to produce satisfactory shapes. Clearly, this scale information is inaccessible in real-world inference, causing severe performance degradation (Fig. 1(b)). We introduce a GT-scale-free point cloud completion task, where the observed partial point cloud is independently normalized, and a prior acquisition module is employed to predict GT-scale cues from the partial observation rather than assuming them.

Method. To solve this new task, we propose *ScaleBlind*, consisting of two core modules - prior acquisition and prior-guided completion. Figure 2 illustrates a method overview, and in what follows, we elaborate on technical details.

2.2 Prior Acquisition

The goal of this module is to estimate the coarse full shape prior from the partial observation in the following four steps.

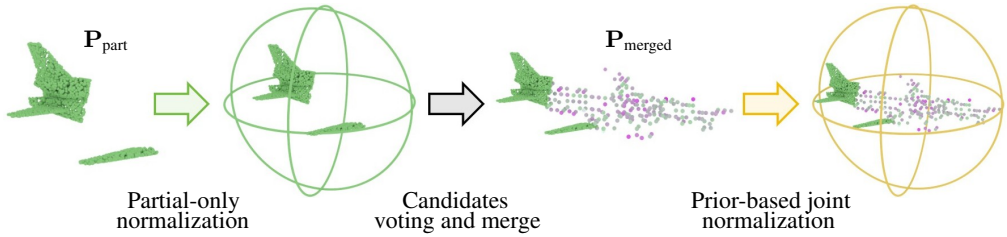


Figure 3: **Illustration of our normalization strategies.** We first apply a *partial-only normalization* to the input partial point cloud $\mathbf{P}_{\text{part}}$. After extracting and merging the candidate points to form $\mathbf{P}_{\text{merged}}$, a *prior-based joint normalization* is performed to dynamically adjust the scale based on the combined spatial extent of the partial input and the generated geometric priors.

Multi-view projection and completion. As illustrated in Fig. 2(top), we sample a set of viewpoints $\mathcal{V}$ on the upper hemisphere surrounding the partial input and place the camera at a sufficiently large distance (default r=5). Unlike conventional pipelines that render from a unit sphere, our framework performs partial-only normalization, thus the global extent of the missing geometry is unknown *a priori*. We therefore adopt a conservative camera distance, chosen empirically for extremely high missing ratios, to provide a sufficiently wide field of view that covers the potential completion region. By maintaining this camera distance, we prevent the VLM-hallucinated structures from extending beyond the image boundaries. Subsequently, we apply Nano-banana Team et al. (2023) to each viewpoint to synthesize completed projections $\{I_{\text{comp},v}\}_{v\in\mathcal{V}}$ that capture the approximate global shape, thereby generating plausible missing regions in the 2D domain. See *supplementary* for the detailed prompt when interacting with the VLM.

To detect the inpainted regions that indicate the missing geometry, we compute a *mask* M_v per view by pixel-wise differencing:

$$M_v(i,j) = \begin{cases} 1, & \text{if } |I_{\text{comp},v}(i,j) - I_{\text{part},v}(i,j)| > \delta_{\text{diff}}, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where δ_{diff} is a fixed intensity threshold.

Geometric consistency check. VLM completion may occasionally introduce view-wise structures and scales that are incompatible with the observed evidence. We therefore filter unreliable viewpoints using a simple overlap test. Let $B_{p,v}$ and $B_{c,v}$ denote the binary foreground masks of the partial and completed projections, respectively. We compute the overlap ratio:

$$r_v = \frac{\sum_{(i,j)} (B_{p,v} \cap B_{c,v})}{\sum_{(i,j)} B_{p,v}}, \tag{2}$$

and keep a view in the valid set $\mathcal{V}_{\text{valid}}$ only if $r_v \geq \tau_{\text{overlap}}$, which suppresses clearly misaligned hallucinations.

Multi-view voting. To get the 3D shape prior, we lift the multi-view completions into 3D using a visual-hull-like voxel voting scheme. Following the partial-only normalization strategy illustrated in Fig. 3, we define a canonical voxel grid $\mathcal{G}$ over the input partial point cloud space. This grid is bounded within a cubic coordinate range of $[-5, 5]^3$ and parameterized by a voxel resolution Δ. For each voxel center $\mathbf{x} \in \mathbb{R}^3$ and each valid view v, we project $\mathbf{x}$ onto the image plane:

$$\mathbf{p}_v = \pi(\mathbf{K}(\mathbf{R}_v\mathbf{x} + \mathbf{t}_v)), \tag{3}$$

where $\pi(\cdot)$ is the perspective projection operator, and $\mathbf{K}$, $(\mathbf{R}_v, \mathbf{t}_v)$ are the camera intrinsics and extrinsics. A voxel $\mathbf{x}$ receives a vote from view v if its projection falls inside the foreground of M_v. The accumulated votes are:

$$S(\mathbf{x}) = \sum_{v\in\mathcal{V}_{\text{valid}}} M_v(\lfloor \mathbf{p}_v \rfloor), \tag{4}$$

where $\lfloor\cdot\rfloor$ maps image coordinates to pixel indices.

We normalize the voting volume into a probabilistic field:

$$\mathrm{P}_{\text{prior}}(\mathbf{x}) = \frac{S(\mathbf{x})}{\max_{\mathbf{x}'\in\mathcal{G}} S(\mathbf{x}')}, \tag{5}$$

where each voxel value $\mathbf{P}_{\text{prior}}(\mathbf{x}) \in [0, 1]$ quantifies the consensus confidence of the missing geometry supported by multi-view VLM completions. This probabilistic field serves as a shape-aware and scale-consistent prior for completion.

Prior-based joint normalization. Given the cubic spatial extent of $[-5, 5]^3$ and the voxel resolution $\Delta = 0.02$, the predicted probabilistic field is formulated as a volumetric tensor $\mathrm{P}_{\text{prior}} \in \mathbb{R}^{500\times500\times500}$. From this tensor, we further extract a concrete high-confidence candidate point set by applying a probability threshold τ_{prob}. These candidate points provide a coarse estimation of the missing geometry. We then jointly normalize the partial point cloud and the candidate set into a unit sphere according to their combined spatial extent, yielding $\tilde{\mathbf{P}}_{\text{part}}$ and $\tilde{\mathbf{P}}_{\text{cand}}$. Importantly, this normalization does not require any GT-scale information, yet yields a faithful global-scale normalization by recovering the coarse missing extent from multi-view priors.

2.3 Prior-guided Completion

Given the normalized partial point cloud $\tilde{\mathbf{P}}_{\text{part}}$ and $\tilde{\mathbf{P}}_{\text{cand}}$, we aim to predict a complete point set that follows the inferred global structure while preserving the observed geometry. We design the completion module in a coarse-to-fine manner using a bespoke neural network – it first predicts a coarse scaffold and then refines it in two cascaded refinement steps (see Fig. 2(bottom)).

Partial and candidate points fusion. To initialize the completion process, we first merge the input partial points with the generated prior candidates to form a fused set:

$$\tilde{\mathbf{P}}_{\text{merged}} = \texttt{FPS}\left(\tilde{\mathbf{P}}_{\text{part}} \cup \tilde{\mathbf{P}}_{\text{cand}}, N\right) \in \mathbb{R}^{N\times4}. \tag{6}$$

Other than the coordinate, each fused point has an extra confidence attribute σ. We set $\sigma = 1.0$ for input partial points, while querying $\mathrm{P}_{\text{prior}}$ for points in $\tilde{\mathbf{P}}_{\text{cand}}$ at their respective locations. Finally, we apply farthest point sampling (FPS) to obtain a compact and uniform set for the completion network.

Visual prior encoding. To extract multi-view visual cues, we utilize a PointNet++ Qi et al. (2017b) encoder to extract a comprehensive global point feature $\mathcal{F}_p \in \mathbb{R}^{256}$ from the fused set $\tilde{\mathbf{P}}_{\text{merged}}$. This feature serves as a geometric reference for subsequent multi-view feature aggregation.

In parallel, we employ an *image pyramid encoder* based on a ResNet backbone He et al. (2016) to process the N_v completed projections in $\mathcal{V}$. This encoder yields a feature pyramid $\{\mathcal{F}_{\text{low}} \in \mathbb{R}^{N_v\times32}, \mathcal{F}_{\text{mid}} \in \mathbb{R}^{N_v\times64}, \mathcal{F}_{\text{global}} \in \mathbb{R}^{N_v\times128}\}$, capturing semantics ranging from the detailed local geometry and appearance to the coarse global shape.

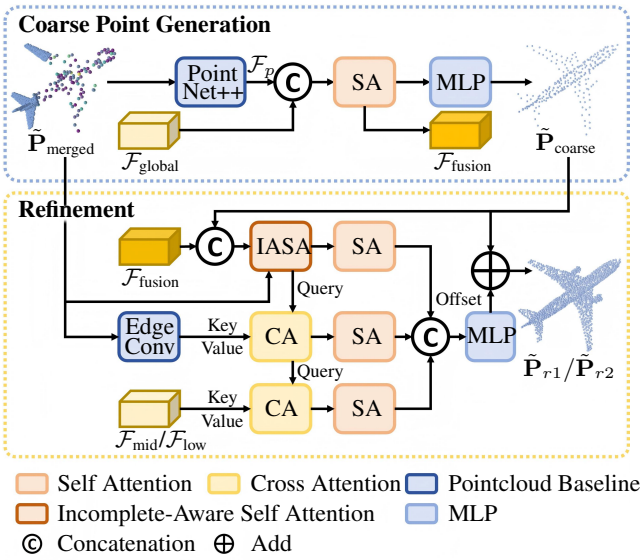


Figure 4: **Coarse points generation and refinement.** We design a coarse point generation module with a cross-attention-based fusion mechanism, and a cascaded two-stage refinement network to produce the final consolidated complete point cloud.

