BEAMSCENE: A PHYSICALLY CONSISTENT FRAMEWORK WITH BEAM-FOOTPRINT GEOMETRY FOR 3D LIDAR SCENE COMPLETION

Anonymous authors
Paper under double-blind review

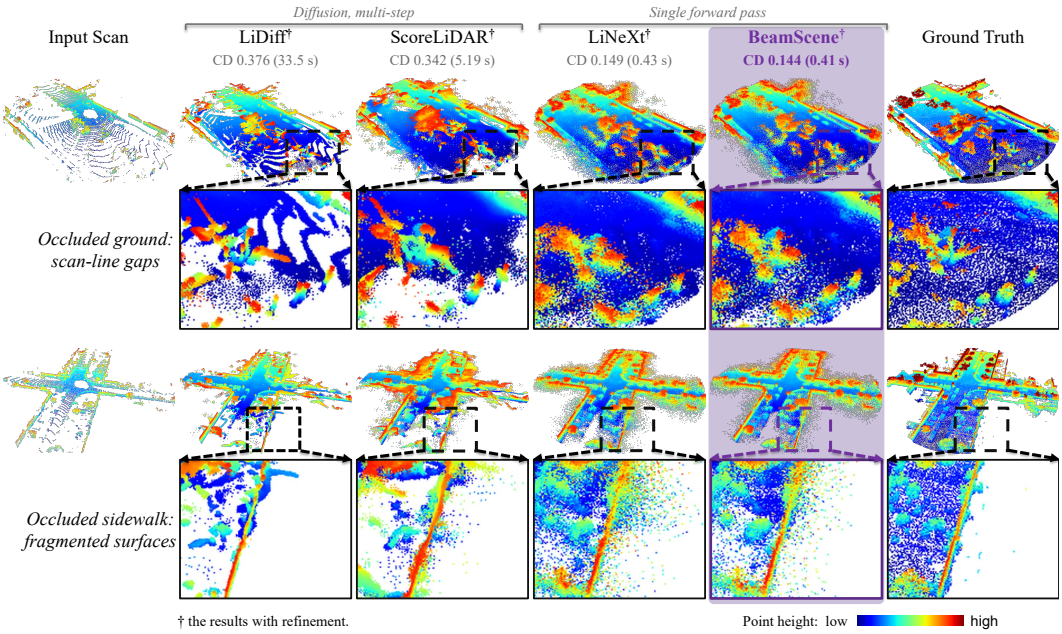

Figure 1: BeamScene completes a sparse LiDAR scan in a single forward pass. Diffusion-based completion (LiDiff, ScoreLiDAR) leaves scan-line gaps and fragmented surfaces in occluded regions (dashed boxes, enlarged below). BeamScene derives candidate placement, feature scale, and visibility supervision from the beam footprint of the sensor, and attains the lowest Chamfer distance (CD) at the lowest runtime. CD and runtime are taken from Table 3 (SemanticKITTI, one RTX 6000 Ada GPU); † denotes the configuration with refinement.

ABSTRACT

Autonomous vehicles depend on LiDAR for metrically accurate 3D perception, yet a single scan is sparse, and occluded regions are left almost empty. LiDAR scene completion recovers the missing geometry, but diffusion-based completion requires many network evaluations per scan, and recent single-pass methods place candidate points, choose feature scales, and supervise visibility with hand-set rules that are unrelated to how the sensor samples the scene. We present **BeamScene**, a single-pass framework that derives each of these decisions from one quantity, the beam footprint of every return, obtained from its range and the angular resolution of the sensor. The Beam-Footprint-Aware Candidate Constructor (BFCC) allocates an exact, bounded candidate budget in proportion to the footprint, spreads candidates in the ray-orthogonal tangent plane, and keeps every measured return as an immutable anchor. The Footprint-Aligned Shared-Scale Encoder (FASSE) shares sparse kernels across eight voxel resolutions and both input branches, and weights the resolutions with a parameter-free footprint-to-grid gate. The Frustum-Ray Visibility Objective (FRVO) adds frustum-local entropic transport and differentiable first-return and free-space terms during training only. On SemanticKITTI, BeamScene raises voxel IoU at 0.2 m and 0.1 m from 19.45 and 6.30 (LiNeXt) to 24.91 and 8.74, in a regime where Chamfer distance has largely saturated; its refined variant is the best of all compared methods on all six

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.

metrics of both SemanticKITTI and KITTI-360. The coarse model has 1.12M parameters, 43.7% fewer than LiNeXt, and runs in 0.234 s per scan, 17–99× faster than diffusion-based completion.

1 INTRODUCTION

Accurate and efficient perception of the surrounding environment is a prerequisite for the safe operation of autonomous vehicles (Gao et al., 2026; Xiong et al., 2026; Khatab et al., 2026). Among onboard sensors, 3D LiDAR is widely adopted because it measures geometry directly and at long range (Krishnakumar et al., 2026; Elhashash et al., 2022), and the same measurements underpin high-definition mapping (Ilci & Toth, 2020) and city-scale digital twins (Raynold & Rao, 2026; Larsen et al., 2026). A single scan, however, samples the scene sparsely: the gap between neighbouring beams widens with range, foreground objects occlude the structures behind them, and a single viewpoint leaves large regions unobserved (Nunes et al., 2024). LiDAR scene completion infers this missing geometry and turns a sparse scan into a dense scene representation for downstream perception.

Diffusion models have become the dominant approach to this task because they train stably and generate high-quality scenes (Nunes et al., 2024; Martyniuk et al., 2025; Zhao et al., 2026). LiDiff applies denoising directly to scene-scale points (Nunes et al., 2024), LiDPM shows that a global formulation with a suitable starting point is effective (Martyniuk et al., 2025), and ScoreLiDAR distills a diffusion teacher into a few-step student (Zhang et al., 2025). All of them still evaluate the network several times per scan, which costs 4 to 34 s on a modern GPU (Fig. 2) and is difficult to reconcile with the latency budget of an autonomous vehicle. Single-pass designs remove this cost (He et al., 2026; Hussian et al., 2026). LiNeXt replicates each return according to manually defined range groups, perturbs the replicas with fixed-variance Gaussian noise, and refines the result in one forward pass (He et al., 2026).

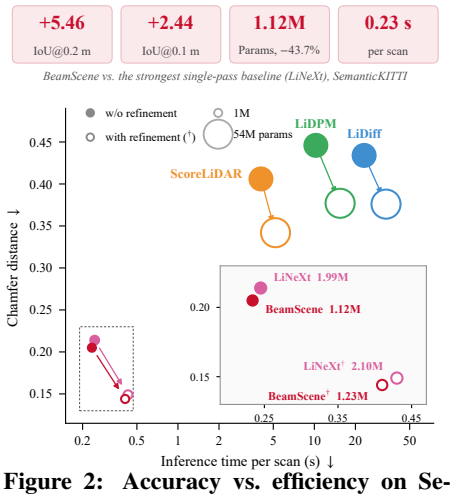

Figure 2: Accuracy vs. efficiency on SemanticKITTI. Marker area ∝ parameters; open markers denote refinement (†). BeamScene is 17–99× faster than diffusion-based completion and uses 43.7% fewer parameters than LiNeXt.

Once the sampling trajectory is removed, the quality of a single-pass method is decided by four coupled design choices: where candidate points are placed, at which scale features are extracted, which observations are preserved, and how the generated geometry is supervised. In existing pipelines each choice rests on its own assumption. Candidate counts follow hand-tuned range groups rather than the range-dependent area that a beam actually covers; voxel sizes are architectural constants shared by a nearby, densely sampled surface and a distant, sparsely sampled one; displacement-based decoders move measured and hypothesised points alike, so the observation is not retained exactly; and Chamfer supervision is invariant to point order, so a prediction can reach a small set distance while lying in front of the surface that the ray actually hit (Agro et al., 2024).

LiDAR acquisition geometry supplies a common basis for all four choices. Combined with the angular resolution of the sensor, a measured return defines an approximate footprint at its range, constitutes a surface observation that should not be displaced, and certifies the space in front of it as empty. We call a completion pipeline *physically consistent* when candidate support, feature scale, retained measurements, and visibility supervision are all derived from this single acquisition model. The term refers to how the pipeline is constructed; some of its consequences hold by construction (Appendix B), while the others are encouraged during training and are assessed through the standard completion metrics.

