VIPER: VISUAL PRIOR RETRIEVAL FOR MULTI-MODAL POINT CLOUD COMPLETION

Anonymous authors
Paper under double-blind review

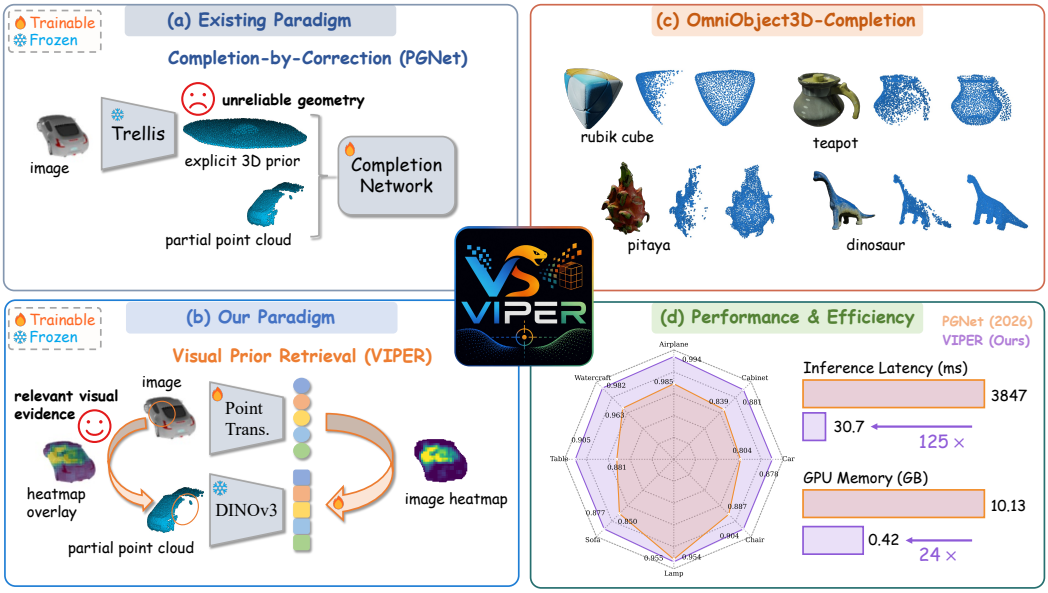


Figure 1: **Motivation and key contributions of VIPER**. **(a)** PGNet exemplifies the *Completion-by-Correction* paradigm, generating an explicit image-derived 3D prior as an intermediate geometric representation; inaccuracies in this prior may introduce misleading cues. **(b)** VIPER offers a feature-space alternative by treating frozen DINOv3 features as queryable visual priors and adaptively integrating point-conditioned evidence. **(c)** We introduce OmniObject3D-Completion, a joint-category benchmark with broad object coverage for developing and evaluating multi-modal point cloud completion methods. **(d)** VIPER achieves better completion performance than PGNet, with a 125× end-to-end inference speedup and a 24× reduction in GPU memory usage.

ABSTRACT

We introduce **VIPER**, a **simple and effective** visual prior retrieval framework for multi-modal point cloud completion. Explicit image-derived 3D priors can provide structural guidance, but inaccuracies in these priors may introduce misleading geometric cues, and generating them incurs substantial computational overhead. VIPER instead treats image representations extracted by a frozen vision foundation model as queryable visual priors, using partial geometry to retrieve evidence relevant to missing or ambiguous structures. The retrieved priors adaptively enhance point representations, which guide global shape recovery and are reused throughout progressive geometric refinement. On ShapeNet-ViPC, VIPER achieves state-of-the-art performance with **16.4% lower average Chamfer Distance** and a **125× end-to-end inference speedup** over PGNet. We further introduce **OmniObject3D-Completion**, a joint-category benchmark with broad object coverage for developing and evaluating multi-modal point cloud completion methods. We will release the code, pretrained model weights, and the OmniObject3D-Completion benchmark upon publication.

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.

1 INTRODUCTION

Point cloud completion (PCC) is a fundamental task in 3D vision that aims to recover complete geometric structures from incomplete, sparse, or occluded point cloud observations. As an important upstream component of 3D perception systems, PCC benefits a wide range of applications, including autonomous driving Chen et al. (2024), archaeological site restoration Li et al. (2025b), object recognition Liang et al. (2023); Du et al. (2024b), 3D detection Liang et al. (2025a;b), robotic perception Iwase et al. (2025), and scene understanding Fang et al. (2026); He et al. (2026; 2025). However, real-world point clouds often suffer from severe information loss caused by limited sensor viewpoints, self-occlusion, and measurement noise, making the recovery of unobserved geometry inherently challenging Guo et al. (2021); Fei et al. (2022).

Single-modal PCC Yuan et al. (2018); Tesema et al. (2024) has achieved substantial progress through advances in point cloud representation learning, Transformer architectures, and geometric priors. Nevertheless, these methods infer missing structures solely from incomplete observations and learned priors. Under severe occlusion or extreme sparsity, such limited evidence can lead to over-smoothed surfaces, incomplete structures, or semantically inconsistent predictions.

Images provide complementary appearance and structural cues to alleviate geometric ambiguity, motivating multi-modal PCC Romanelli et al. (2026). Since ViPC Zhang et al. (2021) introduced image-guided completion, subsequent approaches have mainly improved image-point interaction through cross-modal attention, feature alignment, bidirectional fusion, and local modeling.

More recently, PGNet Luo et al. (2026) introduces the *Completion-by-Correction* paradigm, as illustrated in Fig. 1 (a). PGNet uses Trellis Xiang et al. (2025) to generate an explicit image-derived 3D prior as an intermediate geometric representation and then corrects it using the partial point cloud. Although this design provides a complete structural scaffold and substantially improves completion quality, representing visual information as an explicit 3D prior commits the available visual evidence to a single geometric hypothesis. When this intermediate hypothesis contains distorted structures or inaccurate local geometry, such errors may propagate through the subsequent correction process and bias the final completion, as illustrated in Fig. 1 (a). More fundamentally, the visual evidence required for completion depends on both the missing structures and the geometry already available in the partial observation, and may not need to be fully materialized as a complete 3D shape. This observation motivates a more fundamental question: ***Does visual information need to be converted into explicit 3D priors for point cloud completion?***

We argue that it does not. Instead, image representations extracted by a frozen vision foundation model can serve as visual priors. These representations contain rich semantic and structural cues, motivating point-conditioned retrieval according to the geometric needs of the partial observation.

Based on this principle, we introduce **VIPER**, a visual prior retrieval framework for multi-modal PCC. VIPER maps point tokens and patch-level DINOv3 Siméoni et al. (2026) visual tokens into a shared latent space, where each point token selectively retrieves its top-K relevant visual priors. The retrieved evidence is incorporated through adaptive gated injection, producing prior-enhanced point tokens. These enhanced tokens provide global cues for coarse shape generation and are reused through token-guided refinement to recover local geometry. In this way, VIPER treats DINOv3 image representations as queryable visual priors conditioned on partial geometry, providing an alternative to explicit 3D prior generation.

The contributions are summarized as follows: (1) ***Visual Prior Retrieval Perspective***. We rethink the role of visual information in multi-modal PCC and propose treating pretrained image representations as queryable visual priors and retrieving them based on partial-geometry context, providing an alternative to explicit image-derived 3D prior generation. (2) ***Novel Framework***. We propose VIPER, which combines shared-space projection, point-conditioned top-K visual prior retrieval, adaptive gated integration, and token-guided refinement for coarse-to-fine completion. (3) ***Joint-category Benchmark and Analysis***. We introduce OmniObject3D-Completion, a large-scale benchmark with substantially broader category coverage, enabling joint-category training and systematic evaluation of multi-modal PCC across diverse object types. (4) ***SOTA Performance***. VIPER achieves state-of-the-art performance on multi-modal PCC benchmarks with fewer trainable parameters and substantially lower end-to-end inference cost than the previous best-performing method.

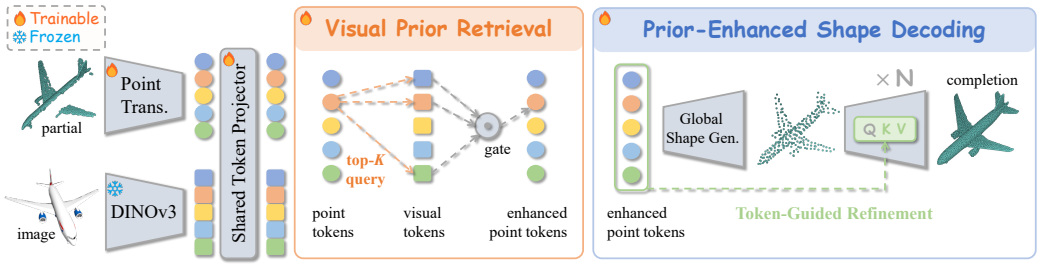


Figure 2: **Overview of VIPER.** A trainable point encoder and a frozen DINOv3 produce point and visual tokens, which are mapped into a shared latent space by a shared token projector. Each point token selectively retrieves top-K visual priors and injects the aggregated evidence through an adaptive channel-wise gate, yielding prior-enhanced point tokens. These tokens support global shape generation and are reused at each stage of token-guided refinement.

2 METHOD

2.1 OVERVIEW

As illustrated in Fig. 2, VIPER consists of cross-modal token embedding, visual prior retrieval, and prior-enhanced shape decoding. A trainable point encoder extracts tokens from the partial point cloud, while a frozen DINOv3 Siméoni et al. (2026) provides image tokens as visual priors. A shared projector maps both modalities into a common latent space. Point-conditioned top-K cross-attention retrieves visual evidence, and a channel-wise gate regulates its contribution to the point tokens. The enhanced point tokens are pooled for coarse shape generation and reused by each refinement stage, preserving access to token-level evidence throughout decoding.

2.2 VISUAL PRIOR RETRIEVAL

We implement visual prior retrieval as top-K cross-attention, using the partial geometry to select evidence from pretrained image representations. Let $\boldsymbol{P} \in \mathbb{R}^{N\times C_0}$ denote the point tokens extracted from the partial point cloud, and let $\boldsymbol{I} \in \mathbb{R}^{M\times C_0}$ denote the patch-level visual tokens extracted by the frozen image encoder. Here, N and M are the numbers of point and visual tokens, respectively. We employ a shared token projector $\Phi(\cdot)$ to map both modalities into a common C-dimensional latent space:

$$\boldsymbol{P}^0 = \Phi(\boldsymbol{P}), \qquad \boldsymbol{I}^0 = \Phi(\boldsymbol{I}), \tag{1}$$

where $\boldsymbol{P}^0 \in \mathbb{R}^{N\times C}$ and $\boldsymbol{I}^0 \in \mathbb{R}^{M\times C}$. Sharing the projection function provides a common transformation for subsequent cross-modal comparison, while preserving the distinct point and visual token sets.