Coarse point generation. Figure 4(top) illustrates the detailed network structure of our point generator. Specifically, we fuse the high-level multi-view feature $\mathcal{F}_{\text{global}}$ with the global point feature $\mathcal{F}_p$ via a self-attention (SA) mechanism. After fusing, an updated fused feature representation $\mathcal{F}_{\text{fusion}} \in \mathbb{R}^{256}$ directly predicts an initial coarse completion $\tilde{\mathbf{P}}_{\text{coarse}} \in \mathbb{R}^{256\times3}$ via a few MLP layers. $\tilde{\mathbf{P}}_{\text{coarse}}$ provides a global scaffold guided by the inferred shape priors, and is structurally plausible, while remaining anchored to the partial points in $\tilde{\mathbf{P}}_{\text{merged}}$.

Cascaded refinement. We next design a cascaded refinement network to *upsample* and *consolidate* $\tilde{\mathbf{P}}_{\text{coarse}}$. Specifically, as shown in Fig. 4(bottom), the network takes as input $\tilde{\mathbf{P}}_{\text{coarse}}$, $\tilde{\mathbf{P}}_{\text{merged}}$, $\mathcal{F}_{\text{fusion}}$, and $\{\mathcal{F}_{low}, \mathcal{F}_{mid}\}$. With all the input, we aim to fuse information from both the point clouds and guiding images to drive the completion. To this end, we design two fusion modes.

Mode one: point cloud feature fusion. Firstly, we concatenate $\mathcal{F}_{\text{fusion}}$ to the point feature of each point in $\tilde{\mathbf{P}}_{\text{coarse}}$. Following Zhu et al. (2023), we incorporate their Incomplete-Aware Self-Attention (IASA) module alongside a standard self-attention mechanism. By explicitly comparing and cross-referencing the coarse prediction $\tilde{\mathbf{P}}_{\text{coarse}}$ against the merged partial input $\tilde{\mathbf{P}}_{\text{merged}}$, these modules produce the fused features. With the IASA and SA, we explicitly measure the discrepancy between the confidence-weighted inferred geometry and the coarse shape. Secondly, we further process $\tilde{\mathbf{P}}_{\text{merged}}$ using a DGCNN Wang et al. (2019) feature extractor (*i.e.*, EdgeConv) to capture more fine-grained local geometric structures, which is used as key and value for the following cross-attention (CA), where the query tokens are obtained from the fused IASA features. A further SA is employed to refine the second feature. With the CA and SA, an enhanced shape-aware feature is produced. The two resulting features in this mode are capable of capturing the missing region and the whole shape.

Mode two: point and image feature fusion. In this mode, the pyramid image feature and the point feature go through a cross-attention and self-attention blocks to produce the fused feature. Since the image features capture shape appearance and geometry, especially $\mathcal{F}_{\text{low}}$, the resulting fused feature is essential to enhance geometric details.

The three features are next concatenated and sent to a few `MLP` layers to produce upsampled (x8) offset vectors Xiang et al. (2021), which are then added to the eight replicated copies of $\tilde{\mathbf{P}}_{\text{coarse}}$. As a result, the first refinement creates an upsampled and consolidated point cloud $\mathbf{P}_{r1} \in \mathbb{R}^{(256\times 8)\times 3}$, while the second refinement produces the final resulting point cloud $\mathbf{P}_{r2} \in \mathbb{R}^{(256\times 8\times 8)\times 3}$. Note that the two refinement blocks have the identical structure, the only difference is that $\tilde{\mathbf{P}}_{\text{coarse}}$ and $\mathcal{F}_{\text{mid}}$ are used in the first refinement block, while $\mathbf{P}_{r1}$ and $\mathcal{F}_{\text{low}}$ are used in the second refinement block.

2.4 Network Training

The completion network, consisting of a point generation block followed by two refinement blocks, is trained end-to-end. Our objective encourages geometric fidelity to the ground-truth shape while strictly preserving consistency with the observed partial input. Specifically, we minimize the Chamfer Distance (CD) between the predicted point sets ($\tilde{\mathbf{P}}_{\text{coarse}}, \tilde{\mathbf{P}}_{r1}, \tilde{\mathbf{P}}_{r2}$) and the ground-truth complete shape $\mathbf{P}_{gt}$. The total loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{CD}}(\tilde{\mathbf{P}}_{\text{coarse}}, \mathbf{P}_{gt}) + \mathcal{L}_{\text{CD}}(\tilde{\mathbf{P}}_{r1}, \mathbf{P}_{gt}) + \mathcal{L}_{\text{CD}}(\tilde{\mathbf{P}}_{r2}, \mathbf{P}_{gt}). \tag{7}$$

3 Experiments and Results

Before delving into the detailed evaluations, it is important to clarify the geometric implications of our partial point cloud generation protocol. Depending on the missing regions, the resulting partial inputs typically fall into two distinct geometric scenarios. The first scenario involves missing regions located primarily in the central or internal parts of the object (*e.g.*, the 25% and 50% missing cases in Fig. F2). Because the outer extremities remain largely intact, these inputs implicitly preserve the original global scale. The second scenario involves missing regions at the extremities or outer boundaries, resulting in a complete loss of the ground-truth global scale. While our proposed framework is highly robust and seamlessly handles both scenarios, we purposefully highlight the more demanding extremity-missing cases in the majority of our qualitative visualizations (*e.g.*, figs. 5, E1, E2, F1, F3 and F4). This design choice intuitively illustrates the severe completion challenges introduced by absolute scale ambiguity. Additional implementation details are provided in Sec. D.2 of the supplementary material.

Dataset. We evaluate our method on PCN Yuan et al. (2018) and ShapeNet-55 Chang et al. (2015). Partial inputs are generated by removing 25%, 50%, or 75% of points around randomly selected seed points using KNN, and the same construction protocol is applied to both datasets. Additional dataset and construction details are provided in Sec. D of the supplementary material.

Evaluation Metric. To quantitatively evaluate the completion performance, we employ three standard metrics: Chamfer Distance (CD), F-Score, and Density-aware Chamfer Distance (DCD) Wu et al. (2021). CD measures the global geometric distance between predicted and ground-truth point sets, but is sensitive to outliers and ignores local density variations. F-Score supplements CD to comprehensively assess visual fidelity and surface reconstruction quality. Unlike vanilla CD, DCD is designed to be more sensitive to the spatial distribution, penalizing uneven point clusters and encouraging more realistic geometric detail.

3.1 Comparison: Direct Point Cloud Completion

To verify the effectiveness of our proposed method, we conduct a comprehensive comparison with several state-of-the-art completion frameworks, including AdaPoinTr Yu et al. (2023), SVDFormer Zhu et al. (2023), CRA-PCN Rong et al. (2024), GeoFormer Yu et al. (2024), SIMECO Wang et al. (2026), and PCDreamer Wei et al. (2025). The compared baselines are grouped according to whether their original frameworks support text prompts as input. AdaPoinTr, SVDFormer, CRA-PCN, GeoFormer, and SIMECO do not accept text prompts, whereas PCDreamer natively supports category-related textual input. In addition, SIMECO explicitly addresses scale ambiguity and can operate without external scale information, while the other baselines do not explicitly recover the unknown object scale. The detailed treatment of scale information under different evaluation protocols is described below. All baseline networks are retrained from scratch on our benchmark using their default settings. We report results under two preprocessing protocols:

(1) *GT-scale-free normalization (unknown scale).* We normalize each partial input using the scale inferred by our *prior acquisition* module, and apply the *same* estimated scale to all competing methods. Note that SIMECO Wang et al. (2026) addresses the GT scale ambiguity, we evaluate them exclusively under the *GT-scale-free* setting (without normalizing their partial inputs using the scale inferred by our *prior acquisition* module). Unless otherwise specified, the visual results shown in this paper are produced under this setting.

(2) *GT-scale-tied normalization (oracle scale).* We additionally report results under the standard synthetic protocol, where normalization uses the ground-truth full-shape scale. Note that, although the setting is impractical, it provides a performance upper bound for all methods, and also *validates the effectiveness of our completion network*.

We additionally consider a third setting, termed *on-the-fly partial-based normalization*, where the input partial observation is normalized using only its own spatial extent at inference time. Under this protocol, existing baselines—originally trained with GT-scale-tied normalization and thus relying on oracle scale cues—suffer from a substantial distribution shift and yield severely degraded completions. For completeness, we report preliminary qualitative results under this setting in Figure F5; however, due to the consistently poor performance and the lack of meaningful quantitative comparability, we do not include this setting in the main quantitative comparison.

When invoking the vision-language model (VLM) for image completion, we include the *object category* in the text prompt to help the VLM better infer and reconstruct the missing regions at the image level (see supplementary). Since the resulting VLM completion is used for scale estimation, the estimated scale is shared with the applicable competing methods under the GT-scale-free protocol described above, rather than being exclusive to our method. Regarding textual input, PCDreamer natively supports category-related prompts and is therefore provided with the same category information as our method, while no additional textual input is introduced to baselines whose original architectures do not support text prompts.