Following this principle, we propose **BeamScene** (Fig. 1). The **B**eam-**F**ootprint-Aware **C**andidate **C**onstructor (**BFCC**) estimates the footprint of each return, allocates an exact global candidate budget under per-return bounds, and distributes candidates with a golden-angle pattern in the ray-orthogonal tangent plane while keeping each measured point as an immutable anchor. The **F**ootprint-**A**ligned

Shared-**S**cale **E**ncoder (**FASSE**) shares one pair of sparse kernels across eight voxel resolutions and both input branches and weights each resolution by its agreement with the footprint. An observation-preserving coarse-to-fine decoder keeps every anchor exactly in both outputs. Finally, the **F**rustum-**R**ay **V**isibility **O**bjective (**FRVO**) combines frustum-local entropic transport (Cuturi, 2013) with first-return agreement and a one-sided free-space penalty. BFCC, the footprint gate, and FRVO contain no trainable parameters, and FRVO is used only during training. Our contributions are:

- We identify cross-stage inconsistency, i.e., candidate support, feature scale, observation retention, and visibility each being governed by an unrelated assumption, as a limitation of single-pass LiDAR completion, and propose **BeamScene**, which derives all four from the beam footprint of the sensor.
- We design **BFCC** and **FASSE**, two parameter-free uses of the same footprint: BFCC builds an exact-size candidate set in the tangent plane with immutable anchors, and FASSE routes shared multi-scale sparse features by footprint-to-grid agreement.
- We design **FRVO**, a training-only objective that adds frustum-local transport, first-return agreement, and a one-sided free-space penalty to order-invariant set supervision, at no inference cost.
- On SemanticKITTI and KITTI-360 (without fine-tuning), BeamScene† is the best of all compared methods on all six metrics of both benchmarks, and the coarse BeamScene improves on the strongest single-pass baseline on every metric while using 43.7% fewer parameters. Component and internal ablations quantify the contribution of each design.

2 Related Work

Representations for LiDAR scene completion. Outdoor scene completion Xu et al. (2026) has been formulated over voxel grids, implicit fields, and point sets. LMSCNet couples a lightweight 2D backbone with 3D completion heads (Roldão et al., 2020); Make it Dense completes a single-scan truncated signed distance field under self-supervision (Vizzo et al., 2022); and LODE conditions an Eikonal implicit field on local shape priors (Li et al., 2023). Their fidelity and memory are tied to the underlying grid, and none returns a point set that retains the measured coordinates. Point-based generative models keep the native representation of the scan. PVD combines point and voxel representations for object-level generation (Zhou et al., 2021); LiDiff reformulates diffusion at scene scale (Nunes et al., 2024); LiDPM removes the local-diffusion approximation (Martyniuk et al., 2025); and ScoreLiDAR distills LiDiff into a few-step student with a structural loss (Zhang et al., 2025). Single-pass designs remove the sampling trajectory altogether (He et al., 2026; Hussian et al., 2026), which shifts the decisive design choice from the denoising schedule to the quality of the initial point set; BeamScene targets exactly this choice.

Coarse-to-fine and multi-scale point processing. PCN decodes a global code into a coarse set and then local patches (Yuan et al., 2018), SnowflakeNet progressively splits parent points (Xiang et al., 2021), and PoinTr casts completion as set-to-set translation (Yu et al., 2021). Scene-scale inputs require sparse operators: submanifold convolution restricts computation to active sites (Graham et al., 2018), and generalized sparse convolution extends it to high-dimensional tensors (Choy et al., 2019). LiNeXt aggregates parallel sparse blocks at a fixed set of voxel sizes, each with its own kernels (He et al., 2026); FASSE instead shares the kernels and selects scales per point from the physical footprint.

Geometric supervision and LiDAR visibility. Completion is usually supervised with permutation-invariant set distances: Chamfer distance (Yu et al., 2021), its hyperbolic variant that down-weights outliers (Lin et al., 2023), or entropic optimal transport solved with Sinkhorn iterations (Cuturi, 2013). None of these encodes depth order along a ray. Ray-based occupancy models such as UnO make this order explicit by rendering LiDAR depth from a learned occupancy field (Agro et al., 2024). FRVO brings the same first-return evidence into point-set supervision without changing the network.

3 Method

3.1 Problem Setup and Overview

Let $\mathscr{P} = \{\mathbf{p}_i\}_{i=1}^{N} \subset \mathbb{R}^3$ be a sparse scan in the sensor frame and $\mathscr{Y}$ the dense reference scene. BeamScene predicts a coarse set $\mathscr{C}$ of M points and a refined set $\mathscr{R}$ of UM points; we use $N = 18{,}000$,

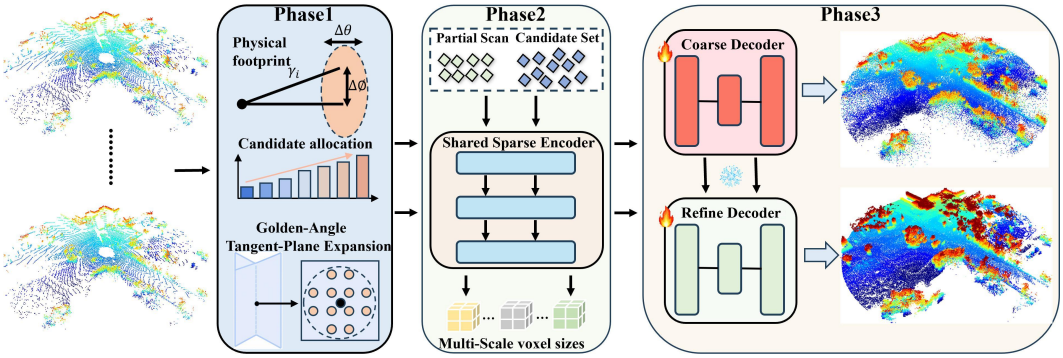

Figure 3: Overview of BeamScene. *Phase 1*: BFCC derives the footprint of each return from its range and the angular resolutions, allocates an exact, bounded candidate budget in proportion to footprint area, and places non-anchor candidates with a golden-angle pattern in the ray-orthogonal tangent plane; each measured return is kept as an anchor. *Phase 2*: the partial scan and the candidate set are encoded by FASSE, which applies the same sparse kernels at eight voxel resolutions and to both branches, and weights the resolutions by footprint-to-grid agreement. *Phase 3*: the observation-preserving coarse decoder reconstructs the scene and is then frozen while the refinement decoder predicts bounded child offsets; anchors are retained exactly in both outputs. FRVO supervises both stages during training and is not used at inference.

$M = 180{,}000$, and $U = 6$. As shown in Fig. 3, BFCC (Sec. 3.2) expands $\mathscr{P}$ into a candidate set of exactly M points; FASSE (Sec. 3.3) encodes the scan and the candidates; the decoder (Sec. 3.4) reconstructs the scene without displacing measurements; and FRVO (Sec. 3.5) adds acquisition-order supervision during training. $\Delta\theta$ and $\Delta\phi$ denote the horizontal and vertical angular resolutions of the sensor, and ε_0 is a small numerical constant.

3.2 Beam-Footprint-Aware Candidate Construction

A single-pass network must first expand the sparse scan into a candidate set whose size matches the output. Replicating each return by range group and perturbing it isotropically has two consequences: the number of candidates per return does not follow the physical support of the return, which grows quadratically with range, and the perturbation moves candidates along the ray, where the measurement is already reliable, instead of across the surface that the beam did not resolve.

Footprint and budget. For a return $\mathbf{p}_i$ at range $r_i = \|\mathbf{p}_i\|_2$, one beam covers an approximate area A_i with equivalent length scale s_i:

$$A_i = r_i^2\,\Delta\theta\,\Delta\phi, \qquad s_i = \max\big(s_{\min},\, r_i\sqrt{\Delta\theta\,\Delta\phi}\big), \tag{1}$$

where $s_{\min}$ prevents degeneration at short range. The scale s_i links the stages of BeamScene: it sets the expansion radius here and the routing weights in FASSE. Because A_i grows quadratically, we clip the areas to the empirical $[q_\ell, q_u]$ quantile range and normalise them, $w_i = \bar{A}_i^{\gamma} / \sum_j \bar{A}_j^{\gamma}$, with $q_\ell = 0.05$, $q_u = 0.95$, and $\gamma = 1$. Every return first receives one anchor candidate, and the remaining $M - N$ candidates are distributed in proportion to w_i by a bounded largest-remainder rule, which yields integer counts $1 \le n_i \le n_{\max}$ with $\sum_i n_i = M$ exactly ($n_{\max} = 24$); no padding or truncation is needed.