We apply modality-specific normalization to the projected tokens and form point queries, visual keys, and visual values:

$$\boldsymbol{Q} = \mathrm{LN}_P(\boldsymbol{P}^0)\boldsymbol{W}_q, \qquad \boldsymbol{K} = \mathrm{LN}_I(\boldsymbol{I}^0)\boldsymbol{W}_k, \qquad \boldsymbol{V} = \mathrm{LN}_I(\boldsymbol{I}^0)\boldsymbol{W}_v, \tag{2}$$

where $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, and $\boldsymbol{W}_v$ are learnable linear projections. For each point query $\boldsymbol{q}_i$, we compute its cosine similarity with all visual keys and retain the indices of the K most relevant visual tokens:

$$\mathcal{S}_i = \mathrm{TopK}_j\left(\frac{\boldsymbol{q}_i^\top \boldsymbol{k}_j}{\|\boldsymbol{q}_i\|_2\|\boldsymbol{k}_j\|_2}, K\right), \tag{3}$$

where $\mathcal{S}_i$ contains the selected K visual-token indices for the i-th point token. This restricts aggregation to tokens ranked by their relevance to the partial-geometry queries, reducing the inclusion of potentially irrelevant visual evidence.

Cosine similarity selects the visual tokens, while scaled dot-product attention assigns their aggregation weights:

$$\alpha_{ij} = \mathrm{Softmax}_{j\in\mathcal{S}_i}\left(\frac{\boldsymbol{q}_i^\top \boldsymbol{k}_j}{\sqrt{d}}\right), \tag{4}$$

where d is the dimensionality of each query and key vector. The retrieved visual message for the i-th point token is then computed as

$$\boldsymbol{m}_i = \sum_{j \in \mathcal{S}_i} \alpha_{ij} \boldsymbol{v}_j. \tag{5}$$

Adaptive Gated Prior Injection. The retrieved visual message may be less informative for some point tokens because of background regions, viewpoint ambiguity, or cross-modal inconsistency. We therefore introduce an adaptive channel-wise gate to regulate the strength of visual-prior injection:

$$\boldsymbol{g}_i = \sigma\left(f_{\text{gate}}\left(\left[\boldsymbol{p}_i^0 \,\|\, \boldsymbol{m}_i\right]\right)\right), \tag{6}$$

where $\boldsymbol{g}_i \in [0,1]^C$, $f_{\text{gate}}(\cdot)$ is a learnable mapping, and $\sigma(\cdot)$ denotes the sigmoid function. Conditioned on both the point representation and the retrieved visual message, the gate modulates the contribution of visual evidence at a channel-wise level. The visual-prior-enhanced point token is then obtained through residual injection:

$$\tilde{\boldsymbol{p}}_i = \boldsymbol{p}_i^0 + \boldsymbol{g}_i \odot f_{\text{prior}}(\boldsymbol{m}_i), \tag{7}$$

where $f_{\text{prior}}(\cdot)$ projects the aggregated visual prior into the point-token space and $\odot$ denotes element-wise multiplication. A residual feed-forward refinement is subsequently applied to obtain the final enhanced point token $\bar{\boldsymbol{p}}_i$. The resulting token set $\bar{\boldsymbol{P}} \in \mathbb{R}^{N \times C}$ preserves the original point-token structure while incorporating visual information with adaptively controlled strength. Visual tokens serve as prior sources, and the enhanced point tokens carry the retrieved evidence into decoding.

2.3 Prior-Enhanced Shape Decoding

The enhanced point tokens $\bar{\boldsymbol{P}} \in \mathbb{R}^{N \times C}$ support decoding at two levels: their pooled representation initializes the coarse shape, while the full token set guides subsequent refinement. The tokens are computed once and reused throughout decoding, allowing each stage to access geometric and visual evidence beyond the pooled global representation.

Global Shape Generation. We first aggregate the prior-enhanced point tokens into a global shape representation using max pooling across tokens:

$$\boldsymbol{z} = \text{MaxPool}(\bar{\boldsymbol{P}}), \tag{8}$$

where $\boldsymbol{z} \in \mathbb{R}^C$ summarizes the global structure of the incomplete object enriched by visual priors. A lightweight seed generator then decodes $\boldsymbol{z}$ into an initial coarse point cloud:

$$\boldsymbol{X}^{(0)} = f_{\text{seed}}(\boldsymbol{z}) \in \mathbb{R}^{n_0 \times 3}, \tag{9}$$

where n_0 denotes the number of coarse points. The coarse point cloud provides a global geometric scaffold for subsequent detail recovery.

Token-Guided Refinement. At each refinement stage, the decoder attends to the same prior-enhanced point tokens $\bar{\boldsymbol{P}}$ using queries derived from its current geometric state. This reuses the previously integrated visual evidence without repeating retrieval from the image tokens. Let $\boldsymbol{X}^{(\ell)} \in \mathbb{R}^{n_\ell \times 3}$ denote the current point cloud. We construct decoder queries from point-wise coordinate features, global geometric context, and the global shape representation:

$$\boldsymbol{H}^{(\ell)} = f_{\text{query}}\left(\left[\phi(\boldsymbol{X}^{(\ell)}) \,\|\, \text{MaxPool}\left(\phi(\boldsymbol{X}^{(\ell)})\right) \,\|\, \boldsymbol{z}\right]\right), \tag{10}$$

where $\phi(\cdot)$ embeds coordinates and $f_{\text{query}}(\cdot)$ is a learnable projection; both global vectors are broadcast to all points before concatenation. Each query aggregates the enhanced point tokens:

$$\boldsymbol{A}^{(\ell)} = \text{Attn}\left(\boldsymbol{H}^{(\ell)}, \bar{\boldsymbol{P}}, \bar{\boldsymbol{P}}\right), \tag{11}$$

where $\text{Attn}(\cdot)$ denotes cross-attention. In contrast to the top-K visual prior retrieval stage, this operation attends to all enhanced point tokens. These tokens already contain selectively retrieved visual evidence, which remains available for geometric refinement.

An adaptive token-conditioning gate regulates the aggregated guidance for each decoder point:

$$\boldsymbol{\Gamma}^{(\ell)} = \sigma\left(f_{\text{cond}}\left(\left[\boldsymbol{H}^{(\ell)} \,\|\, \boldsymbol{A}^{(\ell)}\right]\right)\right), \tag{12}$$

$$\widetilde{\boldsymbol{H}}^{(\ell)} = \boldsymbol{H}^{(\ell)} + \boldsymbol{\Gamma}^{(\ell)} \odot f_{\text{guide}}\left(\boldsymbol{A}^{(\ell)}\right), \tag{13}$$

where $f_{\text{cond}}(\cdot)$ and $f_{\text{guide}}(\cdot)$ are learnable mappings. This mechanism modulates token-level guidance according to the current geometric state of each decoder point.

The token-conditioned features are then processed by a geometry-aware local aggregation operator:

$$\boldsymbol{F}^{(\ell)} = \mathcal{G}_{\text{geo}}\left(\boldsymbol{X}^{(\ell)}, \widetilde{\boldsymbol{H}}^{(\ell)}, \boldsymbol{F}^{(\ell-1)}\right), \tag{14}$$

where $\mathcal{G}_{\text{geo}}(\cdot)$ aggregates local neighborhood features according to the spatial relations among the current points. Here, $\boldsymbol{F}^{(\ell-1)}$ denotes the feature propagated from the preceding refinement stage; for the first stage, the token-conditioned feature is directly used as the initial geometric feature. A learned feature splitting operation predicts child-point features and corresponding coordinate offsets, yielding the next-stage point cloud:

$$\boldsymbol{X}^{(\ell+1)} = \text{Up}\left(\boldsymbol{X}^{(\ell)}\right) + f_{\Delta}\left(\boldsymbol{F}^{(\ell)}\right). \tag{15}$$

Here, Up expands the point coordinates, and f_{Δ} predicts matching child-point offsets. Starting from $\boldsymbol{X}^{(0)}$, we apply two refinement stages, each with an upsampling factor of 2. Token-guided refinement reuses the enhanced tokens beyond global shape initialization. All completion modules are trained jointly with DINOv3 frozen. Appendix B details the output configuration, loss, and training settings.

3 Experiments and Analyses

We organize our experiments and analyses around the following research questions: **RQ1:** How effective and generalizable is VIPER for multi-modal point cloud completion? **RQ2:** How does visual prior retrieval compare with explicit 3D prior generation? **RQ3:** Which design choices contribute to effective visual prior retrieval and utilization?

3.1 Datasets

ShapeNet-ViPC. We conduct the primary evaluation on ShapeNet-ViPC Zhang et al. (2021), a widely adopted benchmark for multi-modal point cloud completion. The dataset contains 38,328 CAD objects spanning 13 categories. Each object is rendered from 24 viewpoints, where every sample consists of an RGB image, a corresponding partial point cloud, and its complete point cloud annotation. We follow the standard eight-category data split and evaluation protocol used by recent methods Xu et al. (2024); Fang et al. (2025); Luo et al. (2026).

OmniObject3D-Completion. To study multi-modal completion under a broad category vocabulary, we construct OmniObject3D-Completion from OmniObject3D Wu et al. (2023). The benchmark contains 3,864 objects from **85 categories** and 24 views per object, yielding 92,736 multi-modal samples. Each sample consists of a view-aligned image and partial point cloud together with a complete point cloud target. We train a single model jointly across all 85 categories, moving beyond conventional category-wise training and enabling systematic evaluation across diverse object types.

KITTI. To further evaluate the generalization ability of our method in real-world scenarios, we conduct experiments on the KITTI dataset Geiger et al. (2012). Following Cross-PCC Wu et al. (2025), we adopt the processed benchmark containing 156 paired samples, each composed of a real LiDAR partial point cloud and the corresponding RGB image. Because KITTI does not provide complete point cloud annotations, we follow the protocol of IAET Fang et al. (2025): the model is trained on the car category of the 3D-EPN dataset Wu et al. (2025) and directly evaluated on KITTI.

3.2 EXPERIMENTAL SETTINGS

Evaluation Metrics. Following prior multi-modal point cloud completion methods Fang et al. (2025); Luo et al. (2026), we use Chamfer Distance (CD) and F-Score.