Evaluation on the PCN dataset. Table 1 reports results under both the GT-scale-free and the conventional GT-scale-tied settings. *ScaleBlind* delivers the strongest performance in both protocols and is the only method that consistently ranks first across all eight categories under the GT-scale-free setting. This setting is practical for real-world applications, and our superior performance indicates that our completion network does not rely on GT-scale canonicalization to recover global extent, structure, and geometry. The gains are most evident from categories that are sensitive to scale cues (e.g., *Plane*, *Lamp*, and *Boat*), where large missing regions often disrupt the object span and amplify the ambiguity of completion. More importantly, *ScaleBlind* is substantially more stable when moving from GT-scale-tied to GT-scale-free preprocessing. While prior methods exhibit a clear performance drop once the oracle scale cue is removed, *ScaleBlind* maintains competitive accuracy with only a modest degradation. Figure 5 shows typical examples from all competing methods. The first five

Table 1: **Quantitative comparison on the PCN dataset under the *GT-scale-free* and *GT-scale-tied* settings**. (ℓ^1 CD $\times 10^3$ and F-Score@1%). Shapes from eight categories are used in the evaluation. We report the CD-Avg↓ for each category, and additionally report the CD-Avg↓, DCD-Avg↓, F1↑ across all categories. The compared methods are divided into two groups according to whether their original frameworks support text prompts as input: methods in the first group do not support text-prompt input, whereas methods in the second group can receive category-related text prompts, as does our method. For each cell, the two numbers are for the GT-scale-free and GT-scale-tied settings, respectively. We use the dark green and dark blue to indicate the best results for evaluations under GT-scale-free and GT-scale-tied settings, while using the light green and light blue to indicate the second best.

Methods	Plane	Cabinet	Car	Chair	Lamp	Couch	Table	Boat	CD-Avg↓	DCD-Avg↓	F1↑
AdaPoinTr*	4.28 / 3.81	8.11 / 7.75	7.58 / 7.12	6.99 / 6.97	9.45 / 7.96	6.51 / 6.32	6.71 / 5.98	6.94 / 5.75	7.07 / 6.46	0.530 / 0.528	0.826 / 0.852
SVDFormer*	4.38 / 3.95	7.98 / 7.39	7.69 / 6.98	7.13 / 6.97	9.42 / 8.14	7.06 / 6.21	6.77 / 5.73	6.96 / 5.82	7.18 / 6.40	0.532 / 0.510	0.818 / 0.857
CRA-PCN*	4.20 / 4.08	7.55 / 6.84	7.65 / 7.15	7.08 / 6.13	8.89 / 7.12	6.67 / 6.04	6.62 / 5.50	6.83 / 5.70	6.94 / 6.10	0.523 / 0.507	0.834 / 0.862
GeoFormer*	4.87 / 3.90	7.85 / 7.18	7.77 / 6.90	7.50 / 6.98	9.17 / 8.09	6.76 / 5.93	6.98 / 5.56	7.09 / 5.78	7.25 / 6.29	0.535 / 0.508	0.814 / 0.869
SIMECO	4.72 / - - -	8.43 / - - -	8.03 / - - -	6.97 / - - -	10.26 / - - -	7.45 / - - -	7.71 / - - -	6.97 / - - -	7.56 / - - -	- - - / - - -	0.774 / - - -
PCDreamer*	4.23 / 3.75	7.88 / 6.93	7.53 / 6.88	7.45 / 6.14	9.03 / 7.81	6.66 / 5.94	6.76 / 5.52	6.85 / 5.69	7.05 / 6.08	0.528 / 0.503	0.829 / 0.871
Ours	3.88 / 3.56	7.39 / 7.03	7.08 / 6.84	6.49 / 6.09	7.34 / 7.10	6.18 / 5.87	5.72 / 5.44	5.89 / 5.66	6.21 / 5.95	0.514 / 0.499	0.871 / 0.883

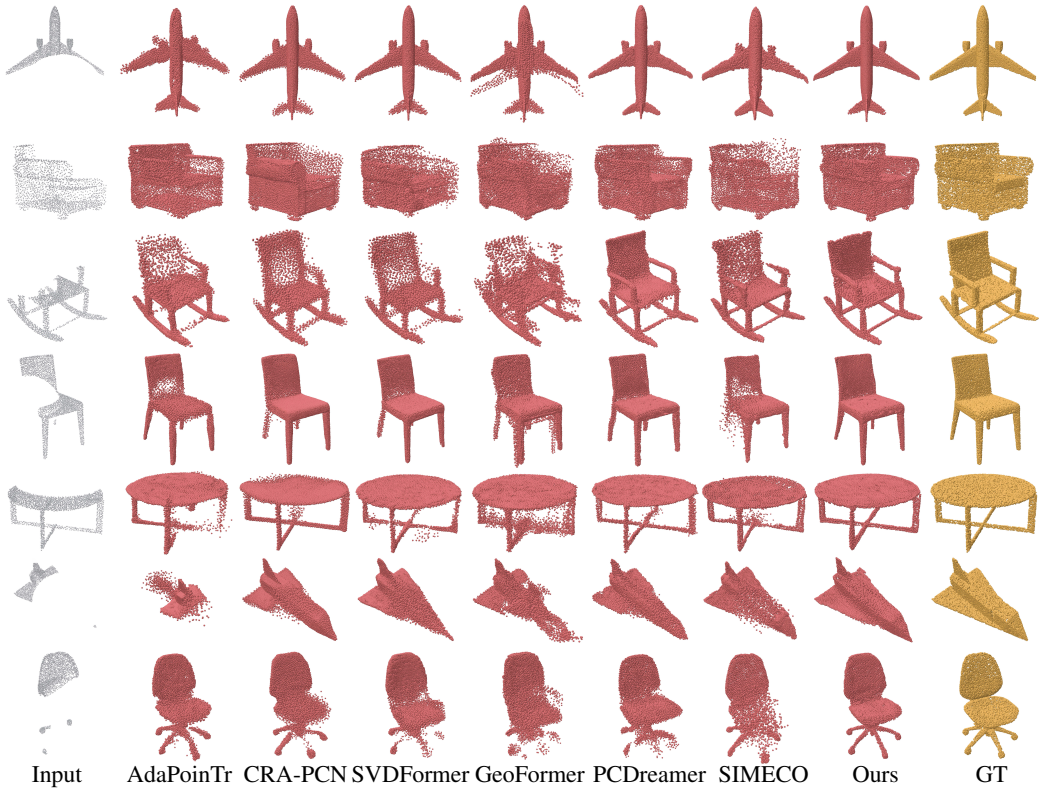


Figure 5: **Qualitative comparison on the PCN dataset under the GT-scale-free setting**. Given the inputs with varying missing regions, baseline methods produce inferior completion results with clear artifacts, while our method completes objects faithfully.

rows correspond to partial inputs with 50% missing, while the last two rows use the more challenging 75% setting. When the partial input still preserves a reliable scale cue (*e.g.*, the *Chair* in the fourth row, where the missing region is mostly internal), most methods perform reasonably well. In contrast, at 75% missing, large holes often remove key evidence about the global extent (*e.g.*, the *Plane* in the sixth row). In this case, many baselines degrade and produce noisy points, over-completion, or missing parts.

Evaluation on the ShapeNet-55 dataset. We also conduct the experiments on the ShapeNet-55 dataset, where we observe similar conclusions from the evaluations as in the PCN benchmark. Due to the page limit, the quantitative and qualitative results are in the supplementary.

3.2 Comparison: Generative Point Cloud Completion

We further compare *ScaleBlind* with Hi3DGen Ye et al. (2025) and Trellis Xiang et al. (2025). Although these image-to-3D generators produce plausible global shapes, they often fail to preserve geometry consistent with the observed partial input. In contrast, *ScaleBlind* explicitly fuses multi-view priors with partial geometry and produces more faithful, observation-aligned completions. Detailed experimental settings and qualitative comparisons are provided in Fig. E2 of the supplementary.

3.3 Ablation Study

We conduct extensive ablation experiments to validate our core design choices. All quantitative and qualitative evaluations are performed on the PCN dataset.

Impact of the number of viewpoints. Image-based guidance is critical for high-fidelity point cloud completion, and the number of views is a key hyperparameter affecting the final results. To investigate this, we generated 1-18 candidate viewpoints for each object and systematically evaluated model performance across different view counts. Detailed results are provided in Fig. E3 of the supplementary material.

Table 2: **Statistical ablation results**. Statistics on CD-Avg↓, DCD-Avg↓, F1↑ are reported from the ablated variants of our method (ℓ^1 CD $\times 10^3$ and F-Score@1%).

Conf. Att.	Hier. Feat.	CD-Avg↓	DCD-Avg↓	F1↑
	✓	6.32	0.519	0.866
✓		6.41	0.522	0.863
✓	✓	**6.21**	**0.514**	**0.871**

Efficacy of the confidence attribute. The confidence attribute (denoted as Conf. Att.) is introduced to modulate the contribution of prior candidates, down-weighting uncertain regions while preserving the reliable evidence and partial geometry. We ablate this component by removing the confidence values and feeding the network only the merged point set with uniform weights. As shown in Tab. 2, performance drops consistently when the confidence attribute is disabled, indicating that explicitly encoding confidence is important for suppressing noisy prior candidates and improving the stability of prior-guided completion under GT-scale-free preprocessing.

Impact of hierarchical image features. To assess the role of the hierarchical image features (denoted as Hier. Feat.), we ablate our feature injection by using only the global image feature ($\mathcal{F}_{\text{global}}$) in the refinement block, while keeping all other settings unchanged. As reported in Tab. 2, this variant consistently underperforms the full model, confirming that hierarchical features are important for detailed reconstruction. In particular, global features primarily constrain the overall extent and coarse structure, whereas mid/low-level cues provide the local geometric evidence needed to refine thin parts and recover fine-grained details.