Tangent-plane expansion. Candidates are placed in the plane orthogonal to the ray, the direction that the beam leaves unresolved. With $\mathbf{u}_i = \mathbf{p}_i/(r_i + \varepsilon_0)$ and an orthonormal basis $(\mathbf{e}_{i,1}, \mathbf{e}_{i,2})$ of the plane normal to $\mathbf{u}_i$, the j-th child ($j = 0, \ldots, n_i - 1$) is

$$\mathbf{q}_{ij} = \mathbf{p}_i + \kappa s_i \sqrt{\frac{j}{\max(n_i - 1, 1)}}\big(\cos\alpha_{ij}\,\mathbf{e}_{i,1} + \sin\alpha_{ij}\,\mathbf{e}_{i,2}\big), \qquad \alpha_{ij} = j\,\pi(3 - \sqrt{5}) + 2\pi\,\mathrm{frac}\big(\xi\,(i-1)\big), \tag{2}$$

where $\kappa = 0.35$ scales the radius, $\pi(3 - \sqrt{5})$ is the golden angle, and the irrational offset ξ decorrelates neighbouring parents. The square-root radial factor covers the disc of radius κs_i with approximately uniform density, and the golden angle spreads the azimuths without repetition. The radial factor vanishes at $j = 0$, so $\mathbf{q}_{i0} = \mathbf{p}_i$; we enforce this equality by assignment. Flattening yields $\mathscr{Q} = \{\mathbf{q}_m\}_{m=1}^{M}$ and an anchor mask $\mathbf{a} \in \{0,1\}^M$; each child inherits s_i. BFCC is deterministic, has no trainable parameters, and adapts to a new sensor through its published angular resolutions (Alg. 1 in the appendix).

3.3 FOOTPRINT-ALIGNED SHARED-SCALE ENCODER

Multi-scale sparse encoding is standard for scene completion, but instantiating separate kernels per resolution and per branch increases the parameter count linearly and still leaves the choice of scale unrelated to how densely the sensor sampled each point. FASSE uses one local operator for all resolutions $\mathcal{G} = \{g_k\}_{k=1}^{K}$ and both branches. At resolution g_k, the projected point features $\mathbf{H}$ are voxelised at $\lfloor \mathbf{p}/g_k \rfloor$ into a sparse tensor $\mathscr{T}_k^{(0)}$, and the same two submanifold convolutions are applied with residual connections, $\mathscr{T}_k^{(l)} = \mathrm{SC}_l(\mathscr{T}_k^{(l-1)}) + \mathscr{T}_k^{(l-1)}$ for $l = 1, 2$. Sharing $\mathrm{SC}_1, \mathrm{SC}_2$ imposes one notion of local geometry, so features from different resolutions are directly comparable. Gathering $\mathscr{T}_k^{(2)}$ back to points gives $\mathbf{f}_i^{(k)}$, which is weighted by the agreement between footprint and voxel size in the log domain:

$$\omega_{ik} = \mathrm{softmax}_k\Big[-\tfrac{(\log g_k - \log s_i)^2}{2\tau_s^2}\Big], \qquad \mathbf{f}_i = \mathbf{W}_o\Big(\textstyle\sum_{k=1}^{K} \omega_{ik}\, \mathbf{f}_i^{(k)}\Big) + \mathbf{b}_o. \tag{3}$$

We use $K = 8$ resolutions $g_k = 0.01 \cdot 2^{k-1}$ m (0.01–1.28 m) and $\tau_s = 0.70$. The gate is a fixed function of s_i and g_k and adds no parameters; all eight resolutions are still evaluated, so sharing reduces storage without shrinking the receptive field. The same FASSE instance encodes $\mathscr{P}$ and $\mathscr{Q}$.

3.4 OBSERVATION-PRESERVING COARSE-TO-FINE RECONSTRUCTION

Decoders that predict a displacement for every point treat measured and hypothesised points alike, so the completed scene drifts from the observation it is conditioned on. We separate the two roles. The shared encoder gives $\mathbf{F}_p = E_{\mathrm{sh}}(\mathscr{P})$ and $\mathbf{F}_q = E_{\mathrm{sh}}(\mathscr{Q})$, and a seed encoder with serialised local aggregation and progressive downsampling condenses the scan into regional seeds $(\mathscr{S}, \mathbf{F}_s) = E_{\mathrm{seed}}(\mathscr{P}, \mathbf{F}_p)$. Conditioned on the seeds, the coarse decoder predicts a residual $\Delta\mathbf{q}_m$ and a feature $\mathbf{z}_m$ for each candidate, and the output is a masked selection,

$$\mathbf{c}_m = a_m\, \mathbf{q}_m + (1 - a_m)\,(\mathbf{q}_m - \Delta\mathbf{q}_m), \qquad \mathbf{r}_{m,u} = \mathbf{c}_m + \beta \tanh\big[D_r(\mathbf{c}_m, \mathbf{z}_m, \mathscr{S}, \mathbf{F}_s, u)\big], \tag{4}$$

where $u = 0, \ldots, U-1$ indexes the children predicted by the refinement decoder D_r and β bounds their offsets. For $a_m = 1$, $\mathbf{c}_m = \mathbf{q}_m$ holds exactly, independently of the predicted residual, and the first child of an anchor is projected back to $\mathbf{r}_{m,0} = \mathbf{c}_m$. Every measured return therefore survives both stages without floating-point drift, while the remaining children recover unobserved structure. The coarse network is trained first and frozen while D_r is trained.

3.5 FRUSTUM-RAY VISIBILITY OBJECTIVE

Chamfer distance measures how close a prediction is to the reference set but not the order in which surfaces are met along a ray, so a point in front of the reference first return, i.e., in space the sensor observed as empty, incurs only a small penalty. FRVO adds two terms that encode this order.

Frustum-local transport. Predictions and references are expressed in spherical coordinates (r, θ, ϕ) and partitioned into common (θ, ϕ, r) cells, so that matching occurs only between points the sensor would observe in the same direction. For cell c with prediction and reference subsets $\mathscr{X}_c, \mathscr{Y}_c$, mean ranges $\bar{r}(\cdot)$, and angular widths $\Delta\theta_c, \Delta\phi_c$, the characteristic length and pairwise cost are

$$\ell_c = \max\Big[\ell_{\min}, \tfrac{\bar{r}(\mathscr{X}_c) + \bar{r}(\mathscr{Y}_c)}{2}\sqrt{\Delta\theta_c \Delta\phi_c}\Big], \qquad M_{ab}^{c} = \frac{\|\mathbf{x}_a - \mathbf{y}_b\|_2^2}{\ell_c^2} + \lambda_{\mathrm{rad}}\frac{|r_a - r_b|}{\ell_c}, \tag{5}$$

where the normalisation makes costs comparable across ranges and the radial term penalises depth disagreement even at small lateral offsets. With uniform marginals, we compute the entropic transport plan $\mathbf{\Pi}_c^{\star}$ with regularisation $\varepsilon_{\mathrm{ot}}$ by six log-domain Sinkhorn iterations (Cuturi, 2013) and aggregate cells by prediction mass, $\mathscr{L}_{\mathrm{FT}} = \sum_c |\mathscr{X}_c| \langle \mathbf{\Pi}_c^{\star}, \mathbf{M}^c \rangle / \sum_c |\mathscr{X}_c|$.

Ray-wise first-return consistency. Transport aligns local structure but does not separate occupied from empty space. We bin points into common angular rays $b \in \mathscr{B}$ that are non-empty for both sets, take the reference first return $r_b^{\star} = \min_{\mathbf{y} \in \mathscr{Y}_b} \|\mathbf{y}\|_2$, and estimate the predicted first return with a soft minimum $\hat{r}_b = \sum_i \pi_{b,i} r_{b,i}$, $\pi_{b,i} = \mathrm{softmax}_i(-r_{b,i}/\tau_r)$, so that gradients reach more than one point. The ray objective is

$$\mathscr{L}_{\mathrm{ray}} = \frac{1}{|\mathscr{B}|}\sum_{b \in \mathscr{B}}\Big[\mathrm{SmoothL1}\big(\hat{r}_b - r_b^{\star}\big) + \lambda_f\big[r_b^{\star} - \delta_f - \hat{r}_b\big]_+^2\Big], \tag{6}$$

where the first term aligns first returns and the one-sided second term activates only when a predicted surface lies more than δ_f in front of the reference first return. Both terms act on subsampled points and bounded numbers of cells and rays (Appendix A), and neither is evaluated at inference.