Baselines. We compare VIPER with representative multi-modal point cloud completion approaches on the ShapeNet-ViPC benchmark, including ViPC Zhang et al. (2021), CSDN Zhu et al. (2024), XMFnet Aiello et al. (2022), CDPNet Du et al. (2024a), EGIInet Xu et al. (2024), DuInNet Liu et al. (2025), SGPRNet Zhou et al. (2026), GAPM Xiu et al. (2025), IAET Fang et al. (2025), and the recent state-of-the-art method PGNet Luo et al. (2026).

Training Details. For a fair comparison on ShapeNet-ViPC, we adopt the category-wise training and optimization protocol of PGNet Luo et al. (2026). Each category is trained separately for 100,000 iterations using AdamW with an initial learning rate of 2×10^{-4} and a cosine annealing schedule. We use a batch size of 128 and conduct all experiments on a single NVIDIA H100 GPU. Training configurations for OmniObject3D-Completion and KITTI are provided in Appendix B.

Table 1: Comparison of CD ($\times 10^{-3}$; $\downarrow$) on ShapeNet-ViPC.
† denotes results directly reported in the corresponding papers.

Methods	Avg.	Airplane	Cabinet	Car	Chair	Lamp	Sofa	Table	Watercraft
ViPC† (CVPR'21)	3.308	1.760	4.558	3.183	2.476	2.867	4.481	4.990	2.197
CSDN† (TVCG'24)	2.570	1.251	3.670	2.977	2.835	2.554	3.240	2.575	1.742
XMFnet (NeurIPS'22)	1.443	0.572	1.980	1.754	1.403	1.810	1.702	1.386	0.945
CDPNet† (AAAI'24)	1.374	0.671	1.902	1.857	1.504	1.024	1.694	1.437	0.904
EGIInet (ECCV'24)	1.211	0.534	1.921	1.655	1.204	0.776	1.552	1.227	0.802
DuInNet† (TMM'25)	1.147	0.545	1.667	1.533	1.153	0.770	1.526	1.217	0.765
SGPRNet† (AAAI'26)	1.110	0.499	1.639	1.597	1.096	0.725	1.425	1.181	0.717
GAPM† (ICCV'25)	1.107	0.492	1.784	1.590	1.094	0.681	1.431	1.048	0.733
IAET (IJCAI'25)	1.090	0.503	1.397	1.648	1.196	0.668	1.401	1.200	0.704
PGNet (AAAI'26)	0.926	0.455	1.111	1.303	1.038	0.578	1.139	1.167	0.615
VIPER (Ours)	**0.774**	**0.398**	**0.955**	**0.918**	**0.931**	**0.527**	**1.026**	**0.966**	**0.471**

Table 2: Comparison of F-Score@0.001 ($\uparrow$) on ShapeNet-ViPC.
† denotes results directly reported in the corresponding papers.

Methods	Avg.	Airplane	Cabinet	Car	Chair	Lamp	Sofa	Table	Watercraft
ViPC† (CVPR'21)	0.591	0.803	0.451	0.512	0.529	0.706	0.434	0.594	0.730
CSDN† (TVCG'24)	0.695	0.862	0.548	0.560	0.669	0.761	0.557	0.729	0.782
XMFnet (NeurIPS'22)	0.796	0.961	0.662	0.691	0.809	0.792	0.723	0.830	0.901
CDPNet† (AAAI'24)	0.814	0.955	0.697	0.692	0.799	0.890	0.738	0.829	0.914
EGIInet (ECCV'24)	0.836	0.969	0.693	0.723	0.847	0.919	0.756	0.857	0.927
DuInNet† (TMM'25)	0.851	0.971	0.745	0.753	0.858	0.919	0.770	0.857	0.935
SGPRNet† (AAAI'26)	0.862	0.978	0.749	0.746	0.876	0.927	0.796	0.875	0.946
GAPM† (ICCV'25)	0.860	0.974	0.759	0.757	0.862	0.927	0.794	0.882	0.923
IAET (IJCAI'25)	0.860	0.980	0.782	0.725	0.850	0.934	0.795	0.867	0.950
PGNet (AAAI'26)	0.895	0.985	0.839	0.804	0.887	0.954	0.850	0.881	0.963
VIPER (Ours)	**0.922**	**0.994**	**0.881**	**0.878**	**0.904**	**0.955**	**0.877**	**0.905**	**0.982**

3.3 PERFORMANCE AND GENERALIZATION (RQ1)

To answer RQ1, we evaluate completion accuracy on ShapeNet-ViPC, joint-category performance on OmniObject3D-Completion, and qualitative transfer to sparse real-world observations on KITTI.

Evaluation on ShapeNet-ViPC. As shown in Tables 1 and 2, VIPER achieves state-of-the-art performance across all eight categories on ShapeNet-ViPC in terms of both CD and F-Score. Compared with the previous best-performing method PGNet Luo et al. (2026), VIPER reduces the average CD by **16.4%**, from 0.926 to 0.774, while improving the average F-Score from 0.895 to 0.922. The gains are particularly pronounced on geometrically challenging categories. For example, CD decreases from 1.303 to 0.918 on Car and from 0.615 to 0.471 on Watercraft, corresponding to relative

reductions of 29.5% and 23.4%, respectively. Consistent improvements are also observed on Cabinet, where CD decreases from 1.111 to 0.955 and F-Score increases from 0.839 to 0.881.

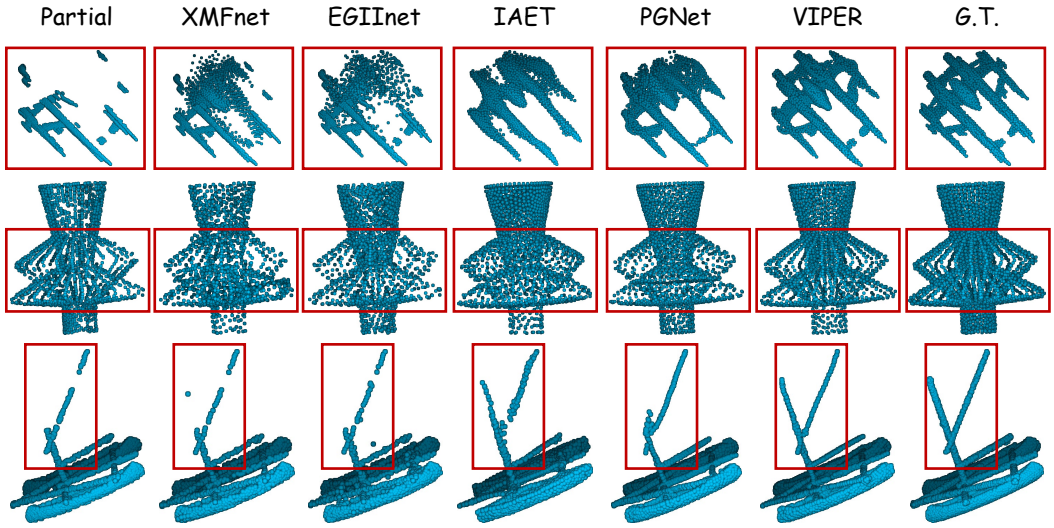


Figure 3: **Qualitative comparisons on ShapeNet-ViPC.** Red boxes highlight representative regions with severe incompleteness or challenging geometric structures.

The qualitative comparisons in Fig. 3 highlight structural continuity and local detail, complementing the quantitative gains. In the first example, VIPER better restores the connected frame while preserving the gaps between slender components, whereas several baselines leave fragmented parts or introduce irregular point clusters. In the second example, VIPER restores the surrounding supports and their connections to the central structure, not just the overall shape. In the third example, VIPER reconstructs both branches of the thin V-shaped structure with clearer continuity, whereas some methods omit one branch or leave disconnected segments. Overall, VIPER better matches the ground truth in connectivity and local geometry, consistent with its CD and F-Score gains. Further qualitative comparisons across diverse geometric structures are provided in Appendix C.2.

Evaluation on OmniObject3D-Completion. VIPER achieves the best overall CD and F-Score under joint training across all 85 categories. It obtains the lowest CD in 62 categories and the highest F-Score in 67 categories, indicating that its gains extend across diverse object types. Complete per-category and macro-averaged results are provided in Appendix C.1.

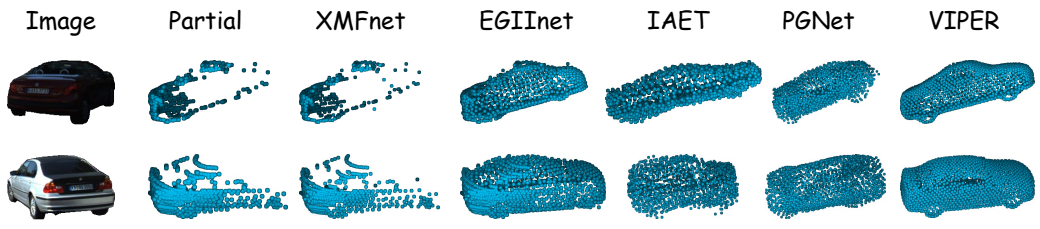


Figure 4: **Qualitative comparisons on KITTI under sparse real-world LiDAR observations.** VIPER produces more complete and structurally coherent vehicle shapes in the illustrated cases.

Evaluation on KITTI. We further evaluate VIPER on KITTI to examine its transfer to sparse real-world LiDAR observations. Since KITTI does not provide complete ground-truth point clouds, we focus on qualitative comparisons following the established protocol. As shown in Fig. 4, existing methods often produce incomplete vehicle shapes, scattered points, or irregular geometry under severe sparsity. VIPER instead generates more compact and structurally coherent profiles. These

comparisons suggest that the retrieved visual evidence remains useful under challenging real-world observations. Additional qualitative comparisons are provided in Appendix C.3.

3.4 Visual Prior Retrieval vs. Explicit 3D Prior Generation (RQ2)

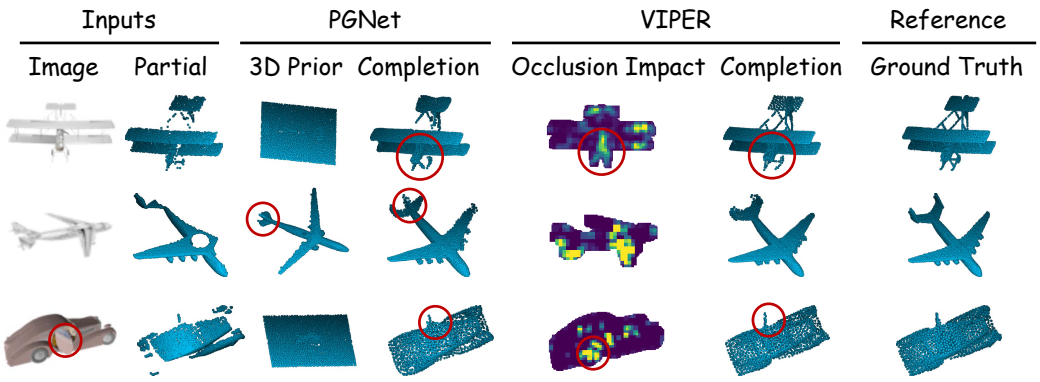


Figure 5: **Case studies comparing explicit 3D priors with visual prior retrieval.** Occlusion-impact maps highlight influential visual regions, while red circles mark geometric differences.