Additional evaluations and discussion. We provide additional results in the supplementary, including single-view scanned partial points (Sec. F.1), robustness to input noise and missing ratios (Sec. F.2), and real-world LiDAR completion on ScanNet Dai et al. (2017) and KITTI Geiger et al. (2013) (Sec. F.3). We also compare with generative completion methods GenPC Li et al. (2025) and SDS-Complete Kasten et al. (2023) under the GT-scale-free setting (Sec. F.4), and further discuss plausible variations from the ground truth (Sec. F.5) and VLM-induced failure cases (Sec. F.7).

4 Conclusion

We propose *ScaleBlind*, a novel point cloud completion framework that achieves breakthroughs on two parallel fronts: establishing a highly effective state-of-the-art 3D completion architecture, and solving the practical GT-scale-free challenge. Our method leverages multi-view rendering and foundation-model-based image completion to infer missing extents, lifts the multi-view evidence into a confidence-aware 3D proxy, and fuses it with the observed partial input via a refinement network, which utilizes a dual-mode fusion strategy to meticulously align the 3D partial geometry with hierarchical 2D visual cues to generate the final completion. Extensive experiments demonstrate that *ScaleBlind* consistently outperforms state-of-the-art completion methods and modern 3D generative baselines under realistic GT-scale-free settings. We hope our work can inspire further efforts on bridging foundation-model priors with 3D shape modeling and completion.

Limitations and future directions are discussed in Appendix C.

REFERENCES

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Anchorformer: Point cloud completion from discriminative nodes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13581–13590, 2023.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Yun-Chih Guo, Tzu-Hsuan Weng, Robin Fischer, and Li-Chen Fu. 3d semantic segmentation based on spatial-aware convolution and shape completion for augmented reality applications. *Computer Vision and Image Understanding*, 224:103550, 2022.

Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, and Pheng-Ann Heng. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following, 2023. URL https://arxiv.org/abs/2309.00615.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Jiahui Huang, Zan Gojcic, Matan Atzmon, Or Litany, Sanja Fidler, and Francis Williams. Neural kernel surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4369–4379, 2023.

Zitian Huang, Yikuan Yu, Jiawen Xu, Feng Ni, and Xinyi Le. Pf-net: Point fractal network for 3d point cloud completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7662–7670, 2020.

Yoni Kasten, Ohad Rahamim, and Gal Chechik. Point cloud completion with pretrained text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:12171–12191, 2023.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

An Li, Zhe Zhu, and Mingqiang Wei. Genpc: zero-shot point cloud completion via 3d generative priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1308–1318, 2025.

Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10072–10083, 2024.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017a.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

Yi Rong, Haoran Zhou, Lixin Yuan, Cheng Mei, Jiahao Wang, and Tong Lu. Cra-pcn: Point cloud completion with intra-and inter-level cross-resolution transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4676–4685, 2024.

Bipasha Sen, Aditya Agarwal, Gaurav Singh, Srinath Sridhar, Madhava Krishna, et al. Scarp: 3d shape completion in arbitrary poses for improved grasping. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3838–3845. IEEE, 2023.

Junshu Tang, Zhijun Gong, Ran Yi, Yuan Xie, and Lizhuang Ma. Lake-net: Topology-aware point cloud completion by localizing aligned keypoints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1726–1735, 2022.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Jacob Varley, Chad DeChant, Adam Richardson, Joaquín Ruales, and Peter Allen. Shape completion enabled robotic grasping. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 2442–2447. IEEE, 2017.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12, 2019.

Yuqing Wang, Zhaiyu Chen, and Xiaoxiang Zhu. Learning generalizable shape completion with sim (3) equivariance. *Advances in Neural Information Processing Systems*, 38:24300–24329, 2026.

Guangshun Wei, Yuan Feng, Long Ma, Chen Wang, Yuanfeng Zhou, and Changjian Li. Pcdreamer: Point cloud completion through multi-view diffusion priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27243–27253, 2025.

Xin Wen, Peng Xiang, Zhizhong Han, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Yu-Shen Liu. 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, 2022.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, Yao Yao, et al. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. *Advances in Neural Information Processing Systems*, 38:170778–170804, 2026.

Tong Wu, Liang Pan, Junzhe Zhang, Tai Wang, Ziwei Liu, and Dahua Lin. Balanced chamfer distance as a comprehensive metric for point cloud completion. *Advances in Neural Information Processing Systems*, 34:29088–29100, 2021.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 21469–21480, June 2025.

Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han. Snowflakenet: Point cloud completion by snowflake point deconvolution with skip-transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5499–5509, 2021.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *ECCV*, 2024.

Xuejun Yan, Hongyu Yan, Jingjing Wang, Hang Du, Zhihong Wu, Di Xie, Shiliang Pu, and Li Lu. Fbnet: Feedback network for point cloud completion. In *European Conference on Computer Vision*, pp. 676–693. Springer, 2022.

Chuanxiang Yang, Yuanfeng Zhou, Guangshun Wei, Long Ma, Junhui Hou, Yuan Liu, and Wenping Wang. Monge-ampere regularization for learning arbitrary shapes from point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25050–25061, October 2025.

Jinpeng Yu, Binbin Huang, Yuxuan Zhang, Huaxia Li, Xu Tang, and Shenghua Gao. Geoformer: Learning point cloud completion with tri-plane integrated transformer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8952–8961, 2024.

Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12498–12507, 2021.

Xumin Yu, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. Adapointr: Diverse point cloud completion with adaptive geometry-aware transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14114–14130, 2023. doi: 10.1109/TPAMI.2023.3309253.

Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 international conference on 3D vision (3DV)*, pp. 728–737. IEEE, 2018.

Xuancheng Zhang, Yutong Feng, Siqi Li, Changqing Zou, Hai Wan, Xibin Zhao, Yandong Guo, and Yue Gao. View-guided point cloud completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15890–15899, 2021.

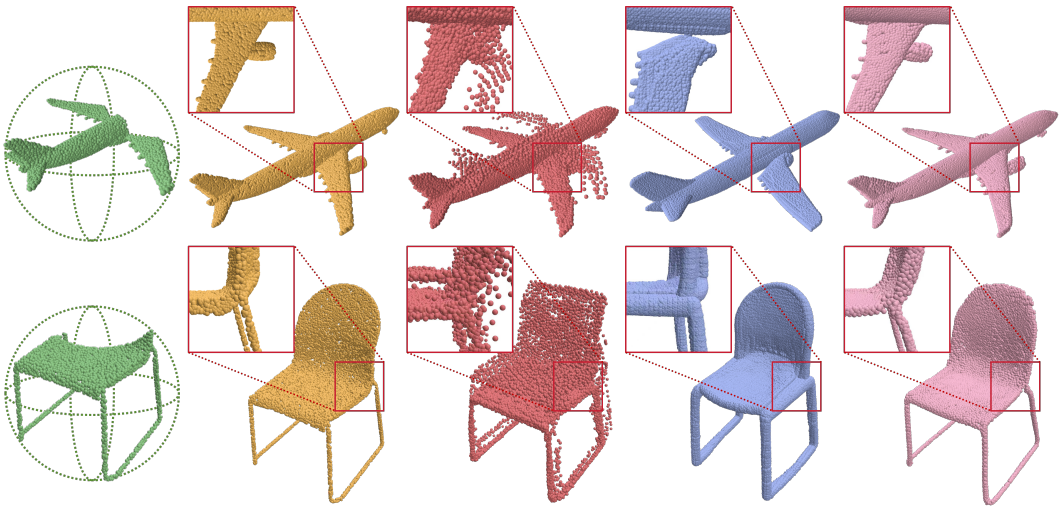
Haoran Zhou, Yun Cao, Wenqing Chu, Junwei Zhu, Tong Lu, Ying Tai, and Chengjie Wang. Seedformer: Patch seeds based point cloud completion with upsample transformer. In *European conference on computer vision*, pp. 416–432. Springer, 2022.

Zhe Zhu, Honghua Chen, Xing He, Weiming Wang, Jing Qin, and Mingqiang Wei. Svdformer: Complementing point cloud via self-view augmentation and self-structure dual-generator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14508–14518, 2023.

A LLM USAGE

In this work, ChatGPT was used solely to improve the grammar, wording, and overall flow of the manuscript after the initial draft had been completed. All AI-assisted revisions were carefully reviewed and verified by the authors. ChatGPT was not used to generate any algorithms, proofs, experimental results, or technical contributions, and no proprietary or sensitive data were shared with the tool. All technical content, analyses, and conclusions are entirely the authors' own work.

B RELATED WORK



(a) Partial point cloud | (b) Ground truth | (c) GeoFormer Yu et al. (2024) | (d) Trellis Xiang et al. (2025) | (e) *ScaleBlind* (Ours)

Figure B1: **Point cloud completion.** Given (a) a partial point cloud, supervised completion methods, such as (c) GeoFormer Yu et al. (2024), often implicitly rely on access to the ground truth shape scale during inference, leading to over-completion artifacts when such information is unavailable. In contrast, (d) Trellis Xiang et al. (2025), a modern 3D generative model, generates structurally plausible complete shapes, yet fails to align with the partial input. We propose (e) *ScaleBlind*, a GT-scale-free framework that faithfully completes the shape while strictly adhering to the partial observation.

Point-based Learning Methods. Deep learning-based point cloud completion Yu et al. (2021); Xiang et al. (2021); Yan et al. (2022); Tang et al. (2022) has emerged as a predominant research direction. PointNet Qi et al. (2017a) established the foundation for directly processing 3D coordinates. Building on this, PCN Yuan et al. (2018) pioneered an end-to-end framework that operates directly on raw point clouds, utilizing a coarse-to-fine strategy to progressively generate complete shapes. Subsequently, Huang et al. (2020) further optimized this framework by refining the multi-stage completion process, significantly enhancing overall performance. Most of these methods assume canonicalized inputs during training and evaluation, and do not explicitly address completion under unknown global scale.