3.6 Training Objective and Model Size

Each stage minimises $\mathscr{L} = \mathscr{L}_{\text{CD}} + \lambda_{\text{FT}}\mathscr{L}_{\text{FT}} + \lambda_{\text{ray}}\mathscr{L}_{\text{ray}}$ with $\lambda_{\text{FT}} = \lambda_{\text{ray}} = 0.02$, where $\mathscr{L}_{\text{CD}}$ is the bidirectional rooted Euclidean Chamfer distance, applied to $\mathscr{C}$ in the first stage and to $\mathscr{R}$ in the second. BFCC, the footprint gate, the anchor projection, and FRVO contain no trainable parameters. The coarse network has 1.12M parameters, and the complete coarse-to-refined model has 1.23M; compared with LiNeXt (1.99M) and LiNeXt† (2.10M), these are reductions of 43.7% and 41.4%, which come mainly from sharing the sparse kernels across resolutions and branches.

4 Experiments

4.1 Experimental Setup

Datasets. We follow the protocol of LiDiff (Nunes et al., 2024). On SemanticKITTI (Behley et al., 2019), the dense reference of each scan is built by aggregating neighbouring scans with the provided poses and removing dynamic objects with the semantic labels; we train on sequences 00–07 and 09–10 and evaluate on sequence 08. On KITTI-360 (Liao et al., 2023) we evaluate on sequence 00 with the model trained on SemanticKITTI and no fine-tuning, which measures cross-dataset generalisation.

Metrics. We report the rooted bidirectional Chamfer Distance (CD) Tenenbaum et al. (1977), the Jensen–Shannon Divergence (JSD) Lin (1991) between the completed and reference scenes in 3D and in bird's-eye view (JSD 3D, JSD BEV), and voxel occupancy IoU at 0.5, 0.2, and 0.1 m. The coarse IoU reflects whether the scene volume is filled, whereas the fine IoUs indicate whether points lie on the reference surfaces and are therefore the most direct measure of how well unobserved regions are recovered. Efficiency is reported as trainable parameters and per-scan inference time on one NVIDIA RTX 6000 Ada GPU with batch size one, using official implementations and checkpoints and excluding data loading. † denotes the configuration with refinement. Implementation details are given in Appendix A.

4.2 Comparison with State-of-the-Art Methods

SemanticKITTI. Table 1 compares BeamScene with voxel, implicit, diffusion-based, and single-pass methods. Without refinement, BeamScene improves on LiNeXt in all six metrics, reducing CD from 0.214 to 0.205 and raising voxel IoU from 41.07/19.45/6.30 to 42.06/24.91/8.74 at 0.5/0.2/0.1 m. Among methods without refinement, only MID attains a higher IoU at 0.1 m, at a 2.5× larger CD. With refinement, BeamScene† reaches 0.144 CD, 0.479 JSD 3D, 0.323 JSD BEV, and 42.87/31.93/15.85 IoU, the best value of all methods in every column.

Where the gain comes from. Among recent single-pass methods CD lies in a narrow band, so the reduction from 0.214 to 0.205 is modest. The same comparison raises IoU at 0.2 m by 5.46 points (+28.1%) and at 0.1 m by 2.44 points (+38.7%). Fig. 4(c) shows that, on both datasets, the relative gain is small for CD, JSD, and IoU at 0.5 m and concentrated at 0.2 and 0.1 m. Since fine-resolution occupancy responds to whether predicted points fall on the reference surfaces, the pattern indicates that BeamScene gains by placing points in previously unfilled regions rather than by redistributing points that were already close to the reference, which is the behaviour that footprint-proportional allocation is designed to produce.

KITTI-360. Without fine-tuning, BeamScene obtains 0.213 CD and 37.68/19.75/6.20 IoU, improving on LiNeXt in every column, and BeamScene† again is best in all six metrics (0.147 CD, 41.94/31.19/14.78 IoU). Because the footprint is computed from sensor geometry rather than from dataset statistics, the candidate allocation and scale routing transfer to the new acquisition trajectory and scene layout without retuning.

Qualitative results. Fig. 5 compares completed scenes. The outputs of LiDiff and ScoreLiDAR remain organised along the original scan lines and leave large occluded regions sparse. LiNeXt

Table 1: Scene completion on SemanticKITTI and KITTI-360. Within each dataset, the upper block lists methods without refinement and the lower block methods with refinement (†). Colours mark the best and second-best results within each block; "–" denotes a result not reported by the corresponding work. KITTI-360 results use the model trained on SemanticKITTI without fine-tuning.

	Method	CD↓	JSD 3D↓	JSD BEV↓	Vox. $IoU_{0.5m}$ ↑	Vox. $IoU_{0.2m}$ ↑	Vox. $IoU_{0.1m}$ ↑
SemanticKITTI	LMSCNet (Roldão et al., 2020)	0.641	–	0.431	30.83	12.09	3.65
	LODE (Li et al., 2023)	1.029	–	0.451	33.81	16.39	5.00
	MID (Vizzo et al., 2022)	0.503	–	0.470	31.58	22.72	13.14
	PVD (Zhou et al., 2021)	1.256	–	0.498	15.91	3.97	0.60
	LiDiff (Nunes et al., 2024)	0.434	0.564	0.444	31.47	16.79	4.67
	LiDPM (Martyniuk et al., 2025)	0.446	0.532	0.440	34.09	19.45	6.27
	ScoreLiDAR (Zhang et al., 2025)	0.406	–	0.425	–	–	–
	LiNeXt (He et al., 2026)	0.214	0.494	0.336	41.07	19.45	6.30
	BeamScene (ours)	0.205	0.487	0.332	42.06	24.91	8.74
	LiDiff† (Nunes et al., 2024)	0.376	0.573	0.416	32.43	22.99	13.40
	ScoreLiDAR† (Zhang et al., 2025)	0.342	–	0.399	–	–	–
	LiDPM† (Martyniuk et al., 2025)	0.376	0.542	0.403	36.59	25.76	14.93
	LiNeXt† (He et al., 2026)	0.149	0.481	0.331	41.97	31.25	15.09
	BeamScene† (ours)	0.144	0.479	0.323	42.87	31.93	15.85
KITTI-360	LMSCNet (Roldão et al., 2020)	0.979	–	0.496	26.17	9.21	2.88
	LODE (Li et al., 2023)	1.565	–	0.483	33.06	15.24	4.68
	MID (Vizzo et al., 2022)	0.637	–	0.476	33.05	21.32	11.30
	LiDiff (Nunes et al., 2024)	0.564	–	0.459	33.23	17.55	4.88
	ScoreLiDAR (Zhang et al., 2025)	0.472	–	0.444	–	–	–
	LiNeXt (He et al., 2026)	0.217	0.508	0.355	36.85	16.91	5.73
	BeamScene (ours)	0.213	0.503	0.345	37.68	19.75	6.20
	LiDiff† (Nunes et al., 2024)	0.517	–	0.446	33.43	22.04	11.84
	ScoreLiDAR† (Zhang et al., 2025)	0.452	–	0.437	–	–	–
	LiNeXt† (He et al., 2026)	0.149	0.499	0.339	41.88	29.34	13.90
	BeamScene† (ours)	0.147	0.491	0.335	41.94	31.19	14.78

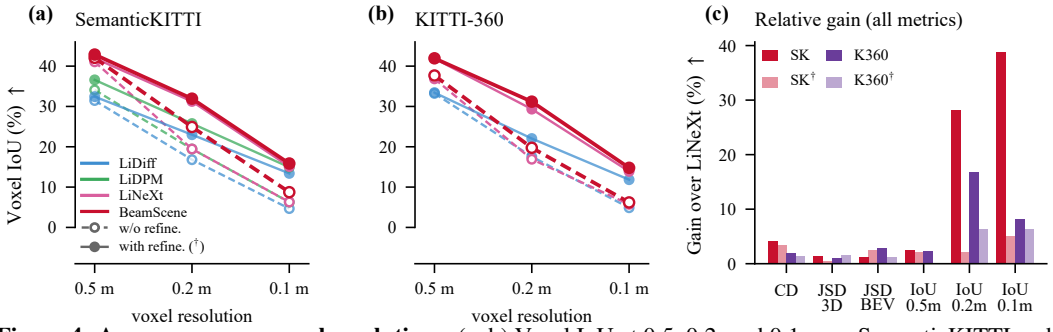

Figure 4: Accuracy across voxel resolutions. (a, b) Voxel IoU at 0.5, 0.2, and 0.1 m on SemanticKITTI and KITTI-360 (values from Table 1; dashed: without refinement, solid: with refinement). (c) Relative improvement of BeamScene over LiNeXt on each metric (for CD and JSD, the relative reduction). On both datasets the gain is concentrated at 0.2 and 0.1 m, i.e., in whether points lie on the reference surfaces.

produces a denser scene but retains empty areas at long range, where fixed-variance perturbation confines candidates to small neighbourhoods around the returns. BeamScene fills these regions more uniformly, and ground and facade surfaces remain continuous rather than breaking into scan-aligned bands; after refinement, the input layout is preserved, as expected from exact anchor retention. These observations agree with the lower CD and JSD and the higher fine-resolution IoU in Table 1.