Input-Dependent Visual Evidence Analysis. Figure 5 compares PGNet's explicit image-derived 3D priors with VIPER's completions and post-hoc occlusion-impact maps. With the partial point cloud fixed, these maps measure the increase in completion error after masking local image regions. This connects visual evidence to the completion objective: larger error increases indicate greater sensitivity to the masked region, while small changes indicate limited influence under the perturbation. In the illustrated cases, inaccuracies in PGNet's intermediate priors correspond to artifacts in its completions. VIPER produces more faithful geometry, and the highlighted aircraft nose and door regions indicate sensitivity to object structures relevant to completion. These cases complement the quantitative comparisons by relating visual input sensitivity to recovered geometry.

The stage-wise comparison on ShapeNet-ViPC in Appendix D.1 further shows that VIPER achieves lower CD and higher F-Score than PGNet from the coarse prediction through both refinement stages.

Table 3: End-to-end inference cost comparison between PGNet and VIPER.

Method	Trainable Params. (M)	Frozen Params. (M)	Inference Latency (ms)	GPU Memory (GB)
PGNet (AAAI'26)	61.25	1200	3847	10.13
VIPER (Ours)	53.77	21.61	30.7	0.42

Inference Pipeline Analysis. For a fair comparison, we evaluate the end-to-end inference efficiency of PGNet and VIPER under the same hardware setting. As shown in Table 3, PGNet incurs an inference latency of 3847 ms and uses 10.13 GB of GPU memory per sample, whereas VIPER requires only 30.7 ms and 0.42 GB. VIPER therefore achieves an approximately 125$\times$ inference speedup while reducing GPU memory usage by a factor of 24. In this comparison, VIPER directly retrieves visual evidence from the frozen encoder's representations without generating an intermediate 3D prior, achieving substantially lower inference cost.

3.5 Analysis of Visual Prior Retrieval and Utilization (RQ3)

We study how visual evidence is selected, integrated, and reused on the Cabinet category. Table 4 reports test results; the retrieval curves in Appendix E.2 trace test-set CD across training checkpoints.

Visual Prior Selection. Replacing the shared projector with independent 1024-dimensional projectors increases test CD from 0.955 to 1.022, supporting the shared transformation for cross-modal retrieval. In the w/o top-K Selection variant, we replace similarity-based top-K selection with random selection while keeping $K = 32$. This increases test CD from 0.955 to 1.089 and reduces

F-Score from 0.881 to 0.856 (Table 4). With the retrieval size fixed, this comparison supports the importance of selecting relevant evidence rather than merely increasing the number of visual tokens. The test-set curves also show slower convergence (Fig. 10).

Gated Integration and Token-Guided Refinement. Removing the image modality increases test CD from 0.955 to 1.341 and reduces F-Score from 0.881 to 0.801, supporting the contribution of visual evidence. Removing the injection gate increases CD to 1.024, while removing decoder token guidance increases it to 1.095. These results support adaptive integration of retrieved visual evidence. During decoding, reusing the enhanced point tokens gives each refinement stage direct access to these representations as the point cloud is updated. The performance drop without decoder token guidance supports the benefit of this reuse beyond the tokens' contribution to coarse shape initialization. Additional component analyses are provided in Appendices E.1–E.4.

Table 4: Ablation studies of visual evidence selection and utilization on the Cabinet test set in terms of CD (10^{-3}, $\downarrow$) and F-Score@0.001 ($\uparrow$).

Variant	CD $\downarrow$	F-Score $\uparrow$
w/o Image Modality	1.341	0.801
w/o top-K Selection	1.089	0.856
w/o Adaptive Injection Gate	1.024	0.868
w/o Decoder Token Guidance	1.095	0.857
VIPER	**0.955**	**0.881**

4 Related Work

Point Cloud Completion. Single-modal methods use geometric observations and learned shape priors (Yuan et al., 2018; Fang et al., 2024), with Transformer-based approaches further improving long-range geometric modeling (Yu et al., 2021; 2023; Li et al., 2023; Zhu et al., 2023; Wang et al., 2024; Zhu et al., 2025). Recent single-modal approaches further exploit diffusion or 3D generative priors through image-space representations derived from the input point cloud (Wei et al., 2025; Li et al., 2025a). Multi-modal methods instead introduce additional visual or semantic observations: ViPC (Zhang et al., 2021) pioneered image-guided completion, followed by improved interaction and alignment in XMFnet (Aiello et al., 2022), EGIInet (Xu et al., 2024), and IAET (Fang et al., 2025). More recent multi-modal approaches exploit stronger priors, including geometric and textual priors (Xiu et al., 2025; Zhou et al., 2026) and PGNet's Completion-by-Correction framework, which generates an explicit image-derived 3D prior as an intermediate representation (Luo et al., 2026). Extended discussion is provided in Appendix A.

Vision Foundation Models. Vision foundation models such as CLIP (Radford et al., 2021), SAM (Kirillov et al., 2023), and DINO (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2026) provide transferable semantic and dense visual representations. Prior work has connected such representations to point cloud analysis and 3D perception (Huang et al., 2023; Wang et al., 2025). VIPER treats frozen DINOv3 patch representations as queryable visual priors, combining point-conditioned top-K cross-attention with gated integration and token-guided refinement.

5 Conclusion

We presented **VIPER**, which uses point-conditioned visual prior retrieval to adaptively enhance point representations and reuses the enhanced tokens throughout progressive coarse-to-fine decoding. On ShapeNet-ViPC, VIPER achieves 16.4% lower average CD and 125$\times$ faster end-to-end inference than PGNet. Our **OmniObject3D-Completion** benchmark supports joint-category evaluation across 85 object categories, on which VIPER achieves the best aggregate results among the compared methods. Qualitative results on KITTI further suggest its ability to transfer to sparse real-world LiDAR observations. These findings support visual prior retrieval and token-guided refinement as an effective way to incorporate pretrained visual representations into multi-modal point cloud completion. Limitations and future directions are discussed in Appendix F.

REFERENCES

Emanuele Aiello, Diego Valsesia, and Enrico Magli. Cross-modal learning for image-guided point cloud shape completion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 37349–37362, New Orleans, USA, November 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, Montreal, Canada, October 2021.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(12):10164–10183, December 2024.

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 (CVPR)*, pp. 13581–13590, Vancouver, Canada, June 2023.

Zhenjiang Du, Jiale Dou, Zhitao Liu, Jiwei Wei, Guan Wang, Ning Xie, and Yang Yang. CDPNet: Cross-modal dual phases network for point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 1635–1643, Vancouver, Canada, February 2024a.

Zijin Du, Jianqing Liang, Jiye Liang, Kaixuan Yao, and Feilong Cao. Graph regulation network for point cloud segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(12):7940–7955, December 2024b.

Fan Duan, Jiahao Yu, and Li Chen. T-CorresNet: Template guided 3D point cloud completion with correspondence pooling query generation strategy. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 99–106, Milan, Italy, September 2024.

Chenghao Fang, Bing Yang, Hailiang Ye, and Feilong Cao. Fast point completion network. *Neural Computing and Applications (NCA)*, 36:10897–10913, June 2024.

Chenghao Fang, Jianqing Liang, Jiye Liang, Hangkun Wang, Kaixuan Yao, and Feilong Cao. Multimodal point cloud completion with interleaved attention enhanced Transformer. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 963–971, Montreal, Canada, August 2025.

Chenghao Fang, Jianqing Liang, Jiye Liang, Zijin Du, and Feilong Cao. Point cloud semantic scene completion with prototype-guided Transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 3822–3830, Singapore, January 2026.

Ben Fei, Weidong Yang, Wen-Ming Chen, Zhijun Li, Yikang Li, Tao Ma, Xing Hu, and Lipeng Ma. Comprehensive review of deep learning-based 3D point cloud completion processing and analysis. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 23(12):22862–22883, December 2022.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361, Providence, USA, June 2012.

Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3D point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(12):4338–4364, December 2021.

Wenzhe He, Xiaojun Chen, Wentang Chen, Hongyu Wang, Ying Liu, and Ruihui Li. RWKV-PCSSC: Exploring RWKV model for point cloud semantic scene completion. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pp. 161–170, Dublin, Ireland, October 2025.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 4672–4680, Singapore, January 2026.

Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson W.H. Lau, Wanli Ouyang, and Wangmeng Zuo. CLIP2Point: Transfer CLIP to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22157–22167, Paris, France, October 2023.

Shun Iwase, Muhammad Zubair Irshad, Katherine Liu, Vitor Guizilini, Robert Lee, Takuya Ikeda, Ayako Amma, Koichi Nishiwaki, Kris Kitani, Rares Ambrus, and Sergey Zakharov. ZeroGrasp: Zero-shot shape reconstruction enabled robotic grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17405–17415, Nashville, USA, June 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, Paris, France, October 2023.

An Li, Zhe Zhu, and Mingqiang Wei. GenPC: Zero-shot point cloud completion via 3D generative priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1308–1318, Nashville, USA, June 2025a.

Aocheng Li, James R. Zimmer-Dauphinee, Rajesh Kalyanam, Ian Lindsay, Parker VanValkenburgh, Steven Wernke, and Daniel Aliaga. Self-supervised large scale point cloud completion for archaeological site restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11759–11768, Nashville, USA, June 2025b.

Shanshan Li, Pan Gao, Xiaoyang Tan, and Mingqiang Wei. ProxyFormer: Proxy alignment assisted point cloud completion with missing part sensitive Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9466–9475, Vancouver, Canada, June 2023.

Hanzhe Liang, Guoyang Xie, Chengbin Hou, Bingshu Wang, Can Gao, and Jinbao Wang. Look inside for more: Internal spatial modality perception for 3D anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 5146–5154, Philadelphia, USA, February 2025a.

Hanzhe Liang, Jie Zhang, Tao Dai, Linlin Shen, Jinbao Wang, and Can Gao. Taming anomalies with down-up sampling networks: Group center preserving reconstruction for 3D anomaly detection. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pp. 7133–7141, Dublin, Ireland, October 2025b.

Jiye Liang, Zijin Du, Jianqing Liang, Kaixuan Yao, and Feilong Cao. Long and short-range dependency graph structure learning framework on point cloud. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(12):14975–14989, December 2023.