Cross-modal Point-Image Fusion. To compensate for the limited geometric cues in partial point cloud scans, researchers have incorporated auxiliary modalities—such as RGB or depth images—to be processed alongside the geometric data. For instance, ViPC Zhang et al. (2021) fused partial point clouds with single-view RGB images to achieve view-guided completion. SVDFormer Zhu et al. (2023) integrated multi-view depth information with point-based features, utilizing Transformer modules to effectively fuse the spatial relationships of neighboring points with visual semantics. GeoFormer Yu et al. (2024) further enhanced global geometric expression by aligning image features from multi-view Consistent Canonical Maps (CCM) with point cloud features via attention mechanisms. PCDreamer Wei et al. (2025) further explores multi-view generative priors to guide completion. At a high level, our framework shares similarity with this line of research by leveraging multi-view visual cues to complement incomplete geometry. However, unlike prior methods that

mainly treat multi-view information as auxiliary features, we aggregate the recovered evidence into a confidence-aware 3D shape prior to provide explicit geometric guidance for completion. More importantly, this prior also helps recover the missing global extent of the object without GT-scale. Despite these advances, a common limitation remains in synthetic benchmarks: the full shape is often canonicalized before partial observations (and any rendered auxiliary views) are generated, which implicitly ties the input to GT-consistent scale. This assumption is problematic in real scans, where the global extent of the missing region is not available at inference.

Vision-Language Models. The rise of Large Vision-Language Models (VLMs) has provided a powerful paradigm for structural reasoning and image-level completion. Unlike traditional generative models, VLMs—leveraging architectures similar to LLaVA Liu et al. (2023) and other multimodal frameworks—possess extensive pre-trained knowledge that enables zero-shot reasoning of object structures. In 3D tasks, researchers have begun exploring VLMs to interpret partial observations and hypothesize missing components. For instance, Point-Bind Guo et al. (2023) and Point-LLM Xu et al. (2024) demonstrate the potential of aligning 3D geometry with VLM-based semantic spaces.

Overall, most existing completion methods inherently rely on GT-scale-tied canonicalization, while pure image-to-3D generative models fail to preserve the observed partial geometry and are highly prone to scale drift. In contrast, our framework achieves breakthroughs on two parallel fronts. First, as a fundamental completion architecture, we propose a state-of-the-art prior-guided coarse-to-fine network that meticulously aligns multi-modal features to faithfully recover intricate 3D structures. Second, to fundamentally resolve the realistic GT-scale-free challenge under unknown scale, we leverage a VLM to inpaint multi-view projections and lift them into a probabilistic 3D prior via voxel voting. By performing prior-based joint normalization alongside our robust completion network, we yield scale-consistent and geometry-preserving shapes without relying on any privileged oracle information.

C Limitations and Future Work

Limitations. From outer boundary cues, the rendered images successfully guide the VLM to generate accurate global profiles; however, lacking explicit depth or internal topological constraints, these 2D renderings cannot adequately capture depressions, indentations, or local concave surfaces. Note that this performance bottleneck in handling local concavities is a direct consequence of the fundamental ambiguity of representing rich 3D internal geometries through flat 2D renderings. Furthermore, the efficacy of our framework currently relies on the capabilities of *closed-source* VLMs. Through extensive evaluations across various open- and closed-source models (*e.g.*, , FLUXLabs et al. (2025) and GPT-Image2), we observed that existing open-source models still struggle to produce view-consistent images, a property that is crucial for robust 3D lifting. In contrast, state-of-the-art proprietary models like Nano-banana Team et al. (2023), GPT-4o, and GPT-Image2 demonstrate the requisite multi-view consistency and high image quality. We hope that in the near future, new open-source models can serve our purpose.

Our current framework focuses on unknown scale and does not explicitly handle arbitrary input rotations. Since rotation ambiguity is orthogonal to scale estimation, a rotation canonicalization module, such as that used in SCARP Sen et al. (2023), could be introduced before prior acquisition to transform the input into a canonical pose. The completed shape can then be rotated back to the original coordinate system. We leave the joint handling of unknown scale and arbitrary rotation for future work.

Future work. There are a few inspiring directions worth exploring in the future. Our current framework assumes rotational alignment and does not explicitly handle rotational misalignment. While rotation-invariant point cloud encoders exist, their performance remains limited in practice; developing more effective rotation-robust representations is an important direction. The multi-view image priors provided by current foundation models are not always perfect, especially under severe occlusions and ambiguous structures. Future work may explore stronger multi-view consistency constraints and more reliable mechanisms for lifting 2D priors into 3D.

D Dataset Construction and Implementation Details

Benchmark datasets. To train and test the network, we employ two benchmark datasets. The PCN dataset is a subset of the ShapeNet dataset, comprising eight object categories where each ground-truth point cloud consists of 16,384 points. It contains a total of 30,974 samples, which are partitioned into training, validation, and test sets following the exact same split ratio as PCDreamer Wei et al. (2025). We construct the dataset as follows. Given a complete shape, we randomly select a seed point and remove a specific proportion of its nearest neighbors using the K-nearest neighbor (KNN) algorithm. By varying the removal ratios (25%, 50%, and 75% of the total points), we generate partial point clouds with different levels of structural incompleteness. At inference time, the VLM inpaints the missing region roughly. To simulate the variation at inference time, in the data construction, we apply a similar process as prior acquisition shown in Fig. 2(top). In the end, each data sample contains the multi-view images, the merged point cloud, the partial points, and the GT points for training the completion network. Please refer to the *supplementary* for more construction details. In contrast to the PCN dataset, which is restricted to only eight categories, ShapeNet-55 provides a significantly larger and more diverse repository encompassing 55 distinct object classes. To evaluate the scalability and robustness of our proposed model on this large-scale benchmark, we strictly follow the same data construction protocol as aforementioned.

D.1 Dataset Construction

Figure D1 demonstrates a few data processing steps of the construction. Specifically, for each shape from the benchmark dataset, we first create the partial observation (green) and the removed points (yellow) following the strategy as described in Sec. 3 in the main paper. Intuitively, from the point partitions, we could follow the steps as in our prior acquisition module to generate the training samples. However, constructing training data by invoking a VLM to complete multi-view projections for every sample is prohibitively expensive. To make construction practical while staying consistent with our inference pipeline, we adopt a lightweight data simulation procedure that mimics VLM-based multi-view completion and injects realistic view-dependent uncertainty.

The key observation is that VLM-completed projections rarely corrupt the underlying object geometry, based on extensive preliminary experiments. Instead, their imperfections mainly arise from deviations in camera viewpoints, misalignment across views, and mild scale drift. Motivated by this observation, we simulate VLM uncertainty by perturbing camera parameters when projecting the geometry. This design preserves the geometric integrity of the object while producing projection artifacts that are consistent with those observed during inference.

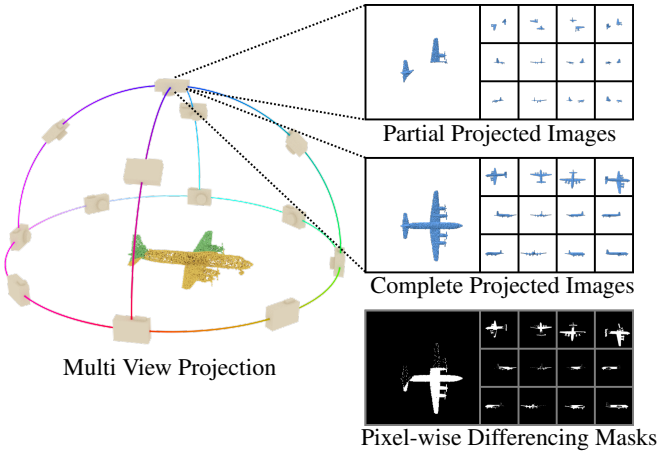


Figure D1: **Dataset construction.** Partial projections are rendered from N_v canonical views of the input point cloud and paired with completed projections (VLM at inference; simulated during training). Hallucination masks are obtained by pixel-wise differencing between the partial and completed projections.

Specifically, as shown in Fig. D1(left), we directly project the partial and the full points from a set of predefined canonical viewpoints, generating the multiview partial and completed projected

images. For each view, random perturbations are applied to both the camera position and the look-at center. For most viewpoints, small Gaussian noise is added to the camera position with a standard deviation of 0.06, and to the look-at center position with a standard deviation of 0.03, simulating mild viewpoint misalignment. For a subset of challenging viewpoints (*e.g.*, the top views), stronger perturbations are applied, with camera position noise of standard deviation 0.15 and target noise of 0.06. In addition, mild scale variations are introduced along the viewing direction for these views, with a random scaling factor sampled from $[0.7, 0.9]$, mimicking the scale drift commonly observed in VLM completions. The two multi-view image sets support the calculation of the pixel-difference masks, from where we apply the remaining steps as in the prior acquisition procedure, including voxel voting, confidence-based shape fusion, and joint normalization (see Figure 2 in the main paper).

After the whole process, for each data sample, we have the paired {*partial points*, *multiview full images*, *ground truth full points*} for training and testing.

D.2 Implementation Details

Training details. Our network is trained for 300 epochs using four NVIDIA RTX 4090 GPUs with a batch size of 24. We optimize the parameters using the Adam optimizer with an initial learning rate of 1×10^{-4}. It is decayed by a factor of 0.7 every 40 epochs on the PCN dataset Yuan et al. (2018) and by a factor of 0.98 every 2 epochs on the ShapeNet-55 dataset Chang et al. (2015). For all experiments, the hyperparameters are fixed at $\tau_{\text{overlap}} = 0.7$, $\tau_{\text{prob}} = 0.8$, $\Delta = 0.02$, $\delta_{\text{diff}} = 30$, $N = 2048$. The number of rendering viewpoints is set to 13, and the render image size is 256×256.