4.3 Ablation Studies

Component ablation. Table 2(a) removes one component at a time, and Fig. 6 shows the corresponding completions. Every component contributes to all six metrics. Removing *FRVO* causes the largest degradation in CD and both JSDs (0.224, 0.503, 0.349): trained with an order-invariant set distance only, the network scatters points in front of observed surfaces, and road boundaries lose contrast in Fig. 6. Removing *BFCC* causes the largest loss at 0.1 m (−1.06 points), consistent with its role of deciding where candidates exist before any reconstruction; the distant arms of the intersection become sparse, where the footprint is largest and allocation matters most. Removing *FASSE* produces

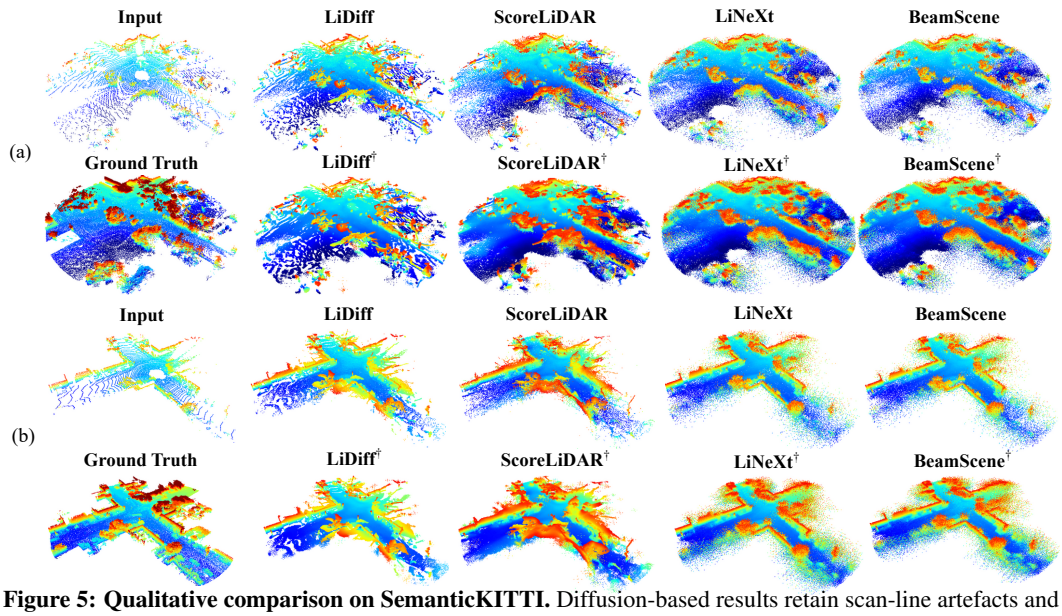

Figure 5: Qualitative comparison on SemanticKITTI. Diffusion-based results retain scan-line artefacts and leave occluded regions sparsely populated; LiNeXt leaves gaps at long range. BeamScene recovers denser and more uniform geometry there and preserves the layout of the input scan; refinement (†) increases coverage without displacing measured returns. More examples are given in Appendix C.

Table 2: Ablations on SemanticKITTI (without refinement). (a) Removing one component at a time: "w/o BFCC" uses uniform replication with fixed-variance perturbation, "w/o FASSE" uses independent per-scale kernels with concatenation, and "w/o FRVO" trains with the Chamfer loss only. (b)–(d) Internal ablations of each component. FT: frustum-local transport; Ray: first-return term plus free-space penalty. Colours mark the best and second-best results in (a); bold marks the best value within each of (b)–(d), determined before rounding.

	Variant	CD↓	JSD 3D↓	JSD BEV↓	Vox. $IoU_{0.5m}$ ↑	Vox. $IoU_{0.2m}$ ↑	Vox. $IoU_{0.1m}$ ↑
(a)	**BeamScene** (full)	0.205	0.487	0.332	42.06	24.91	8.74
	w/o BFCC	0.216	0.501	0.344	41.07	23.89	7.68
	w/o FASSE	0.208	0.493	0.341	41.72	24.32	8.27
	w/o FRVO	0.224	0.503	0.349	40.94	23.61	7.99
(b) BFCC	full (footprint alloc. + tangent plane)	**0.205**	**0.487**	**0.332**	**42.06**	**24.91**	**8.74**
	w/ distance-grouped allocation	0.211	0.489	0.341	41.79	24.73	8.68
	w/ isotropic 3D expansion	0.212	0.490	0.343	41.44	24.31	8.42
(c) FASSE	full (shared kernels + footprint gate)	**0.205**	0.487	**0.332**	42.06	24.91	8.74
	w/o kernel sharing	0.208	0.482	0.337	**42.27**	25.46	**9.53**
	w/o footprint routing (uniform)	0.207	**0.472**	0.335	42.26	**25.51**	9.44
(d) FRVO	CD only	0.224	0.503	0.349	40.94	23.61	7.99
	CD + FT	0.210	**0.478**	0.336	**42.20**	**25.70**	**9.52**
	CD + Ray	0.216	0.478	0.356	39.83	24.37	8.35
	CD + FT + first-return	0.211	0.488	0.340	41.76	24.79	8.65
	CD + FT + Ray (full)	**0.205**	0.487	**0.332**	42.06	24.91	8.74

smaller but uniform losses, showing that footprint-conditioned routing of shared features adds a benefit that plain multi-scale concatenation does not provide.

Inside BFCC. Both halves of BFCC are needed (Table 2(b)): distance-grouped allocation mainly degrades JSD BEV (0.341), i.e., the global layout, whereas isotropic expansion degrades every metric and fine IoU most (8.42), confirming that spreading candidates along the ray wastes them where the measurement is already reliable.

Inside FASSE and FRVO. The internal ablations reveal trade-offs rather than uniform gains (Table 2(c, d)). Independent kernels or uniform scale weighting reach higher fine IoU and lower JSD 3D but worse CD and JSD BEV; the independent-kernel variant additionally requires more parameters. Removing both, i.e., independent kernels with plain concatenation ("w/o FASSE"), degrades all six metrics. We therefore interpret the footprint gate as regularising scale selection toward the physical sampling support, which improves point-level accuracy and BEV layout at a small cost in fine occupancy. In FRVO, frustum-local transport alone yields the largest fine-IoU gain over CD only (+1.53 at 0.1 m), while the ray term is what lowers CD from 0.210 to 0.205 and JSD BEV from 0.336

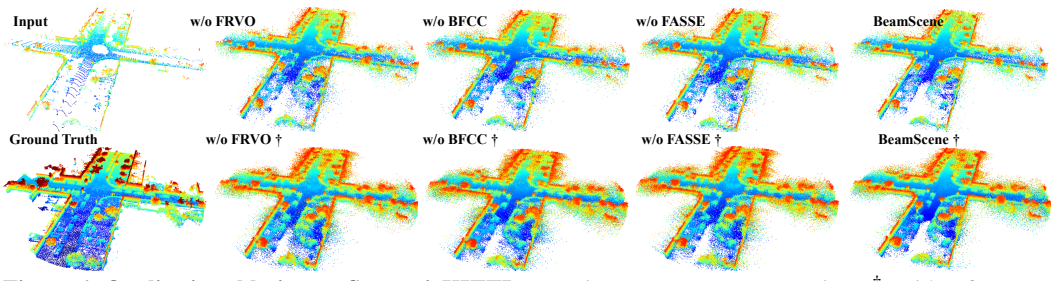

Figure 6: Qualitative ablation on SemanticKITTI, removing one component at a time (†: with refinement). Without FRVO, points scatter in front of observed surfaces; without BFCC, distant road arms remain sparse; without FASSE, roadside structures are reproduced less sharply.

to 0.332; the first-return term without the free-space penalty does not improve on CD + FT (0.211 CD), so the gain requires the one-sided free-space penalty. The ray term alone is insufficient (0.356 JSD BEV), which we attribute to first-return depth being poorly constrained before frustum-local transport aligns the local structure. The full objective gives the best CD and JSD BEV, the metrics most sensitive to spurious points in front of surfaces.

4.4 Efficiency Analysis

Table 3 and Fig. 2 report accuracy together with model size and runtime. BeamScene reaches 0.205 CD with 1.12M parameters in 0.234 s, compared with 0.214 CD, 1.99M, and 0.245 s for LiNeXt, i.e., 43.7% fewer parameters. BeamScene† reaches 0.144 CD with 1.23M parameters in 0.41 s, against 0.149 CD, 2.10M, and 0.43 s for LiNeXt† (41.4% fewer parameters).