Xinpu Liu, Baolin Hou, Hanyun Wang, Ke Xu, Jianwei Wan, and Yulan Guo. DuInNet: Dual-modality feature interaction for point cloud completion. *IEEE Transactions on Multimedia (TMM)*, 27:8525–8540, September 2025.

Wang Luo, Di Wu, Hengyuan Na, Yinlin Zhu, Miao Hu, and Guocong Quan. Rethinking multi-modal point cloud completion: A completion-by-correction perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 7680–7688, Singapore, January 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, January 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pp. 8748–8763, Virtual, July 2021.

Alessio Romanelli, Michaela Servi, Roberto Magherini, Monica Carfagni, and Yary Volpe. Survey of latest advancements in deep learning for point cloud completion. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 32(7):7321–7340, July 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel HAZIZA, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research (TMLR)*, April 2026.

Keneni W. Tesema, Lyndon Hill, Mark W. Jones, Muneeb I. Ahmad, and Gary K.L. Tam. Point cloud completion: A survey. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 30 (10):6880–6899, October 2024.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, Nashville, USA, June 2025.

Jun Wang, Ying Cui, Dongyan Guo, Junxia Li, Qingshan Liu, and Chunhua Shen. PointAttN: You only need attention for point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 5472–5480, Vancouver, Canada, February 2024.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27243–27253, Nashville, USA, June 2025.

Lintai Wu, Qijian Zhang, Junhui Hou, and Yong Xu. Leveraging single-view images for unsupervised 3D point cloud completion. *IEEE Transactions on Multimedia (TMM)*, 27:940–953, February 2025.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. OmniObject3D: Large-vocabulary 3D object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 803–814, Vancouver, Canada, June 2023.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21469–21480, Nashville, USA, June 2025.

Jingqiao Xiu, Yicong Li, Na Zhao, Han Fang, and Xiang Wang. Geometric alignment and prior modulation for view-guided point cloud completion on unseen categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27435–27444, Honolulu, USA, October 2025.

Hang Xu, Chen Long, Wenxiao Zhang, Yuan Liu, Zhen Cao, Zhen Dong, and Bisheng Yang. Explicitly guided information interaction network for cross-modal point cloud completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 414–432, Milan, Italy, September 2024.

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 ACM International Conference on Multimedia (ACM MM)*, pp. 8952–8961, Melbourne, Australia, October 2024a.

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 (ICCV)*, pp. 12478–12487, Montreal, Canada, October 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 (TPAMI)*, 45(12):14114–14130, December 2023.

Xumin Yu, Yanbo Wang, Jie Zhou, and Jiwen Lu. ProtoComp: Diverse point cloud completion with controllable prototype. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 270–286, Milan, Italy, September 2024b.

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

Wenxiao Zhang, Huajian Zhou, Zhen Dong, Jun Liu, Qingan Yan, and Chunxia Xiao. Point cloud completion via skeleton-detail Transformer. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 29(10):4229–4242, October 2023.

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 (CVPR)*, pp. 15890–15899, Virtual, June 2021.

Zhensheng Zhou, Jianqing Liang, Jiye Liang, Zijin Du, and Chenghao Fang. Semantic guided part relation-aware network for point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 13871–13879, Singapore, January 2026.

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 (ICCV)*, pp. 14508–14518, Paris, France, October 2023.

Zhe Zhu, Liangliang Nan, Haoran Xie, Honghua Chen, Jun Wang, Mingqiang Wei, and Jing Qin. CSDN: Cross-modal shape-transfer dual-refinement network for point cloud completion. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 30(7):3545–3563, July 2024.

Zhe Zhu, Honghua Chen, Xing He, and Mingqiang Wei. PointSea: Point cloud completion via self-structure augmentation. *International Journal of Computer Vision (IJCV)*, 133:4770–4794, March 2025.

Appendix

A Extended Related Work

A.1 Single-modal Point Cloud Completion

Single-modal point cloud completion Yuan et al. (2018); Fang et al. (2024); Fei et al. (2022) reconstructs complete 3D shapes from partial observations using only geometric information. Early methods typically adopt encoder–decoder architectures to learn global representations and generate complete point clouds, but often lose fine-grained details due to limited representation capacity. Recently, Transformer-based methods Yu et al. (2021); Zhang et al. (2023); Yu et al. (2023); Chen et al. (2023); Li et al. (2023); Zhu et al. (2023); Yu et al. (2024a); Duan et al. (2024); Yu et al. (2024b); Wang et al. (2024); Zhu et al. (2025) have advanced this field by leveraging self-attention to model long-range dependencies and global structural relationships. PCDreamer Wei et al. (2025) and GenPC Li et al. (2025a) further exploit diffusion or 3D generative priors through image-space representations derived from the input point cloud. Because these representations originate from the geometric input rather than independent visual observations, we categorize both approaches as single-modal PCC. Despite these advances, single-modal methods remain constrained by incomplete geometric observations, making it difficult to recover ambiguous or severely missing regions without complementary semantic and structural guidance from independent visual observations.

A.2 Multi-modal Point Cloud Completion

Multi-modal point cloud completion introduces RGB images to provide complementary semantic and structural cues for recovering missing geometry. ViPC Zhang et al. (2021) pioneers this direction by using single-view images to supplement global shape information. Subsequent methods mainly improve cross-modal interaction and reconstruction strategies: XMFnet Aiello et al. (2022), EGIInet Xu et al. (2024), DuInNet Liu et al. (2025), and IAET Fang et al. (2025) enhance image-point fusion through cross-attention, explicit alignment, dual-modality interaction, or interleaved attention, while CSDN Zhu et al. (2024) and CDPNet Du et al. (2024a) exploit coarse-to-fine or patch-wise modeling. More recent multi-modal methods introduce stronger priors: GAPM Xiu et al. (2025) exploits 3D lifting and local geometric priors, SGPRNet Zhou et al. (2026) uses text-based relational priors, and PGNet Luo et al. (2026) introduces the Completion-by-Correction framework, which generates an explicit 3D shape prior and corrects it using the partial observation.

B Implementation Details

Output Configuration. On ShapeNet-ViPC, OmniObject3D-Completion, and KITTI, VIPER uses the same output configuration: a 512-point coarse prediction followed by two refinement stages, each with an upsampling factor of 2, yielding 1024 and 2048 points, respectively.

Loss. We supervise all three VIPER prediction stages using the Chamfer Distance as the training objective. Given a predicted point set $\boldsymbol{X}$ and the complete ground-truth set $\boldsymbol{Y}$, the CD is defined as

$$\mathcal{L}_{\mathrm{CD}}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{|\boldsymbol{X}|} \sum_{\boldsymbol{x} \in \boldsymbol{X}} \min_{\boldsymbol{y} \in \boldsymbol{Y}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{1}{|\boldsymbol{Y}|} \sum_{\boldsymbol{y} \in \boldsymbol{Y}} \min_{\boldsymbol{x} \in \boldsymbol{X}} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \tag{16}$$

We apply multi-stage supervision by matching each prediction $\boldsymbol{X}^{(0)}$, $\boldsymbol{X}^{(1)}$, and $\boldsymbol{X}^{(2)}$ to the complete ground-truth point set. The overall training objective is

$$\mathcal{L} = \frac{1}{3} \sum_{\ell=0}^{2} \mathcal{L}_{\mathrm{CD}} \left(\boldsymbol{X}^{(\ell)}, \boldsymbol{Y} \right). \tag{17}$$

This multi-stage supervision encourages both the coarse scaffold and the subsequent refinement stages to progressively approach the target geometry.

OmniObject3D-Completion. VIPER, XMFnet Aiello et al. (2022), EGIInet Xu et al. (2024), IAET Fang et al. (2025), and PGNet Luo et al. (2026) are each trained jointly across all 85 categories for 10,000 iterations, with the same split and 2,048-point outputs. RGB inputs are resized to 224×224; partial and complete point clouds each contain 2,048 points. A fixed view-based split assigns approximately 80% of each object's 24 views to training and the remainder to testing, yielding 74,189 training samples and 18,547 test samples. All methods retain their original architectures and training objectives. We use AdamW with an initial learning rate of 2×10^{-4}, a weight decay of 0.01, and a global batch size of 128. The learning rate follows a linear warmup followed by cosine decay to 1×10^{-5}. All experiments are conducted on a single NVIDIA H100.

KITTI. Since KITTI does not provide complete point cloud annotations, we follow the evaluation protocol of IAET Fang et al. (2025), training VIPER on the car category of the 3D-EPN dataset released by Wu et al. (2025) and directly evaluating it on KITTI. We adopt the same 100-epoch training budget as IAET while retaining the optimization settings used by VIPER. Specifically, the model is trained with AdamW using a batch size of 128, an initial learning rate of 2×10^{-4}, and a weight decay of 0.01. The learning rate is linearly warmed up for the first 5 epochs and then decayed to 1×10^{-5} following a cosine schedule. Each partial point cloud contains 2,048 points and is paired with an RGB image resized to 224×224 for image-conditioned point cloud completion.

Table 5: Comparison of CD ($\times 10^{-4}$; $\downarrow$) on OmniObject3D-Completion.
The left and right blocks continue the same alphabetical category list.

