Rethinking Multi-Modal Point Cloud Completion: Query-Aware Gating Attention and Gramian Volume Alignment

Jiye Liang , Haoyu Wang , Chenghao Fang , Kaixuan Yao*
Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University
ljy@sxu.edu.cn, 15534423891@163.com, chenghaofang1014@163.com, ykx@sxu.edu.cn

Abstract

Multi-modal point cloud completion aims to recover complete 3D geometric structures from partial observations by integrating auxiliary data. Although image-guided techniques are well-established, the potential of natural language as a source of high-level semantic cues remains underexplored. Therefore, effectively introducing natural language and synergizing it with the image modality to assist 3D structure recovery is key to breaking current performance bottlenecks. To address this, we propose **Q**uery-Aware **G**ating Attention and **G**ramian Volume **A**lignment (**QGGA**), a framework for multi-modal point cloud completion. QGGA employs a novel query-dependent bidirectional attention mechanism to dynamically filter information flows between modalities and adaptively regulate fusion intensity. To facilitate alignment among the three heterogeneous modalities, we utilize the volume of the parallelotope spanned by multi-modal local features as a consistency metric, constraining geometric alignment by minimizing this volume. Finally, the fused and aligned features are fed into a coarse-to-fine decoder to first predict a coarse point cloud and then progressively recover the complete shape through stepwise upsampling and refinement. Furthermore, we construct ShapeNet-ViPC-Desc, a benchmark dataset enriched with detailed text descriptions. Experimental results demonstrate that QGGA achieves new state-of-the-art (SOTA) performance on this benchmark. The source code is freely accessible at https://github.com/whysq/QGGA.

1 Introduction

Point clouds serve as a fundamental representation in 3D vision , widely applied in autonomous driving [Geiger *et al.*, 2012; Huang *et al.*, 2020; Guo *et al.*, 2026b], robotic manipulation, and digital twinning [Liang *et al.*, 2023; Song *et al.*, 2017]. However, real-world scans are inevitably constrained by self-occlusion [Du *et al.*, 2024b; Liang *et al.*, 2025b; Liang *et al.*, 2026], inter-object occlusion, and sensor resolution. Consequently, observed point clouds are often sparse and incomplete [Tesema *et al.*, 2024], making them insufficient for downstream tasks and necessitating robust algorithms to recover complete geometric structures from partial observations [Liang *et al.*, 2025a]. Early single-modal completion methods relied solely on partial inputs for inference. However, they struggle to resolve topological ambiguities when information is severely deficient, often resulting in structural discontinuities or blurred details [Tchapmi *et al.*, 2019; Huang *et al.*, 2020; Xie *et al.*, 2020; Li *et al.*, 2023; Yu *et al.*, 2021; Guo *et al.*, 2024; Guo *et al.*, 2026a].

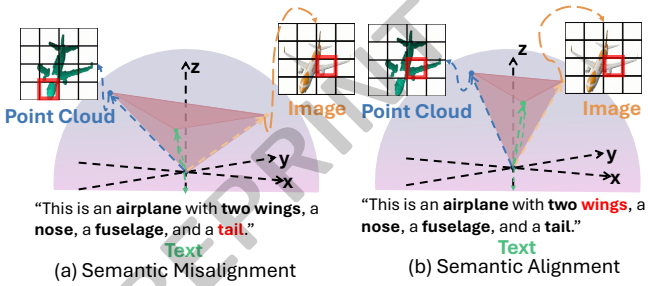

Figure 1: (a) Semantic misalignment across different parts yields a large volume. (b) Semantic alignment of all modalities on the same part results in a small volume.

To address the limitations of single-modal inference, recent research has shifted toward multi-modal point cloud completion, leveraging single-view RGB images to provide complementary appearance cues and object-level semantic priors. ViPC pioneered sequential fusion based on image reconstruction priors [Zhang *et al.*, 2021]. XMFnet successfully fused visual texture and geometric structure using cross-modal attention mechanisms, significantly improving completion accuracy [Aiello *et al.*, 2022]. EGIINet further optimized this process via explicit guided information interaction [Xu *et al.*, 2024]. Building upon previous cross-modal fusion strategies, IAET recently introduced an interleaved attention mechanism designed to significantly enhance the bidirectional information flow between modalities [Fang *et al.*, 2025].