Table 3: Accuracy, size, and runtime on SemanticKITTI (RTX 6000 Ada, batch size one).

Method	CD↓	Params (M)↓	Time (s)↓
LiDiff	0.434	32.67	23.17
LiDPM	0.446	32.67	10.26
ScoreLiDAR	0.406	32.67	4.06
LiNeXt	0.214	1.99	0.245
BeamScene	0.205	1.12	0.234
LiDiff†	0.376	54.40	33.53
LiDPM†	0.377	54.40	15.45
ScoreLiDAR†	0.342	54.40	5.19
LiNeXt†	0.149	2.10	0.43
BeamScene†	0.144	1.23	0.41

The gap to diffusion-based completion is much larger. BeamScene is 17.4× faster than ScoreLiDAR, 43.8× faster than LiDPM, and 99.0× faster than LiDiff, and LiDiff uses 29.2× as many parameters; with refinement, BeamScene† is 12.7× faster than ScoreLiDAR† and 81.8× faster than LiDiff†, with 44.2× fewer parameters, while also attaining lower CD. Because BFCC, the footprint gate, and FRVO add no trainable parameters and FRVO is used only in training, none of the accuracy gains increases deployment cost.

4.5 Limitations

First, the footprint model of Eq. (1) assumes the nominal, uniform angular resolution of a mechanical scanner; sensors with non-uniform or adaptive scan patterns require a per-beam footprint before allocation and routing apply. Second, FRVO supervises visibility only along rays from the acquisition viewpoint of the input scan; multi-view visibility supervision is left to future work. Third, only the properties in Appendix B hold by construction; the absence of free-space violations is encouraged by FRVO during training, not enforced at inference.

5 Conclusion

We presented BeamScene, a single-pass LiDAR scene completion framework that derives candidate placement, feature scale, observation retention, and visibility supervision from the beam footprint of the sensor. It improves fine-resolution occupancy by up to 38.7% over the strongest single-pass baseline with 43.7% fewer parameters and is 17–99× faster than diffusion-based completion. Future work will add synchronised camera images and extend the footprint model to non-uniform scan patterns.

AI USE STATEMENT

Generative AI tools were used for language editing and LaTeX formatting and layout. They were not used to generate research ideas, implement or run experiments, or produce any numerical result. All AI-assisted suggestions were reviewed, verified, and revised by the authors. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work uses the publicly available SemanticKITTI and KITTI-360 datasets under their respective licences and involves no human subjects. Improved LiDAR completion may support safer autonomous driving; completed geometry is an inference and should not replace measured data in safety-critical decisions without validation.

REPRODUCIBILITY STATEMENT

The architecture is specified in Sec. 3, BFCC is given as pseudo-code in Alg. 1, all hyperparameters are listed in Table 4, and the data protocol and timing procedure are described in Sec. 4.1 and Appendix A.

REFERENCES

Ben Agro, Quinlan Sykora, Sergio Casas, Thomas Gilles, and Raquel Urtasun. UnO: Unsupervised occupancy fields for perception and forecasting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14487–14496, 2024.

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jürgen Gall. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9297–9307, 2019.

Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3075–3084, 2019.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

Mostafa Elhashash, Hessah Albanwan, and Rongjun Qin. A review of mobile mapping systems: From sensors to applications. *Sensors*, 22(11):4262, 2022. doi: 10.3390/s22114262.

Yuan Gao, Mattia Piccinini, Yuchen Zhang, Dingrui Wang, Korbinian Moller, Roberto Brusnicki, Baha Zarrouki, Alessio Gambi, Jan Frederik Totz, Kai Storms, et al. Foundation models in autonomous driving: A survey on scenario generation and scenario analysis. *IEEE Open Journal of Intelligent Transportation Systems*, 2026.

Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9224–9232, 2018. doi: 10.1109/CVPR.2018.00961.

Wenzhe He, Xiaojun Chen, Ruiqi Wang, Ruihui Li, Huilong Pi, Jiapeng Zhang, Zhuo Tang, and Kenli Li. LiNeXt: Revisiting LiDAR completion with efficient non-diffusion architectures. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(6):4672–4680, 2026. doi: 10.1609/aaai.v40i6.42468.

Azhar Hussian, Martin Vossiek, and Vasileios Belagiannis. Towards real-time and adaptable LiDAR scene completion. *arXiv preprint arXiv:2608.16490*, 2026.

Veli Ilci and Charles Toth. High definition 3d map creation using GNSS/IMU/LiDAR sensor integration to support autonomous vehicle navigation. *Sensors*, 20(3):899, 2020. doi: 10.3390/s20030899.

Esraa Khatab, Fares Fathy, Abdallah AlKholy, and Omar Shalash. Sensor fusion and perception for autonomous driving: A critical review of modalities, ai models, algorithms, and industry configurations. *Machine Learning and Knowledge Extraction*, 8(7):199, 2026.

N Krishnakumar, Jaya Sreevalsan-Nair, and Prashant Goswami. Road and building reconstruction from 3d lidar point clouds: A scoping review. *Archives of Computational Methods in Engineering*, pp. 1–36, 2026.

Sofia E Larsen, Mei-Lin Huang, and Kwame Adjei. City-scale digital twins for smart urban governance: Frameworks, implementation challenges, and strategic pathways. *Global Intelligence Technology Horizons*, 1(1), 2026.

Pengfei Li, Ruowen Zhao, Yongliang Shi, Hao Zhao, Jirui Yuan, Guyue Zhou, and Ya-Qin Zhang. LODE: Locally conditioned eikonal implicit scene completion from sparse LiDAR. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8269–8276, 2023.

Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (3):3292–3310, March 2023. doi: 10.1109/TPAMI.2022.3179507.

Fangzhou Lin, Yun Yue, Songlin Hou, Xuechu Yu, Yajun Xu, Kazunori D. Yamada, and Ziming Zhang. Hyperbolic chamfer distance for point cloud completion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14595–14606, 2023.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Tetiana Martyniuk, Gilles Puy, Alexandre Boulch, Renaud Marlet, and Raoul de Charette. LiDPM: Rethinking point diffusion for LiDAR scene completion. In *IEEE Intelligent Vehicles Symposium (IV)*, pp. 555–560, 2025. doi: 10.1109/IV64158.2025.11097538.

Lucas Nunes, Rodrigo Marcuzzi, Benedikt Mersch, Jens Behley, and Cyrill Stachniss. Scaling diffusion models to real-world 3D LiDAR scene completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14770–14780, 2024.

Bricen Raynold and Dhananjai M Rao. Accelerating work-building assignments for large-scale urban digital twins using regression models. In *Proceedings of the 40th ACM SIGSIM International Conference on Principles of Advanced Discrete Simulation*, pp. 169–176, 2026.

Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. LMSCNet: Lightweight multiscale 3D semantic completion. In *International Conference on 3D Vision (3DV)*, pp. 111–119, 2020.

RC Bolles HG Barrow JM Tenenbaum, RC Bolles, HC Wolf, et al. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *Proceedings of the 5th international joint conference on Artificial intelligence*, pp. 659–663, 1977.

Ignacio Vizzo, Benedikt Mersch, Rodrigo Marcuzzi, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. Make it dense: Self-supervised geometric scan completion of sparse 3D LiDAR scans in large outdoor environments. *IEEE Robotics and Automation Letters*, 7:8534–8541, 2022.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5499–5509, 2021.

Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, Kun Ma, et al. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. In *International Conference on Learning Representations*, volume 2026, pp. 157518–157556, 2026.

Qingdong Xu, Jiajun Zhu, Shilin Zhu, Xinjing He, Chao Lu, Huanran Wang, and Jiyao Zhang. Patchscene: Patch-based voxel diffusion for large-scale scene completion. *arXiv preprint arXiv:2606.03915*, 2026.

Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. PoinTr: Diverse point cloud completion with geometry-aware transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12498–12507, 2021.

Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. PCN: Point completion network. In *International Conference on 3D Vision (3DV)*, pp. 728–737, 2018.

Shengyuan Zhang, An Zhao, Ling Yang, Zejian Li, Chenye Meng, Haoran Xu, Tianrun Chen, AnYang Wei, Perry Pengyun Gu, and Lingyun Sun. Distilling diffusion models to efficient 3D LiDAR scene completion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5007–5016, 2025.

An Zhao, Shengyuan Zhang, Zejian Li, Ling Yang, Pei Chen, Jiale Wu, Haoran Xu, AnYang Wei, Perry Pengyun Gu, and Lingyun Sun. Diffusion distillation with direct preference optimization for efficient 3d lidar scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13070–13078, 2026.