Category	VIPER	PGNet	XMFnet	EGIInet	IAET	Category	VIPER	PGNet	XMFnet	EGIInet	IAET
Apple	4.969	**4.769**	5.940	8.866	6.206	Mushroom	**3.141**	3.386	4.656	5.741	4.848
Bamboo shoots	**1.684**	1.909	2.443	2.899	2.583	Onion	**3.512**	3.994	4.619	5.512	4.471
Banana	**1.847**	1.926	2.571	2.489	2.436	Orange	4.545	**4.445**	5.555	7.712	6.325
Biscuit	**2.331**	2.333	3.083	3.227	2.851	Peach	**4.212**	4.662	5.599	8.114	5.667
Box	**2.814**	3.213	3.903	4.052	3.391	Peanut	**2.158**	2.244	3.125	3.281	2.865
Bread	**3.959**	4.059	5.244	6.550	5.803	Pear	**3.982**	4.209	5.132	6.063	4.929
Bun	**4.200**	4.382	5.244	6.023	5.282	Pie	3.521	**3.403**	4.350	4.735	4.640
Candy	**2.205**	2.467	3.050	3.198	3.069	Pillow	**2.495**	2.668	3.421	3.328	3.347
Carrot	**1.368**	1.553	1.788	2.103	1.726	Pitaya	4.286	**4.262**	5.742	6.869	6.247
Cheese	3.417	**3.258**	4.676	5.410	5.281	Pizza	2.893	**2.770**	3.776	3.993	4.051
Chili	**1.962**	2.158	3.099	4.066	3.065	Pomegranate	**3.464**	3.745	4.237	4.774	4.229
Chocolate	**2.949**	3.247	3.908	5.160	3.997	Popcorn	6.143	**5.243**	9.057	10.046	10.585
Clock	**3.712**	4.157	5.749	6.009	5.765	Potato	3.391	**3.367**	4.304	5.098	4.710
Coconut	**4.178**	4.686	5.598	6.302	5.520	Power strip	**1.902**	2.192	2.743	3.091	2.682
Corn	**1.763**	1.898	2.016	2.179	1.967	Pumpkin	**4.164**	4.770	5.725	8.495	5.713
Cucumber	**1.636**	1.818	2.173	2.603	2.096	Remote control	**1.526**	1.651	1.984	2.044	1.884
Cup	**4.440**	5.477	7.256	8.751	6.663	Rubik cube	**5.287**	6.715	8.163	8.562	7.053
Dinosaur	2.020	**1.682**	3.202	3.048	3.081	Sausage	**1.530**	1.645	1.957	2.106	1.865
Doll	**2.662**	2.674	4.430	4.653	4.491	Shampoo	**1.685**	1.948	2.319	2.490	2.019
Donut	3.999	**3.935**	5.753	6.607	6.098	Shoe	3.329	**2.878**	4.623	4.875	4.585
Durian	4.829	**4.549**	5.899	6.936	6.187	Shrimp	2.233	**2.229**	2.872	2.898	3.201
Egg	**2.751**	3.122	3.622	4.157	3.425	Soap	4.428	**3.572**	5.317	5.570	5.807
Eraser	**2.525**	2.834	3.287	3.459	3.038	Squash	**2.200**	2.433	3.104	3.636	3.043
Facial cream	**2.103**	2.496	2.920	3.283	2.554	Steamed bun	**4.190**	4.390	5.800	6.324	5.889
Fig	**2.532**	3.004	3.412	3.791	3.200	Sushi	**3.987**	4.130	5.620	6.179	5.589
Garage kit	3.476	**3.388**	5.918	6.001	5.938	Sweet potato	**1.151**	1.215	1.517	1.659	1.560
Ginger	2.878	**2.494**	4.508	4.616	5.081	Teapot	**2.906**	3.162	4.878	5.197	5.062
Glasses case	**1.467**	1.626	1.984	2.036	1.702	Tissue	**3.149**	3.572	4.264	4.350	3.747
Hair dryer	**1.526**	1.804	2.293	2.139	2.005	Tomato	**3.966**	4.284	5.145	6.214	5.377
Hamburger	**4.499**	4.597	6.497	7.350	6.913	Tooth paste	**1.051**	1.114	1.435	1.479	1.316
Hand cream	**1.350**	1.500	1.858	1.920	1.812	Toy animals	2.884	**2.849**	5.165	5.124	5.493
Handbag	**3.929**	4.303	5.622	6.302	5.516	Toy boat	2.323	**2.161**	3.487	3.298	3.105
Hat	**3.686**	4.121	6.083	6.745	6.456	Toy bus	**2.515**	2.681	3.461	3.572	3.068
Haw thorn	**3.500**	3.931	4.344	4.901	4.262	Toy motorcycle	3.039	**2.918**	4.938	4.944	4.779
Hot dog	2.615	**2.558**	3.540	3.742	3.522	Toy plane	1.838	**1.781**	2.868	2.930	2.451
House	**4.938**	5.452	8.547	8.673	9.158	Toy train	**3.130**	3.153	4.580	4.558	4.060
Lemon	**2.913**	3.150	3.613	3.922	3.252	Toy truck	3.883	**3.806**	6.178	6.258	5.824
Light	**2.994**	3.359	4.496	4.917	4.401	Waffle	4.882	**4.623**	5.953	6.054	6.177
Lotus root	**2.831**	3.009	3.940	4.409	3.891	Wallet	**2.411**	2.707	3.097	3.265	3.000
Mango	**2.227**	2.433	2.800	3.230	2.711	Walnut	**3.941**	4.096	5.289	6.454	5.065
Mangosteen	**3.820**	4.109	5.169	5.287	4.707	Watermelon	**3.045**	3.199	4.274	4.846	4.398
Medicine bottle	**2.485**	2.781	3.370	3.464	2.989	Yam	**1.789**	2.097	2.282	2.573	2.248
Mouse	**1.995**	2.166	2.632	2.836	2.564						
Overall	**3.054**	3.199	4.276	4.775	4.272						
Macro avg.	**3.036**	3.185	4.236	4.737	4.231						

C Additional Performance and Generalization Results (RQ1)

C.1 Evaluation on OmniObject3D-Completion

OmniObject3D-Completion is designed to further investigate the performance of multi-modal point cloud completion methods under a broad joint-category setting, where a single model is trained across all 85 categories. Tables 5 and 6 present the complete per-category comparisons in terms of CD and F-Score, respectively. Overall scores average all test samples, whereas macro averages give equal weight to the 85 category means, regardless of the number of samples in each category.

VIPER achieves the best overall CD and F-Score, as well as the best macro-averaged results, demonstrating consistent advantages at both the dataset and category levels. Specifically, VIPER obtains an overall CD of 3.054 and F-Score of 0.470, together with a macro-averaged CD of 3.036 and F-Score of 0.479. Moreover, it ranks first in 62 of the 85 categories for CD and in 67 categories for F-Score. These category-wise results show that VIPER's gains are not confined to a small subset of object types, but extend broadly across the diverse category vocabulary of OmniObject3D-Completion.

Table 6: Comparison of F-Score@0.01 (↑) on OmniObject3D-Completion.
The left and right blocks continue the same alphabetical category list.

Category	VIPER	PGNet	XMFnet	EGIInet	IAET	Category	VIPER	PGNet	XMFnet	EGIInet	IAET
Apple	**0.254**	0.229	0.250	0.237	0.207	Mushroom	**0.455**	0.432	0.404	0.398	0.361
Bamboo shoots	**0.746**	0.708	0.680	0.659	0.672	Onion	**0.322**	0.286	0.296	0.284	0.256
Banana	**0.753**	0.737	0.643	0.666	0.692	Orange	**0.281**	0.252	0.267	0.256	0.233
Biscuit	**0.554**	0.547	0.488	0.485	0.486	Peach	0.274	0.260	**0.275**	0.266	0.222
Box	**0.438**	0.390	0.367	0.367	0.366	Peanut	**0.573**	0.565	0.492	0.493	0.481
Bread	**0.343**	0.338	0.337	0.334	0.277	Pear	**0.284**	0.262	0.274	0.263	0.229
Bun	0.266	0.260	**0.278**	0.269	0.217	Pie	0.366	**0.372**	0.369	0.366	0.291
Candy	**0.608**	0.571	0.556	0.549	0.536	Pillow	**0.515**	0.482	0.452	0.460	0.422
Carrot	**0.805**	0.781	0.726	0.700	0.749	Pitaya	0.295	0.290	**0.300**	0.294	0.222
Cheese	**0.562**	0.545	0.497	0.495	0.481	Pizza	0.496	**0.502**	0.468	0.468	0.407
Chili	**0.677**	0.662	0.594	0.574	0.573	Pomegranate	**0.352**	0.315	0.322	0.313	0.283
Chocolate	**0.461**	0.430	0.414	0.397	0.381	Popcorn	0.178	0.205	**0.220**	0.219	0.120
Clock	**0.330**	0.299	0.286	0.283	0.227	Potato	**0.390**	0.386	0.379	0.370	0.327
Coconut	**0.248**	0.219	0.246	0.237	0.191	Power strip	**0.708**	0.683	0.605	0.596	0.642
Corn	**0.695**	0.656	0.625	0.612	0.649	Pumpkin	**0.303**	0.290	0.293	0.283	0.241
Cucumber	**0.750**	0.731	0.673	0.662	0.686	Remote control	**0.766**	0.744	0.655	0.653	0.683
Cup	**0.277**	0.223	0.234	0.230	0.192	Rubik cube	**0.228**	0.163	0.183	0.176	0.146
Dinosaur	0.727	**0.758**	0.641	0.650	0.605	Sausage	**0.759**	0.741	0.679	0.670	0.709
Doll	**0.496**	0.490	0.431	0.430	0.365	Shampoo	**0.727**	0.663	0.606	0.598	0.654
Donut	0.321	**0.324**	0.305	0.299	0.246	Shoe	0.460	**0.466**	0.412	0.404	0.367
Durian	0.281	0.301	**0.311**	0.307	0.227	Shrimp	0.647	**0.652**	0.590	0.596	0.550
Egg	**0.420**	0.371	0.361	0.346	0.359	Soap	**0.388**	0.363	0.357	0.350	0.332
Eraser	**0.536**	0.500	0.459	0.462	0.488	Squash	**0.631**	0.585	0.557	0.538	0.557
Facial cream	**0.655**	0.600	0.553	0.545	0.580	Steamed bun	**0.277**	0.256	0.275	0.267	0.214
Fig	**0.494**	0.405	0.416	0.401	0.402	Sushi	**0.376**	0.368	0.366	0.367	0.320
Garage kit	**0.375**	0.374	0.346	0.347	0.262	Sweet potato	**0.873**	0.856	0.767	0.749	0.776
Ginger	0.526	**0.553**	0.478	0.479	0.409	Teapot	**0.440**	0.409	0.369	0.367	0.305
Glasses case	**0.757**	0.711	0.625	0.616	0.704	Tissue	**0.355**	0.308	0.311	0.307	0.294
Hair dryer	**0.755**	0.681	0.619	0.637	0.677	Tomato	**0.277**	0.258	0.274	0.263	0.212
Hamburger	**0.259**	0.253	0.258	0.254	0.189	Tooth paste	**0.900**	0.880	0.787	0.785	0.847
Hand cream	**0.822**	0.786	0.707	0.705	0.727	Toy animals	**0.488**	0.486	0.424	0.431	0.344
Handbag	**0.312**	0.278	0.292	0.286	0.228	Toy boat	0.585	**0.594**	0.507	0.512	0.490
Hat	**0.343**	0.310	0.303	0.300	0.261	Toy bus	**0.484**	0.451	0.405	0.402	0.408
Haw thorn	**0.331**	0.294	0.314	0.305	0.269	Toy motorcycle	0.423	**0.443**	0.374	0.373	0.305
Hot dog	0.521	**0.527**	0.474	0.471	0.435	Toy plane	0.683	**0.695**	0.572	0.570	0.581
House	**0.241**	0.203	0.217	0.217	0.148	Toy train	**0.432**	0.422	0.383	0.384	0.355
Lemon	**0.386**	0.348	0.346	0.334	0.343	Toy truck	**0.340**	0.326	0.305	0.306	0.248
Light	**0.439**	0.398	0.381	0.378	0.342	Waffle	0.242	0.250	**0.271**	0.270	0.181
Lotus root	**0.475**	0.449	0.432	0.421	0.394	Wallet	**0.525**	0.491	0.466	0.460	0.442
Mango	**0.535**	0.503	0.479	0.462	0.471	Walnut	0.283	0.271	**0.287**	0.280	0.232
Mangosteen	**0.324**	0.278	0.286	0.284	0.258	Watermelon	**0.408**	0.382	0.374	0.362	0.321
Medicine bottle	**0.503**	0.445	0.423	0.416	0.431	Yam	**0.684**	0.645	0.620	0.606	0.625
Mouse	**0.606**	0.571	0.527	0.515	0.513						
Overall	**0.470**	0.448	0.421	0.416	0.393						
Macro avg.	**0.479**	0.456	0.429	0.424	0.402						

C.2 Additional Qualitative Analysis on ShapeNet-ViPC

Figure 6 provides additional qualitative comparisons on ShapeNet-ViPC Zhang et al. (2021) across diverse geometric structures. Overall, the examples are consistent with VIPER's quantitative advantage in recovering both global structure and fine-grained geometry.