Despite the promising prospects of multi-modal fusion, two key challenges remain unresolved. First, the potential of natural language remains largely underexplored. Unlike

*Corresponding author

images, which provide dense pixel-level texture, language offers abstract, high-level semantic descriptions that are invariant to viewpoint changes and lighting conditions [Zhou *et al.*, 2026]. Ignoring text limits the model's ability to capture the global semantic essence of objects. Second, since point clouds contain only partial geometric information while images and text depict the complete object form, directly fusing the three may lead to unstable matching in regions where geometric information is missing. Traditional Cross-Attention mechanisms force every Query feature to find a corresponding Key in the auxiliary modality and assign probability weights [Vaswani *et al.*, 2017]. When geometric features correspond to occluded or missing regions, this "Forced Alignment" causes the network to erroneously attend to irrelevant areas, reducing the effectiveness and controllability of cross-modal information [Fang *et al.*, 2026; Fang *et al.*, 2024; He *et al.*, 2026].

To tackle these challenges, we propose **Q**uery-Aware **G**ating Attention and **G**ramian Volume **A**lignment (**QGGA**), explicitly modeling multi-modal point cloud completion as a problem involving both "Adaptive Modulation" and "Geometric Alignment." For active selection, we introduce a Query-Dependent Gate within the interleaved cross-modal interaction. This mechanism uses query-driven gating scores to modulate attention outputs, dynamically controlling the injection intensity of multi-modal information. When specific visual or textual cues lack constructiveness for the current partial geometry, the gate suppresses their influence, preventing noise propagation caused by invalid alignment.For geometric alignment, we introduce an alignment metric based on Gramian Volume. Consistency is characterized by the volume of the parallelotope spanned by multi-modal feature vectors; this volume is calculated via the determinant of the Gram matrix, where a smaller volume indicates better alignment. As shown in Figure 1, we impose a Gramian Volume alignment constraint at the local level by minimizing the Gramian volume of triplets formed by local features of point clouds, images, and text. Furthermore, we calculate the Gramian consistency of tri-modal features at the global level to derive a confidence score for scaling the fusion gate. When overall tri-modal consistency is low, the system automatically reduces the intensity of cross-modal injection, making the completion process more robust. Additionally, we utilize a Large Language Model to automatically generate structured text descriptions for ShapeNet-ViPC images, introducing ShapeNet-ViPC-Desc. This enables QGGA to operate as a tri-modal system. In summary, our contributions are as follows:

- We introduce natural language descriptions into multi-modal point cloud completion, enabling the model to simultaneously leverage visual appearance cues and high-level textual semantics to assist in the geometric recovery of partial point clouds.
- We propose a fusion and alignment framework for tri-modal completion. For fusion, we employ interleaved cross-modal interaction with a query-dependent gating mechanism to make information injection more controllable. For alignment, we introduce a local consistency constraint based on Gramian Volume, combined with global consistency confidence to regulate fusion intensity, explicitly enhancing the stability of cross-modal interaction.
- We construct the ShapeNet-ViPC-Desc dataset and conduct systematic experiments on this benchmark. Extensive experiments demonstrate that our method achieves state-of-the-art (**SOTA**) performance.