Image inpainting prompt. In the following, we show the inpainting prompt when interacting with Nanobanana.

```
Input: a rendering of a partial point cloud observation.

Task_Prompt:

Task Objective: Based on a visualization of an incomplete chair point
cloud, generate a structurally complete visualization of that same point
cloud.

Viewpoint Specification: the output image must strictly adhere to this
exact viewpoint and camera angle.

General Principles:
-Analyze Valid Information: Comprehensively analyze the geometry, object
contours, and spatial distribution.
-Infer Complete Stucture: Infer the most probable geometry of the missing
 parts.
-Smooth Continuity: Ensure smooth and natural transitions.

Mandatory Constraints:
-Rough Shape Completion: Only complete the approximate overall shape.
-Avoid Excessive Detail: Do not reconstruct fine textures or minor sub-
components.
-Prohibit Invention: Do not invent structures not implied by the data.

The output Format and Content:
-Format Consistency: the output style and format should be identical to
the input.
-Pure Content: Completed object only, white background.

Output:
An inpainted image depicting the full shape.
```

E Additional Experimental Results

E.1 Evaluation on ShapeNet-55

Table E1: **Quantitative comparison on the ShapeNet-55 dataset under the *GT-scale-free* and *GT-scale-tied* settings**. (ℓ^2 CD $\times 10^3$ and F-Score@1%). Shapes from ten categories are used in the evaluation. We report the CD-Avg↓ for each category, and additionally report the CD-Avg↓, DCD-Avg↓, F1↑ across all categories. The compared methods are divided into two groups according to whether their original frameworks support text prompts as input: methods in the first group do not support text-prompt input, whereas methods in the second group can receive category-related text prompts, as does our method. For each cell, the two numbers are for the GT-scale-free and GT-scale-tied settings, respectively. We use the dark green and dark blue to indicate the best results for evaluations under GT-scale-free and GT-scale-tied settings, while using the light green and light blue to indicate the second best.

Methods	Table	Chair	Plane	Car	Sofa	Bird house	Bag	Microwave	Mailbox	Dishwasher	CD-Avg↓	DCD-Avg↓	F1↑
AdaPointr*	1.28 / 0.84	0.88 / 0.81	0.55 / 0.47	0.96 / 0.93	1.17 / 0.86	1.62 / 1.34	1.13 / 0.74	2.08 / 1.41	0.85 / 0.71	1.51 / 1.15	1.40 / 1.08	0.658 / 0.619	0.298 / 0.408
SVDFormer*	1.38 / 0.78	1.05 / 0.77	0.64 / 0.45	1.05 / 0.75	1.27 / 0.86	1.87 / 1.19	1.51 / 0.72	2.21 / 1.19	1.14 / 0.59	1.77 / 1.04	1.45 / 0.93	0.661 / 0.568	0.288 / 0.437
CRA-PCN*	1.54 / 0.83	1.07 / 0.83	0.62 / 0.58	1.09 / 0.96	1.24 / 0.94	1.78 / 1.38	1.21 / 0.78	2.29 / 1.38	1.18 / 0.63	1.84 / 1.59	1.37 / 1.10	0.657 / 0.623	0.295 / 0.405
GeoFormer*	1.42 / 0.75	0.99 / 0.76	0.69 / 0.42	1.01 / 0.84	1.18 / 0.83	1.54 / 1.12	1.12 / 0.68	1.97 / 1.14	1.18 / 0.63	1.45 / 0.98	1.31 / 0.88	0.624 / 0.556	0.335 / 0.454
PCDreamer*	1.12 / 0.72	0.98 / 0.74	0.59 / 0.41	0.99 / 0.79	1.13 / 0.85	1.46 / 1.26	1.03 / 0.65	2.17 / 1.09	0.91 / 0.58	1.54 / 0.96	1.29 / 0.86	0.628 / 0.551	0.322 / 0.442
Ours	0.89 / 0.69	0.81 / 0.74	0.51 / 0.40	0.83 / 0.73	0.98 / 0.78	1.17 / 1.05	0.90 / 0.58	1.41 / 1.02	0.64 / 0.53	1.22 / 0.92	1.18 / 0.84	0.615 / 0.549	0.347 / 0.458

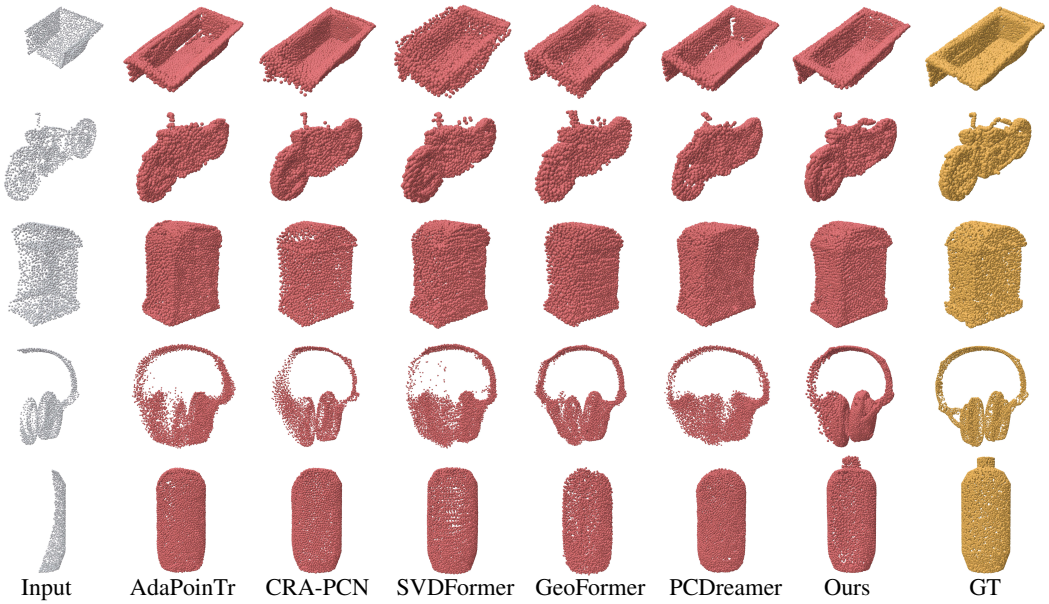


Figure E1: **Qualitative comparison on the ShapeNet-55 dataset under the GT-scale-free setting**. Given the inputs with varying missing regions, baseline methods produce inferior completion results with clear artifacts, while our method completes objects faithfully.

Other than the PCN dataset, we have evaluated the performance of all the methods on the ShapeNet-55 dataset. Specifically, Tab E1 reports results under both the GT-scale-free and GT-scale-tied settings. Compared to PCN, ShapeNet-55 spans more categories and exhibits larger intra-class shape variation, which makes GT-scale-free completion more challenging. Despite this broader diversity, *ScaleBlind* still achieves the strongest overall results in both protocols and remains stable when moving from GT-scale-tied to GT-scale-free preprocessing.

Visual results are shown in Fig. E1. Our method produces superior completions.

E.2 Comparison: Generative Point Cloud Completion

We further compare *ScaleBlind* with recent image-to-3D generators, including Hi3DGen Ye et al. (2025) and Trellis Xiang et al. (2025). Both models are pretrained to synthesize complete 3D assets from a single RGB input, thus providing a strong alternative solution for completion. For a fair

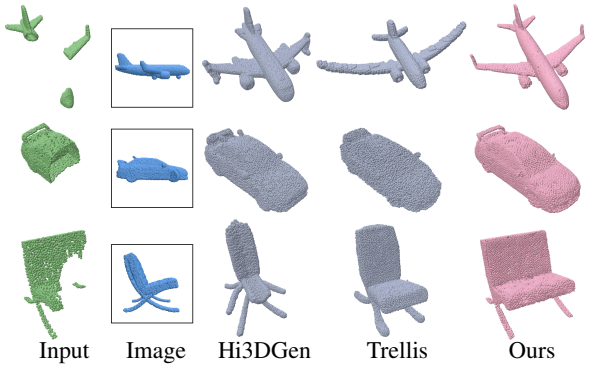


Figure E2: **Generative completion.** Given the partial input, we use VLMs to complete partial renderings and send them to two representative image-to-3D generation models to reconstruct the full shape. Although their results are structurally meaningful and visually plausible, they fail to align with the partial observation. Our results are of high fidelity and respect the inputs.

comparison, we feed the 13 VLM-completed images produced by our pipeline into these models to generate meshes, choose the best result from the 13 candidates, and uniformly sample points from the output via Poisson disk sampling for evaluation. Note that we run several rounds of each generation, since they are sometimes unstable to produce reasonable results.

Visual comparison is shown in Fig. E2, where these generative baselines often yield plausible global shapes but misaligned geometry compared with the partial observation, such as the wings of the airplane or thin hollow structures on car spoilers, due to the limited information in a single view. In the future, a more powerful generative model taking as input multi-view images could potentially solve this issue. In contrast, *ScaleBlind* leverages consistent evidence across multiple views and recovers more faithful shapes with higher fidelity.