The first three cases emphasize the recovery of thin, articulated, or weakly observed structures. XMFnet and EGIInet frequently produce fragmented components or scattered points, while IAET and PGNet improve global completeness but may still exhibit incomplete or geometrically inconsistent local structures. In comparison, VIPER more faithfully reconstructs missing supports and structural connections, yielding predictions closer to the ground-truth geometry.

The remaining cases cover different geometric ambiguities. In the fourth case, VIPER better preserves the closed contour and its local continuity, whereas competing methods exhibit incomplete or distorted structures. For objects involving narrow stems, articulated supports, or complex local components, VIPER produces cleaner geometry with fewer isolated points and better preserves connectivity and spatial organization. The final case shows that VIPER recovers richer internal structures from severely incomplete observations rather than merely completing the global outer shape. Together with the quantitative results in Tables 1 and 2, these comparisons support the consistent advantage of VIPER in recovering challenging missing geometry across diverse object structures.

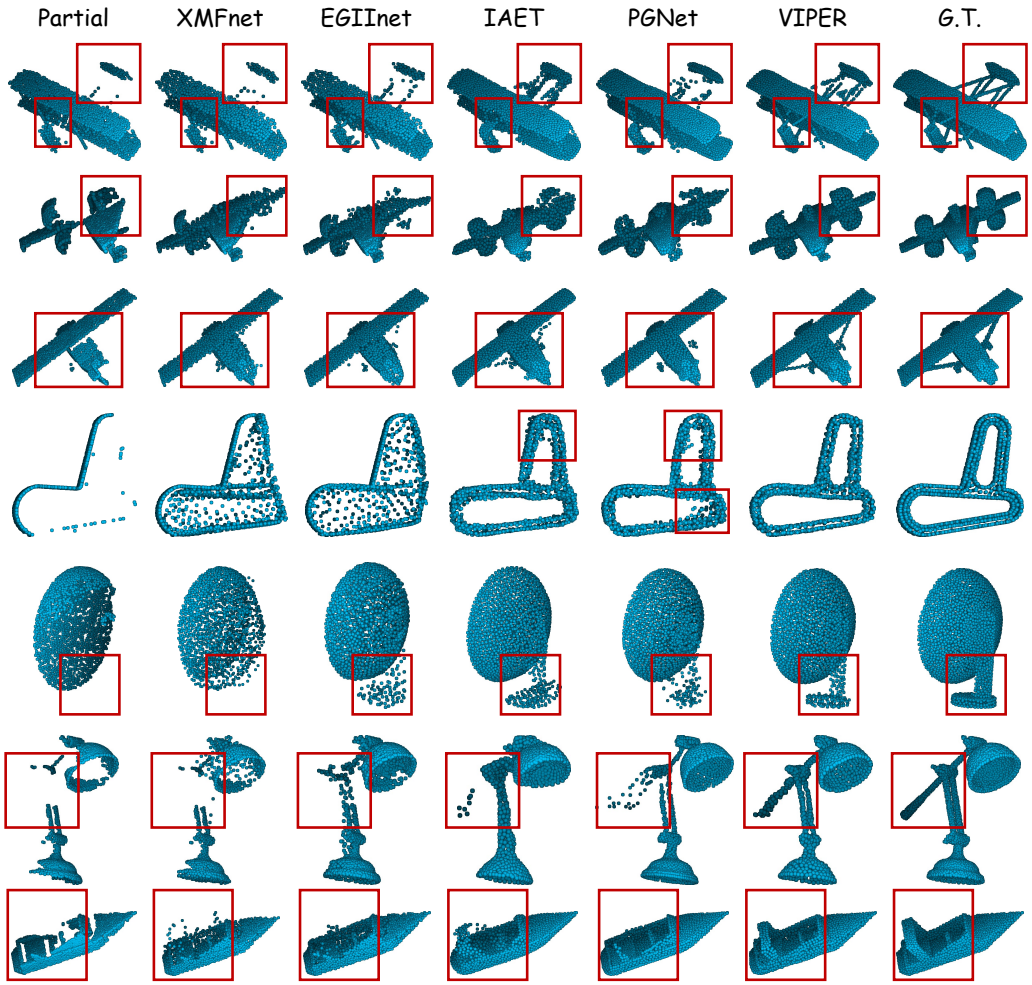


Figure 6: **Additional qualitative comparisons on ShapeNet-ViPC.** Red boxes highlight representative regions with severe incompleteness or challenging local geometry.

C.3 Additional Qualitative Analysis on KITTI

Figure 7 provides additional qualitative comparisons on KITTI under diverse vehicle appearances and sparse LiDAR observations. Across these cases, XMFnet often preserves substantial incompleteness from the partial input and fails to recover a coherent vehicle structure. EGIInet and IAET generally reconstruct a more complete vehicle extent, but may introduce irregular surfaces, scattered points, or locally distorted geometry. PGNet further improves overall completeness, although elongated body profiles and locally inconsistent structures can still be observed in several cases. These comparisons highlight the difficulty of recovering plausible vehicle geometry from highly sparse real-world observations, even when complementary image information is available.

VIPER exhibits more stable geometric recovery across these cases. In the first two examples, it reconstructs compact and continuous vehicle bodies despite substantial missing observations, while preserving clearer overall profiles with fewer isolated artifacts. In the third case, several competing methods produce irregular vehicle surfaces, while VIPER recovers a cleaner and more coherent profile. The last two cases involve different vehicle instances and observation patterns, yet VIPER consistently maintains plausible body proportions, structural continuity, and cleaner local geometry. Although complete ground truth is unavailable on KITTI, these qualitative results suggest that VIPER generates structurally coherent completions across diverse real-world cases when conditioned on paired images and partial point clouds at inference time.

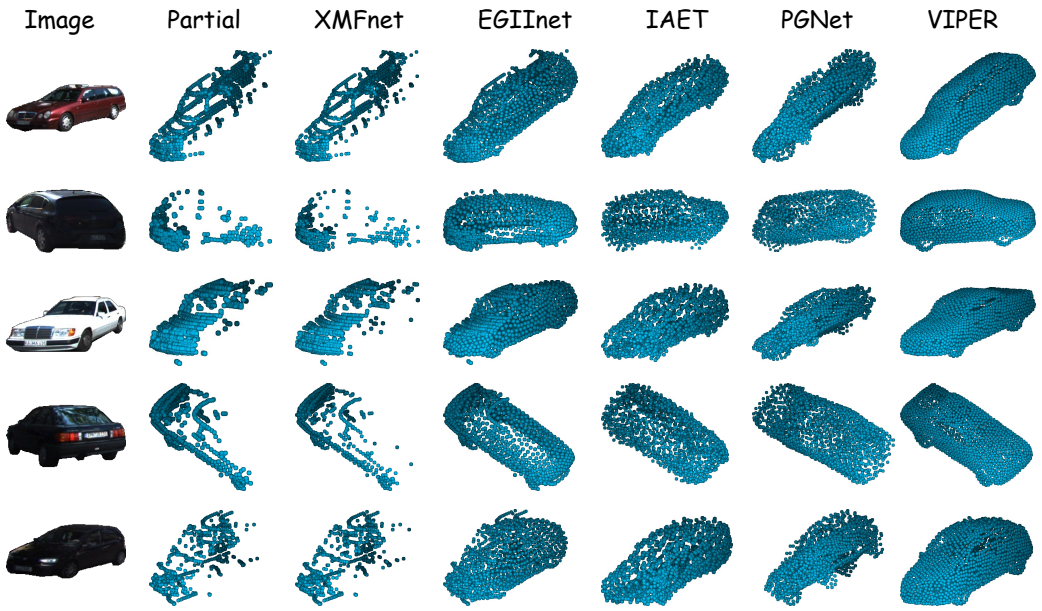


Figure 7: **Additional qualitative comparisons on KITTI.** The examples cover diverse vehicle appearances, viewpoints, and missing patterns in real-world LiDAR observations.

D Further Analysis of Visual Prior Retrieval vs. Explicit 3D Prior Generation (RQ2)

D.1 Stage-wise Recovery Dynamics

To examine where the performance difference between PGNet and VIPER emerges, we compare their predictions at 512, 1024, and 2048 points. Both methods progressively refine a 512-point coarse scaffold into the final 2048-point completion. Matching the output resolutions allows stage-wise comparison without confounding differences in output point counts. Table 7 reports the macro-averaged results over all eight ShapeNet-ViPC categories, while Fig. 8 visualizes the progression.

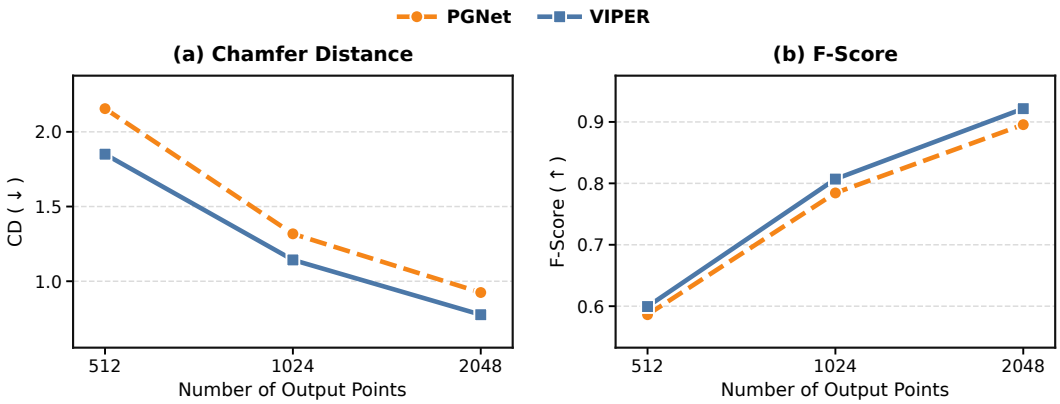


Figure 8: **Stage-wise recovery dynamics of PGNet and VIPER on ShapeNet-ViPC.** VIPER consistently achieves lower CD and higher F-Score throughout the coarse-to-fine completion process.