2 Related Work

2.1 Single-modal Point Cloud Completion

Early data-driven methods primarily relied on 3D self-priors to recover geometric structures from incomplete inputs. PCN [Yuan *et al.*, 2018] pioneered a folding-based decoder but struggled to recover fine local details. To address this, SnowflakeNet [Xiang *et al.*, 2023] introduced Snowflake Point Deconvolution, progressively revealing geometric details through a coarse-to-fine generation process. Entering the Transformer era, PoinTr [Yu *et al.*, 2021] and its adaptive variant AdaPoinTr [Yu *et al.*, 2023] modeled the completion task as a set-to-set translation problem, utilizing geometry-aware attention mechanisms to capture long-range dependencies. Recently, SVDFormer [Zhu *et al.*, 2023] leveraged Self-View Decomposition to understand global shape structure, while SymmCompletion [Yan *et al.*, 2025] explicitly incorporated symmetry priors to guide high-fidelity recovery. However, single-modal methods are fundamentally limited by semantic ambiguity when facing severe occlusion, lacking external guidance to determine the true topology of missing regions [Xiu *et al.*, 2025].

2.2 Multi-modal Point Cloud Completion

To eliminate geometric ambiguity, researchers introduced auxiliary image modalities to provide semantic guidance. ViPC [Zhang *et al.*, 2021] established the view-guided completion paradigm but relied solely on simple global feature concatenation. Subsequent works like XMFnet [Aiello *et al.*, 2022] and CSDN [Zhu *et al.*, 2024] significantly improved performance by projecting modalities into a shared latent space for cross-attention fusion and style transfer. EGIINet [Xu *et al.*, 2024] and CDPNet [Du *et al.*, 2024a] further refined this interaction through explicit guidance and dual-stage mechanisms. Recent advances focus on alignment robustness: IAET [Fang *et al.*, 2025] employs an interleaved attention mechanism to enable bidirectional information flow. Recently, GRAM [Cicchetti *et al.*, 2025] proposed a novel alignment metric that enforces synchronous geometric consistency across n modalities in high-dimensional space by minimizing the Gramian volume of the feature parallelotope. In this work, we apply this theoretical insight to 3D completion by proposing QGGA. This method aims to enforce structural consistency among linguistic, visual, and geometric modalities, thereby effectively bridging the "partial-complete" gap.

3 Method

3.1 Overview

The proposed QGGA framework reconstructs complete and high-fidelity point clouds P_{out} from partial observations P_{in}