E.3 Impact of the Number of Viewpoints

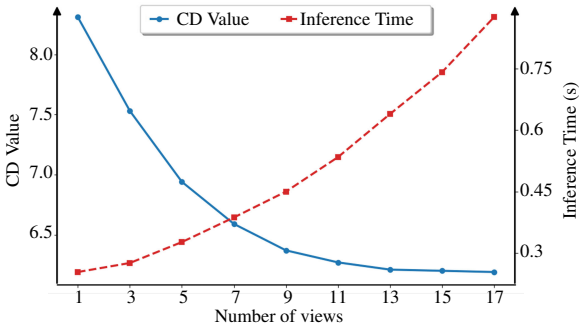


Figure E3: **Viewpoint numbers vs. CD and inference time.**

Image-based guidance is critical for high-fidelity point cloud completion, and the number of views is a key hyperparameter affecting the final results. To investigate this, we generate 1–18 candidate viewpoints for each object and systematically evaluate model performance across different view counts. As shown in Fig. E3, the horizontal axis denotes the number of views, while the left and right vertical axes report the CD value and inference time, respectively. The error decreases significantly as more views are added and then gradually saturates, while the VLM inference cost continues to increase.

Table F1: **Quantitative comparison using single-view point clouds rendered from depth maps.** (ℓ^1 CD $\times 10^3$ and F-Score@1%). Following the same data generation protocol as PCDreamer Wei et al. (2025), the partial inputs are acquired from single-view depth renderings. We report the CD-Avg↓ for each category, and additionally report the CD-Avg↓, DCD-Avg↓, and F1↑ across all categories. The compared methods are divided into two groups according to whether their original frameworks support text prompts as input: methods in the first group do not support text-prompt input, whereas methods in the second group can receive category-related text prompts, as does our method. For each cell, the two numbers are for the GT-scale-free and GT-scale-tied settings, respectively. We use dark green and dark blue to indicate the best results for evaluations under GT-scale-free and GT-scale-tied settings, while using light green and light blue to indicate the second best.

Methods	Plane	Cabinet	Car	Chair	Lamp	Couch	Table	Boat	CD-Avg↓	DCD-Avg↓	F1↑
AdaPoinTr*	4.86 / 3.53	9.34 / 8.99	7.37 / 6.92	7.84 / 6.57	9.50 / 5.57	9.68 / 8.34	7.79 / 5.96	7.12 / 5.93	7.94 / 6.48	0.595 / 0.530	0.693 / 0.852
SVDFormer*	4.63 / 3.68	9.27 / 8.73	7.52 / 7.10	8.64 / 6.95	9.63 / 5.64	9.94 / 8.58	7.68 / 6.26	7.83 / 5.91	8.14 / 6.61	0.673 / 0.534	0.674 / 0.848
CRA-PCN*	4.82 / 3.62	9.15 / 8.77	7.30 / 7.00	8.74 / 6.92	9.83 / 5.46	10.53 / 8.59	8.05 / 6.27	8.23 / 5.86	8.33 / 6.56	0.631 / 0.537	0.672 / 0.846
GeoFormer*	5.16 / 3.70	10.03 / 9.08	7.24 / 6.93	9.92 / 7.19	10.30 / 6.30	10.67 / 8.97	8.61 / 7.14	8.13 / 5.87	8.75 / 6.90	0.697 / 0.563	0.659 / 0.821
SIMECO	4.80 / - - -	12.11 / - - -	7.13 / - - -	8.88 / - - -	7.62 / - - -	11.32 / - - -	8.07 / - - -	6.50 / - - -	8.30 / - - -	- - - / - - -	0.674 / - - -
PCDreamer*	4.26 / 3.52	9.25 / 8.72	7.21 / 6.89	7.96 / 6.71	9.81 / 5.64	9.69 / 8.32	7.94 / 6.24	7.62 / 5.84	7.97 / 6.49	0.614 / 0.518	0.681 / 0.859
Ours	4.10 / 3.46	8.43 / 8.31	7.16 / 6.90	6.97 / 6.69	8.21 / 5.85	8.87 / 7.93	7.27 / 6.21	6.48 / 5.93	7.19 / 6.41	0.538 / 0.513	0.722 / 0.855

F Additional Discussion and Analysis

F.1 Evaluation on Single-View Scanned Partial Points

In addition to generating partial point clouds via point cropping, we further evaluate our method using single-view point clouds from the virtual scanning processing, strictly following the data generation protocol of PCDreamer Wei et al. (2025). To comprehensively validate the effectiveness of our approach, this evaluation is conducted under both the traditional *GT-scale-tied* setting and the more challenging *GT-scale-free* setting. As shown in Tab. F1, under the GT-scale-tied assumption, our method consistently outperforms other state-of-the-art completion networks. This demonstrates that our core architecture possesses robust feature extraction and geometric reasoning capabilities, even when confronted with highly self-occluded single-view observations. Furthermore, when extended to the GT-scale-free setting, our approach maintains its overall superiority. Although it yields slightly sub-optimal results in a small minority of specific object categories, our method secures the best overall performance across all average metrics (i.e., CD-Avg, DCD-Avg, and F1). This confirms that our framework can effectively leverage visual priors to overcome absolute scale ambiguity, achieving robust and comprehensive completions across diverse shapes.

F.2 Robustness Analysis

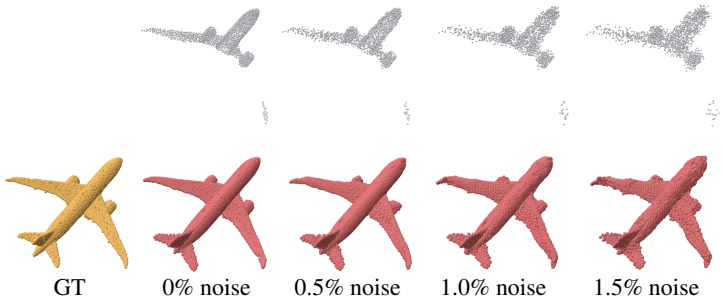


Figure F1: **Completion under varying noise levels.** Given an input with increasing noise levels, our method completes shapes plausibly with a consistent global structure and geometry. Minor artifacts can be observed only at the highest noise level.

Robustness to noise. Fig. F1 visualizes *ScaleBlind* under increasing input noise. To simulate sensor noise, we add zero-mean Gaussian perturbations to the coordinates of partial point clouds, with standard deviations of 0.5%, 1.0%, and 1.5% relative to the object scale. Across all settings, *ScaleBlind* preserves the overall structure, and even at 1.5% noise it reconstructs the aircraft with correct global extent and coherent geometry. This robustness stems from the multi-view prior: the

VLM-completed views provide consistent global evidence that stabilizes prior acquisition and guides the coarse-to-fine network beyond local coordinate jitter.

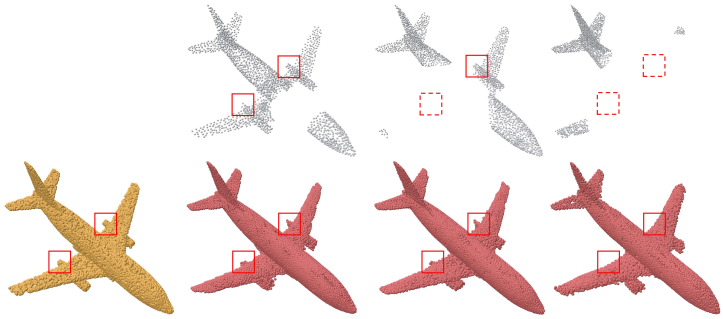


Figure F2: **Completion under varying missing ratios.** Given an input partial shape with varying missing ratios, our method robustly reconstructs the shape across different levels of incompleteness.

Robustness to missing ratios. Fig. F2 visualizes *ScaleBlind* under varying levels of structural incompleteness. To evaluate the performance across different degrees of data loss, we downsample the partial point clouds by 25%, 50%, and 75%. Across all settings, *ScaleBlind* demonstrates consistent stability, reconstructing complete geometry with accurate proportions and coherent global extent. As highlighted by the red boxes, the wing engines remain visible at 25% and 50% missing ratios, and *ScaleBlind* reconstructs them consistently with the ground truth. At 75% missing, the wing engine evidence is largely absent, so the prediction can deviate from the GT while still remaining geometrically plausible. This indicates that *ScaleBlind* is strong across missing ratios, and under severe incompleteness it increasingly relies on the scale-consistent multi-view prior inferred from VLM-completed renderings.

F.3 LiDAR Point Completion

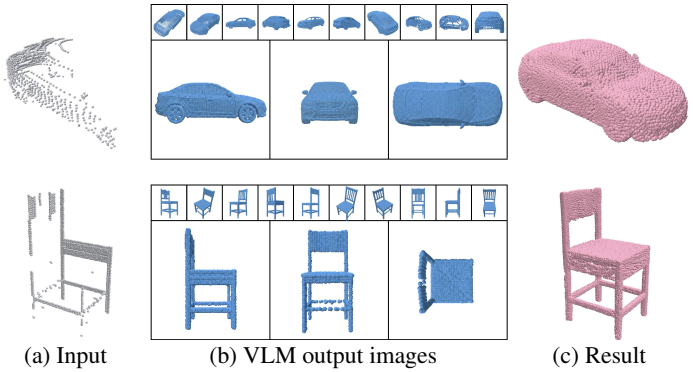


Figure F3: **LiDAR point completion**. Given the partial LiDAR point cloud with a large proportion of missing regions, our method can produce structurally valid and geometrically plausible completion.

Figure F3 presents qualitative results on challenging real-world datasets—indoor scenes from ScanNet Dai et al. (2017) and outdoor driving scenes from KITTI Geiger et al. (2013). These sensor observations are typically sparse and noisy (Fig. F3(a)). However, the pretrained VLMs can still leverage their learned shape and semantic knowledge to generate high-quality multi-view 2D prior images (Fig. F3(b)). The final results (Fig. F3(c)) indicate that our network effectively utilizes these 2D priors as guidance to recover complete, geometrically consistent 3D shapes without further finetuning on specific datasets. This experiment demonstrates the generalization capability and potential for practical applications of our approach.