Linqi Zhou, Yilun Du, and Jiajun Wu. 3D shape generation and completion through point-voxel diffusion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5826–5835, 2021.

APPENDIX

A IMPLEMENTATION DETAILS

Data. Each training pair contains a preprocessed partial point set of $N = 18{,}000$ points and a reference set of $M = 180{,}000$ points, obtained offline by adaptive voxel downsampling and sampling without replacement following prior work (Nunes et al., 2024; He et al., 2026).

Network and objective. All hyperparameters are listed in Table 4. BFCC uses nominal horizontal and vertical angular resolutions of 0.20° and 0.42°, retains one zero-offset candidate per input point, and disables radial jitter. The refinement decoder produces $U = 6$ children per coarse point, i.e., 1,080,000 refined points. FRVO operates in the LiDAR-centred frame, but its spherical partitions are independent of the BFCC footprint discretisation. The transport term retains at most 8,192 points from each set, at most 512 common frustum cells, and at most 64 points per cell; the ray term retains at most 16,384 points, 1,024 common rays, and 128 points per ray. These bounds cap the additional pairwise memory of FRVO.

Training. The coarse and refinement stages are trained separately; the trained coarse network is frozen while the refinement decoder is optimised. Optimiser: We employ the Adam optimization algorithm, with the initial learning rates set at 0.0002 and 0.0005 respectively;epochs: Coarse / refinement are trained for 10/5 epochs respectively; batch size: 16; training time:48 hours. All models are trained on 4 NVIDIA RTX 6000 Ada GPUs and evaluated on a single GPU .

Parameter count. All parameter counts include every trainable weight of the deployed network. The coarse network has 1.12M parameters and the complete coarse-to-refined model 1,221,273 (1.23M). Relative to an otherwise identical two-stage backbone with independent kernels per resolution and per branch, kernel sharing reduces the parameters of the complete model by 41.80%; relative to LiNeXt† (2.10M) the reduction is 41.4%, and relative to LiNeXt (1.99M) the coarse model is 43.7% smaller.

Timing. Inference times of all methods are measured on the same GPU with batch size one using the official implementations and released checkpoints, excluding data loading.

B PROPERTIES THAT HOLD BY CONSTRUCTION

The term *physically consistent* in this paper describes how the pipeline is constructed. Three of its consequences are exact properties of the architecture and therefore require no empirical verification; they hold for every input, independently of training.

Proposition 1 (Exact budget)**.** *If $N \le M \le N n_{\max}$, BFCC returns exactly M candidates with $1 \le n_i \le n_{\max}$ for every return.*

Table 4: Hyperparameters of BeamScene.

Module	Symbol	Meaning	Value
Data	N, M, U	input points, coarse points, children per point	18,000, 180,000, 6
BFCC	$\Delta\theta$, $\Delta\phi$	horizontal / vertical angular resolution	0.20°, 0.42°
	q_ℓ, q_u, γ	area quantile clipping, allocation exponent	0.05, 0.95, 1
	$n_{\max}$	maximum candidates per return	24
	κ, ξ	expansion radius factor, azimuth offset	0.35, 0.754877666
FASSE	K, g_k	number of resolutions, voxel sizes	8, $0.01 \cdot 2^{k-1}$ m
	τ_s	routing bandwidth	0.70
FRVO (FT)	$\Delta\theta_c \times \Delta\phi_c$, Δr_c	frustum cell size	$5^\circ \times 3^\circ$, 5 m
	λ_{rad}, Sinkhorn iters	radial cost weight, iterations	0.25, 6
FRVO (Ray)	$\Delta\theta_r \times \Delta\phi_r$	ray bin size	$1^\circ \times 1^\circ$
	τ_r, δ_f	softmin temperature, free-space tolerance	0.10, 0.10 m
Objective	λ_{FT}, λ_{ray}	loss weights	0.02, 0.02

Algorithm 1 Beam-Footprint-Aware Candidate Construction (BFCC)

Require: scan $\{\mathbf{p}_i\}_{i=1}^N$, budget M, angular resolutions $\Delta\theta, \Delta\phi$, parameters $s_{\min}, q_\ell, q_u, \gamma, n_{\max}, \kappa, \xi$
Ensure: candidates $\{\mathbf{q}_m\}_{m=1}^M$, anchor mask $\mathbf{a}$, footprint scales $\{s_m\}$

for $i = 1, \dots, N$ **do**
  $r_i \leftarrow \|\mathbf{p}_i\|_2$; $A_i \leftarrow r_i^2 \Delta\theta\Delta\phi$; $s_i \leftarrow \max(s_{\min}, r_i\sqrt{\Delta\theta\Delta\phi})$ ▷ Eq. (1)
end for
$\bar{A}_i \leftarrow \text{clip}(A_i, Q_{q_\ell}, Q_{q_u})$; $w_i \leftarrow \bar{A}_i^\gamma / \sum_j \bar{A}_j^\gamma$
$n_i \leftarrow \min\big(n_{\max}, 1 + \lfloor (M-N)\, w_i \rfloor\big)$ ▷ one anchor per return;
distribute the remaining $M - \sum_i n_i$ candidates by largest remainder among returns with $n_i < n_{\max}$
for $i = 1, \dots, N$ **do**
  $\mathbf{u}_i \leftarrow \mathbf{p}_i / (r_i + \varepsilon_0)$; build orthonormal $(\mathbf{e}_{i,1}, \mathbf{e}_{i,2}) \perp \mathbf{u}_i$
  for $j = 0, \dots, n_i - 1$ **do**
    $\alpha_{ij} \leftarrow j\pi(3-\sqrt{5}) + 2\pi\,\text{frac}(\xi(i-1))$; $\rho_{ij} \leftarrow \kappa s_i \sqrt{j / \max(n_i - 1, 1)}$
    $\mathbf{q}_{ij} \leftarrow \mathbf{p}_i + \rho_{ij}(\cos\alpha_{ij}\mathbf{e}_{i,1} + \sin\alpha_{ij}\mathbf{e}_{i,2})$ ▷ Eq. (2)
  end for
  $\mathbf{q}_{i0} \leftarrow \mathbf{p}_i$; $a_{i0} \leftarrow 1$; $a_{ij} \leftarrow 0$ for $j > 0$ ▷ exact anchor
end for
return flattened $\{\mathbf{q}_{ij}\}$, $\{a_{ij}\}$, and inherited scales $\{s_i\}$

Proof sketch. Each return receives one anchor, so $\sum_i n_i \geq N$. The bounded largest-remainder step assigns the remaining candidates one at a time to returns with $n_i < n_{\max}$; such a return exists whenever $\sum_i n_i < M \leq N n_{\max}$, so the step terminates with $\sum_i n_i = M$. □

Proposition 2 (Observation retention)**.** *Every measured return $\mathbf{p}_i$ of the input appears exactly, i.e., bit-identically, in the coarse output $\mathscr{C}$ and in the refined output $\mathscr{R}$.*

Proof. By construction $\mathbf{q}_{i0} = \mathbf{p}_i$ with $a_{i0} = 1$. Eq. (4) gives $\mathbf{c}_m = \mathbf{q}_m$ for $a_m = 1$, and the stored coordinate is selected rather than recomputed, so no floating-point error is introduced. The first child of an anchor parent is assigned $\mathbf{r}_{m,0} = \mathbf{c}_m$. Hence $\mathbf{p}_i \in \mathscr{C}$ and $\mathbf{p}_i \in \mathscr{R}$. □

Proposition 3 (Footprint-matched scale selection)**.** *For every point, the routing weights ω_{ik} of Eq. (3) are maximal at the resolution g_k closest to s_i in the log domain and decrease monotonically with $|\log g_k - \log s_i|$.*

Proof. The logits $-(\log g_k - \log s_i)^2 / (2\tau_s^2)$ are a strictly decreasing function of $|\log g_k - \log s_i|$, and softmax preserves their order. □

The remaining consequence, the absence of predicted surfaces in observed free space, is not guaranteed at inference; FRVO penalises such configurations during training, and its effect is measured through CD, JSD, and IoU in Table 2(d).

C Additional Qualitative Results

Figs. 7–9 show further SemanticKITTI scenes. In each figure the upper row shows the input and the results without refinement and the lower row shows the ground truth and the results with refinement (†).