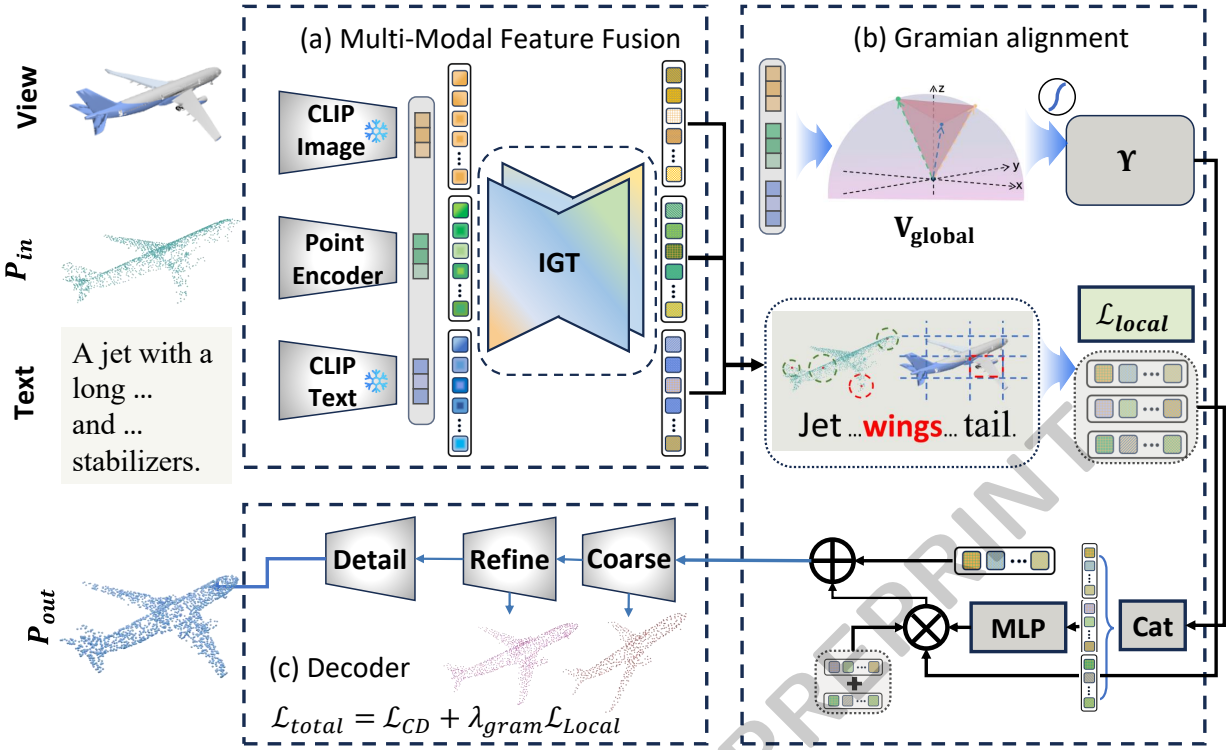

Figure 2: Overview of the proposed QGGA framework. The architecture consists of three main components: (a) Multi-Modal Feature Fusion, which extracts features using frozen CLIP encoders and a Point Encoder, integrating them via the Interleaved Gated Transformer (IGT) for deep cross-modal interaction; (b) Gramian Alignment, which enforces geometric consistency at two levels: a global branch computes Gram weights to dynamically modulate the feature flow, while a local branch minimizes the Local Gram Loss to align fine-grained geometric details; and (c) Decoder, which reconstructs the complete high-fidelity point cloud from the aligned features using a coarse-to-fine strategy.

by leveraging single-view images and text descriptions as complementary auxiliary modalities. As illustrated in Figure 2, the framework consists of three stages. First, in the multi-modal fusion stage, PointNet++ extracts point tokens F_P [Qi *et al.*, 2017], while frozen CLIP encoders produce image tokens F_I and text tokens F_T [Radford *et al.*, 2021]. These tokens are progressively fused by the Interleaved Gated Transformer (IGT), which performs controlled cross-modal interaction through Query-Dependent Attention (QDA). Second, in the Gramian alignment stage, we impose tri-modal consistency during training by minimizing the Gramian volume induced by aligned local feature groups. Third, a coarse-to-fine decoder progressively reconstructs dense complete geometry from the fused point features.

3.2 Multi-Modal Feature Fusion

In multi-modal point cloud completion, the input point cloud is partial, whereas the image and text modalities describe the object in a more complete manner. Direct fusion or standard cross-attention fusion may result in insufficient integration of useful information from various modalities. To address this, we propose the Interleaved Gated Transformer (IGT).

Query Dependent Attention. The core computational unit of IGT is Query Dependent Attention (QDA). Given a query matrix Q, a key matrix K, and a value matrix V, the standard attention is calculated as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

We introduce a learnable gating mechanism based on the query Q at the output of the attention module, as shown in Figure 3. The gating coefficient g is derived via a linear pro-

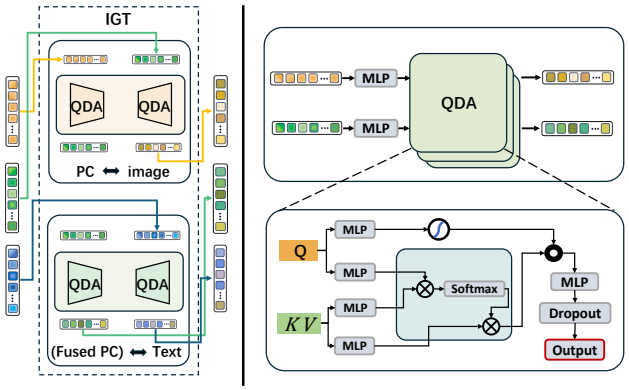

Figure 3: Architecture of the IGT module.

jection and a Sigmoid activation function

$$g(Q) = \sigma(QW_{gate} + b_{gate}) \tag{2}$$

where W_{gate} and b_{gate} are learnable parameters. The final output of QDA is the element-wise product of the standard attention output and the gating coefficient

$$Y_{gated} = \text{Attention}(Q, K, V) \odot g(Q) \tag{3}$$

This gating operation Y_{gated} explicitly controls the amount of cross-modal information allowed for each query token, thereby effectively suppressing the injection of erroneous information and enhancing the fusion robustness of multi-modal completion.