Table F2: **Comparison with generative completion methods without GT-scale on PCN.** We report ℓ^1 CD ($\times 10^3$), DCD, and F-Score@1%. GenPC is evaluated on the complete PCN test set. Due to its substantial computational cost, SDS-Complete is evaluated on two samples from each of the eight categories. For each baseline, our method is evaluated on the same corresponding test samples. Results on the 16-sample SDS-Complete subset are reported separately from the full-test-set results.

Methods	Plane	Cabinet	Car	Chair	Lamp	Couch	Table	Boat	CD-Avg↓	DCD-Avg↓	F1↑
GenPC (full test set)	37.97	97.18	47.48	69.28	80.61	110.20	111.65	45.93	75.04	0.917	0.122
Ours (full test set)	**4.10**	**8.43**	**7.16**	**6.97**	**8.21**	**8.87**	**7.27**	**6.48**	**7.19**	**0.538**	**0.722**
SDS-Complete (2 samples/category)	103.10	152.77	81.24	68.36	55.43	152.31	62.22	188.11	107.94	0.913	0.087
Ours (same samples)	**2.94**	**7.17**	**8.31**	**5.77**	**6.55**	**9.44**	**5.21**	**3.74**	**6.14**	**0.525**	**0.794**

F.4 Comparison with Generative Completion Methods without GT-Scale

Table F2 compares *ScaleBlind* with the generative completion methods GenPC Li et al. (2025) and SDS-Complete Kasten et al. (2023). Rather than learning shape priors solely from the completion dataset, these methods leverage powerful generative models pretrained on large-scale data: GenPC builds upon TRELLIS, while SDS-Complete utilizes a pretrained Stable Diffusion model. Both GenPC and SDS-Complete operate without access to GT-scale information, making them suitable for comparison with *ScaleBlind*. Both methods are designed around view-dependent partial observations and rely on a meaningful acquisition viewpoint for rendering the input. Their pipelines are therefore not directly matched to arbitrary cropped partial point clouds that do not correspond to a specific observation viewpoint. In contrast, *ScaleBlind* supports both single-view and cropped partial inputs. To accommodate the input assumptions of these methods and ensure a fair comparison, we conduct this evaluation under the PCN single-view setting.

GenPC is evaluated on the complete PCN test set. As shown in Table F2, *ScaleBlind* substantially outperforms GenPC across all eight categories, reducing the overall CD-Avg from 75.04 to 7.19 and DCD-Avg from 0.917 to 0.538, while improving F1 from 0.122 to 0.722. In our experiments, SDS-Complete requires approximately 40 hours to process a single sample, making evaluation on the complete test set computationally impractical. We therefore evaluate SDS-Complete on two random samples from each of the eight PCN categories, resulting in 16 samples in total, and report the category-wise and overall averages. *ScaleBlind* is evaluated on exactly the same subset. On this subset, *ScaleBlind* reduces CD-Avg from 107.94 to 6.14 and DCD-Avg from 0.913 to 0.525, while increasing F1 from 0.087 to 0.794. These results demonstrate that, even under the PCN single-view setting tailored to the view-dependent pipelines of these generative baselines, *ScaleBlind* achieves substantially higher metric completion accuracy while additionally supporting more general cropped partial inputs.

F.5 Plausibility versus overfitting

Furthermore, the exact unobserved geometry cannot always be deterministically recovered from a highly incomplete partial scan. Rather than simply memorizing the ground-truth shapes—a common symptom of overfitting—our network learns to hallucinate geometrically plausible and structurally sound completions. As illustrated in Fig. F4, our results may occasionally exhibit slight structural variations from the exact GT in severely occluded regions. For instance, the generated edges of the chair (first row) and the joint structure between the backrest and seat cushion (second row) differ subtly from the original GT. Similarly, natural variations can be observed in the generated car windows (third row) and bumpers (fourth row). These discrepancies are highly reasonable and expected; they provide evidence that our model can generalize to synthesize valid geometric alternatives based on robust priors, rather than rigidly overfitting to the specific local details of the training data.

F.6 On-the-fly Partial-based Normalization

This setting is practical, as it normalizes the input partial point cloud solely based on the observation itself, without accessing any scale information from the full shape. To evaluate existing completion methods under this realistic protocol, we normalize the partial input on-the-fly and directly feed it into their networks. However, most prior methods implicitly assume GT scale during both training and inference through GT-scale-tied canonicalization, resulting in a significant distribution shift once

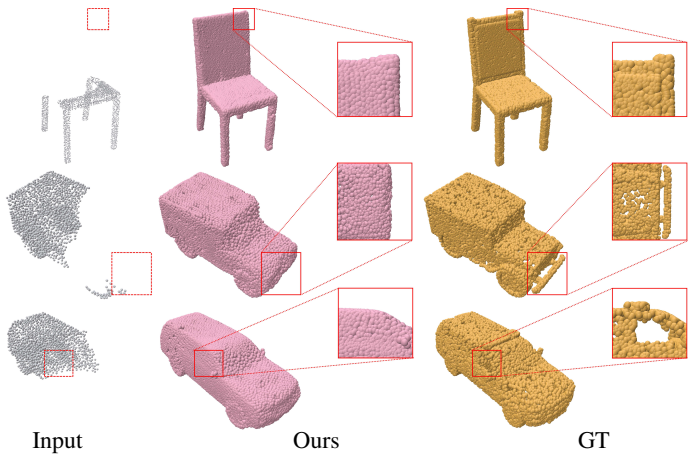


Figure F4: **Qualitative results showing plausible variations from the ground truth.** When dealing with unknown structures, our method synthesizes geometrically reasonable alternatives rather than producing results completely identical to the exact GT.

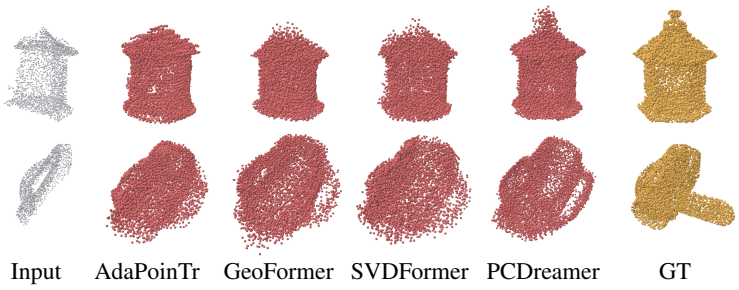


Figure F5: **Qualitative results of SoTA methods under the partial-based normalization setting.** When baseline methods are forced to normalize inputs using the observed partial bounding box instead of the ground truth, they suffer from severe geometric distortion and scale collapse.

this oracle cue is removed. Consequently, naively applying their pretrained models under partial-only normalization yields unreasonable completions, as shown in Fig. F5.

As expected, the performance degrades substantially. Without access to full-shape scale (either from GT or reliable estimation), completion becomes considerably harder, as the model must jointly infer the global placement/scale of the partial observation and hallucinate the missing geometry. These results further validate the motivation of our method, which explicitly recovers the missing global extent and scale under the GT-scale-free setting.

F.7 Failure Case

Figure F6 visually illustrates the decisive influence of the 2D images hallucinated by the VLM on our final 3D completion results. We analyze this dependency by examining the recovery of the specific connection structure between the chair's backrest and seat cushion under different input conditions. In Fig. F6(a), geometric cues of the connecting structure remain partially visible in the input point cloud. Benefiting from these hints, the VLM successfully generates accurate and plausible 2D images, which in turn guide the model to produce high-quality 3D completion results that closely match the ground truth. Conversely, in Fig. F6(b), this connecting structure is entirely missing from the partial input. Lacking local geometric constraints, the VLM hallucinates an alternative, incorrect structural arrangement. Consequently, the final 3D reconstruction faithfully follows this erroneous guidance from the VLM prior, resulting in a less ideal geometric shape that deviates from the original object. In Fig. F6(c), although the connection is also completely absent in the input, the VLM utilizes its learned semantic priors to hallucinate a plausible (even if not identical to ground truth) connecting structure. As a result, our model reconstructs a 3D shape that is geometrically consistent with the

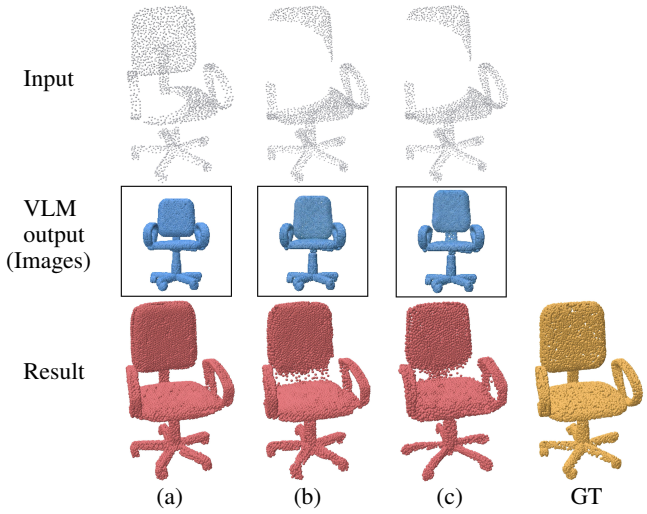


Figure F6: **VLM inpainting and 3D shape completion.** Columns (a)–(c) show completion results conditioned on different VLM-inpainted images. Due to varying visibility of the joint between the backrest and the seat in the partial input point cloud, VLMs may hallucinate different geometric details, leading to distinct 3D completions.

VLM-provided image. Note that the noisy points in (b) and (c) may stem from inconsistencies in the VLM-inpainted images in the connecting region.