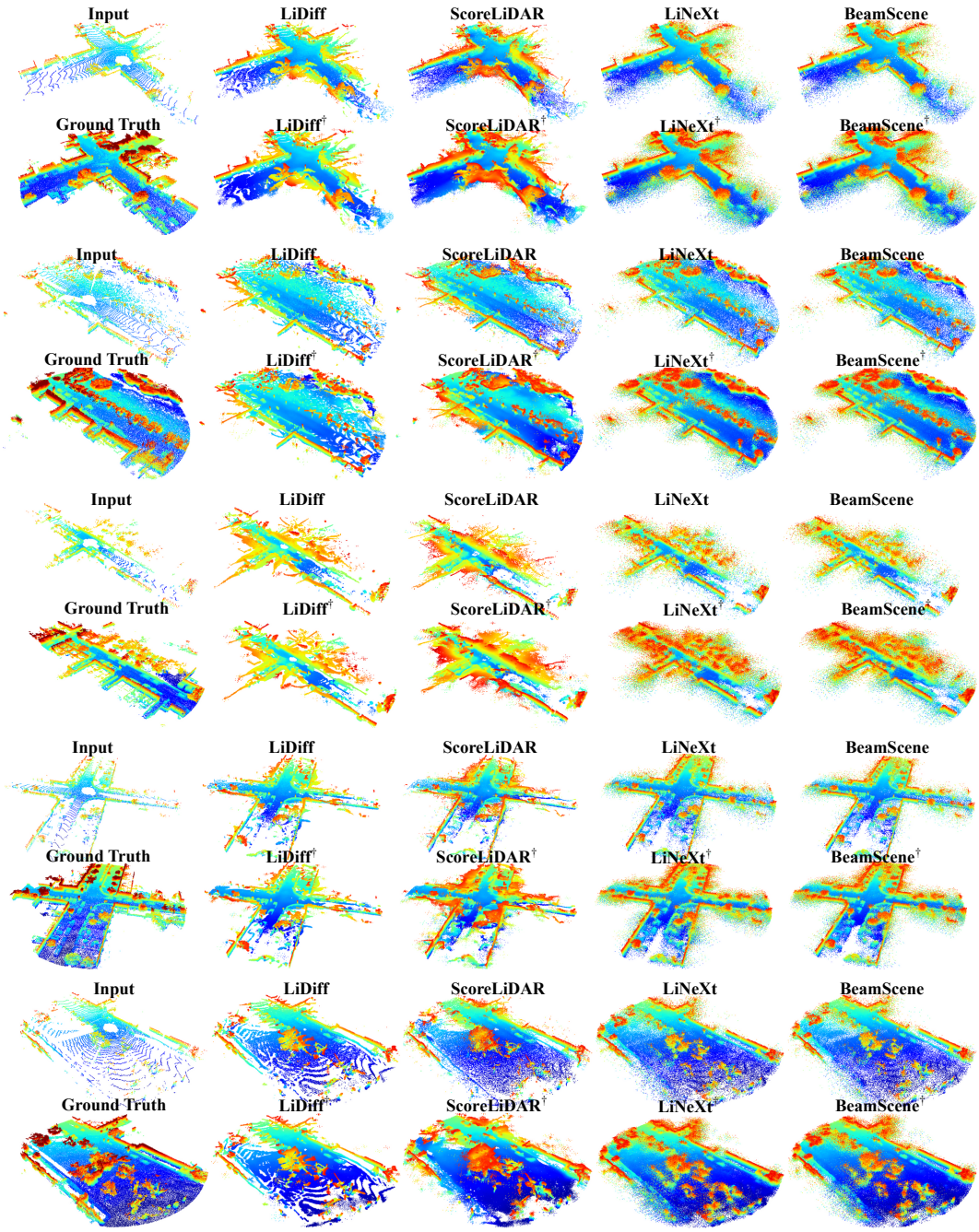

Figure 7: Additional qualitative comparisons on SemanticKITTI (scenes 1–5).

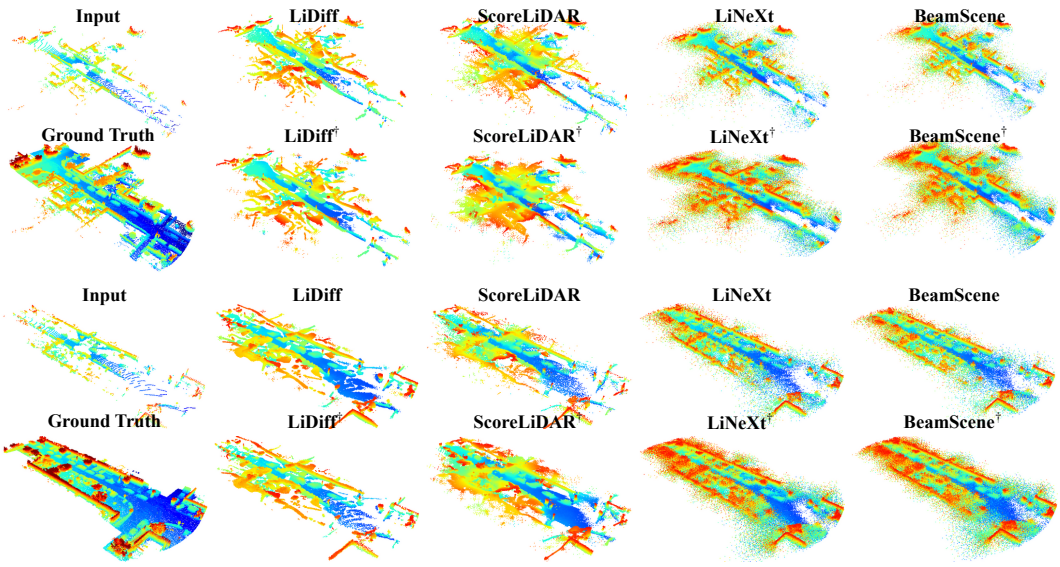

Figure 8: Additional qualitative comparisons on SemanticKITTI (scenes 6–7).

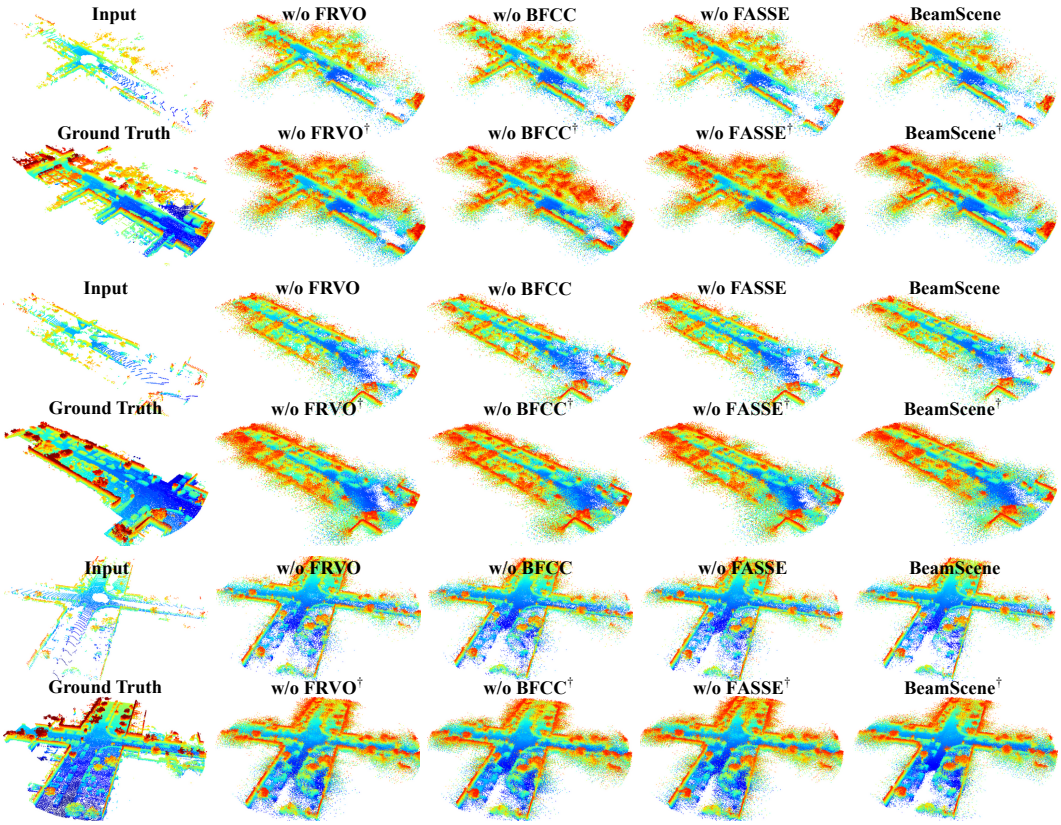

Figure 9: Additional qualitative ablations on SemanticKITTI, removing one component at a time.