VIPER consistently outperforms PGNet throughout all decoding stages. At 512 points, VIPER reduces CD from 2.155 to 1.850 and improves F-Score from 0.586 to 0.599. This 14.2% relative CD reduction provides a more accurate starting point for refinement, establishing an advantage before either upsampling stage. At 1024 points, the advantage is preserved, with CD decreasing from 1.317 to 1.143 and F-Score increasing from 0.784 to 0.807. At 2048 points, VIPER achieves a CD of 0.774 versus 0.926 for PGNet, a 16.4% relative reduction, while the F-Score gap reaches 0.027.

Table 7: Stage-wise recovery dynamics of PGNet and VIPER on ShapeNet-ViPC in terms of CD (10^{-3}, ↓) and F-Score@0.001 (↑).

	CD ↓			F-Score ↑		
Method	512	1024	2048	512	1024	2048
PGNet	2.155	1.317	0.926	0.586	0.784	0.895
VIPER	**1.850**	**1.143**	**0.774**	**0.599**	**0.807**	**0.922**

The refinement dynamics provide further insight into how this advantage is maintained. During the first refinement from 512 to 1024 points, PGNet and VIPER obtain comparable relative CD improvements of 38.9% and 38.2%, respectively. During the subsequent 1024-to-2048 refinement, VIPER achieves a larger CD reduction of 32.3%, compared with 29.7% for PGNet. Meanwhile, the F-Score gap increases from 0.013 at 512 points to 0.023 at 1024 points and 0.027 at 2048 points. These results show that VIPER combines a stronger coarse prediction with continued improvements during refinement, even though the absolute CD gap narrows as both methods become more accurate.

This stage-wise pattern is consistent with VIPER's use of enhanced point tokens for both global shape generation and token-guided refinement. The comparison characterizes the complete pipelines rather than isolating individual components; the decoder-guidance ablation in Table 4 provides complementary evidence that reusing these tokens remains beneficial beyond coarse shape initialization.

E Further Analysis of Visual Prior Retrieval and Utilization (RQ3)

E.1 Analysis of the Shared Token Projector and Latent Dimension

Figure 9 compares projector designs and latent dimensions on Cabinet. Convergence curves throughout this section retrospectively trace test-set CD across checkpoints after training.

Replacing the shared projector with two independent 1024-dimensional projectors increases test CD from 0.955 to 1.022 despite slightly more trainable parameters, supporting a shared transformation for cross-modal retrieval between point tokens and pretrained visual representations.

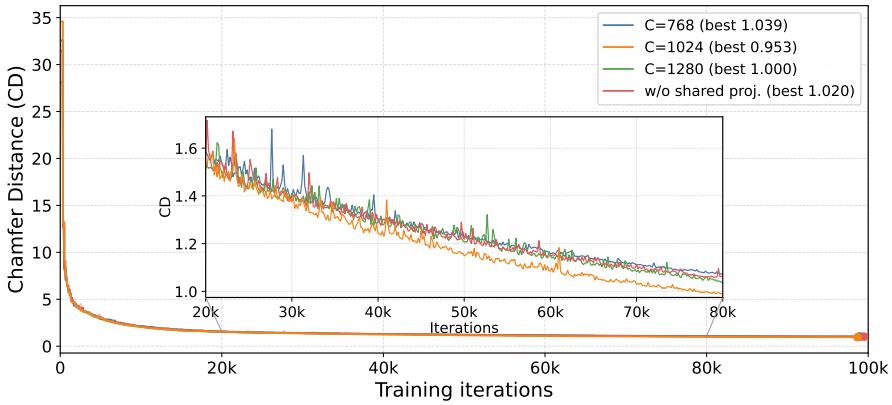


Figure 9: **Analysis of the shared token projector and latent feature dimension on the Cabinet category.** Curves show test-set CD; legends report the best observed values.

Latent dimensions of 768, 1024, and 1280 yield test CD values of 1.040, 0.955, and 1.002, respectively. We therefore use 1024 dimensions among these settings.

E.2 Analysis of Selective Visual Prior Retrieval

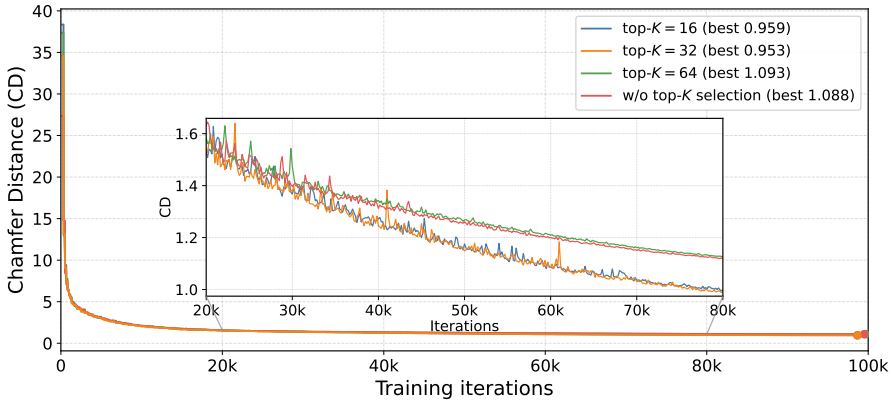


Figure 10: **Analysis of visual prior retrieval on the Cabinet category.** The test-set CD curves compare different retrieval sizes and the w/o top-K Selection variant.

Figure 10 compares retrieval sizes and selection strategies on Cabinet. Increasing K from 16 to 32 slightly improves the best observed test CD from 0.959 to 0.953, whereas $K = 64$ yields 1.093 and slower convergence. Retrieving more visual tokens therefore does not necessarily improve completion. The w/o top-K Selection variant replaces similarity-based selection with random selection at the same $K = 32$, increasing the best observed test CD to 1.088 and slowing convergence. Together, these comparisons support both an appropriate retrieval size and relevance-based selection; we use top-K retrieval with $K = 32$ as the default setting for subsequent visual evidence aggregation.

E.3 Gated Integration and Token-Guided Refinement

Table 4 reports the Cabinet test results. Removing images increases CD from 0.955 to 1.341 and reduces F-Score from 0.881 to 0.801, supporting complementary visual evidence. Removing the injection gate raises CD to 1.024 (+7.2%) and lowers F-Score to 0.868, supporting adaptive integration. Without decoder token guidance, CD rises to 1.095 (+14.7%) and F-Score falls to 0.857, supporting token reuse beyond coarse shape initialization.

E.4 Effect of Vision Foundation Model Capacity

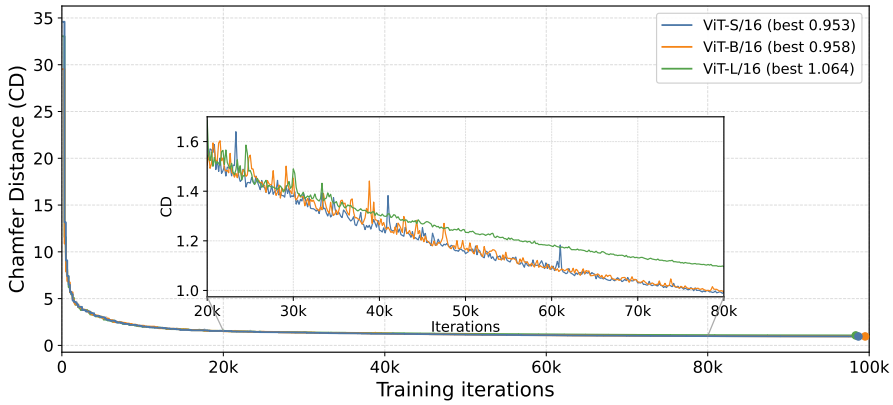


Figure 11: **Effect of DINOv3 backbone capacity on the Cabinet category.** CD curves show similar convergence for ViT-S/16 and ViT-B/16, while ViT-L/16 does not improve performance.

Table 8: Effect of vision foundation model capacity on the Cabinet test set in terms of CD (10^{-3}, ↓) and F-Score@0.001 (↑). The variants compare three frozen DINOv3 backbones within VIPER.

Vision Backbone	Frozen Params. (M)	CD ↓	F-Score ↑
ViT-S/16	21.61	**0.955**	**0.881**
ViT-B/16	85.70	0.958	0.880
ViT-L/16	303.23	1.067	0.863

We compare frozen DINOv3 ViT-S/16, ViT-B/16, and ViT-L/16 on Cabinet while retaining VIPER's retrieval and decoding framework. Table 8 shows that greater backbone capacity does not consistently improve completion. ViT-S/16 achieves a CD of 0.955 and an F-Score of 0.881; the larger ViT-B/16 performs comparably at 0.958 and 0.880. ViT-L/16 instead yields a higher CD of 1.067 and a lower F-Score of 0.863 in this Cabinet completion experiment.

The curves in Fig. 11 show similar convergence for ViT-S/16 and ViT-B/16, while ViT-L/16 retains a higher CD at later checkpoints. Among the evaluated backbones on Cabinet, ViT-S/16 achieves the best completion scores with the fewest frozen parameters, supporting its use as a compact default backbone for extracting the visual representations queried by VIPER.

F Limitations and Future Work

VIPER assumes informative paired RGB images and uses a fixed retrieval size. Robustness to weak cross-modal correspondence, severe image occlusion, appearance-domain shifts, and uninformative images has not been systematically evaluated. Future work could explore confidence-aware or variable-size retrieval adapted to partial geometry and image quality, extend visual prior retrieval to 3D reconstruction, shape understanding, and multi-modal 3D generation, and evaluate broader families of vision and multi-modal foundation models.

G AI use statement

Large language models were used to assist with translation, language editing, manuscript organization, and LaTeX formatting. They were not used for method implementation, dataset generation or preprocessing, or mathematical derivations or proofs. The authors reviewed and revised the AI-assisted content and retain full responsibility for the accuracy and integrity of all text, scientific claims, and experimental findings presented in the final manuscript.