Interleaved Gated Transformer. To fully exploit visual texture and textual semantics, we construct the fusion structure in an interleaved manner, as shown in Figure 3. Since both images and point clouds possess spatial structures, their interaction first establishes a robust geometric skeleton. Abstract textual semantics, which lack spatial coordinates, are subsequently injected to refine the visually enhanced features. First, we set $Q = F_I$ and $K = V = F_P$ to compute the updated image features F'_I

$$F'_I = \text{QDA}(Q = F_I, K = F_P, V = F_P) \tag{4}$$

Subsequently, the point cloud queries the image to inject reliable visual information into the geometric features. We set $Q = F_P$ and $K = V = F'_I$ to compute the updated point cloud features F'_P

$$F'_P = \text{QDA}(Q = F_P, K = F'_I, V = F'_I) \tag{5}$$

Next, we perform similar operations to interact the visually enhanced point cloud tokens F'_P with the text tokens F_T to integrate high-level semantics. We set $Q = F_T$ and $K = V = F'_P$ to compute the updated text features F'_T

$$F'_T = \text{QDA}(Q = F_T, K = F'_P, V = F'_P) \tag{6}$$

Finally, we set $Q = F'_P$ and $K = V = F'_T$ to compute the fully updated point cloud features F''_P

$$F''_P = \text{QDA}(Q = F'_P, K = F'_T, V = F'_T) \tag{7}$$

The final outputs of IGT are the fused point cloud features F''_P, image features F'_I, and text features F'_T.

3.3 Gramian Alignment

Recent advances in multi-modal representation learning demonstrate a strong correlation between geometric alignment in high-dimensional space and semantic consistency. We introduce a novel local-global Gramian mechanism for point cloud completion. Unlike the original GRAM which focuses solely on global representation alignment, we extend this theory to fine-grained local feature alignment and global confidence gating, ensuring the completion process is guided by geometrically consistent multi-modal cues. As demonstrated by [Cicchetti *et al.*, 2025], the degree of alignment among a set of k modality vectors $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \in \mathbb{R}^d$ can be quantified by the volume of the k-dimensional parallelotope they span. This volume is calculated via the determinant of their Gram matrix $\mathbf{G}$, where $\mathbf{G}_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle$

$$Vol(\mathcal{V}) = \sqrt{\det(\mathbf{G}(\mathbf{v}_1, \ldots, \mathbf{v}_k))} \tag{8}$$

We apply L_2 normalization to all feature vectors before constructing the Gram matrix to prevent trivial solutions where feature magnitudes collapse to zero. A smaller volume indicates stronger semantic correlation and alignment among the modality vectors. We leverage this geometric property to enforce consistency among the partial point cloud, the reference image, and the text description.

Soft Alignment and Top-k Filtering. Since there is naturally no one-to-one correspondence between point cloud tokens, image tokens, and text tokens, we must first establish a shared feature space where every local point cloud token possesses corresponding visual and textual representations before performing geometric alignment. To address this, we propose an Attention-Soft Alignment strategy. As shown in Figure 4, let $\mathbf{F''_P} = \{\mathbf{f}_1^P, \ldots, \mathbf{f}_{N_P}^P\} \in \mathbb{R}^{N_P \times d}$, $\mathbf{F'_I} = \{\mathbf{f}_1^I, \ldots, \mathbf{f}_{N_I}^I\} \in \mathbb{R}^{N_I \times d}$, and $\mathbf{F'_T} = \{\mathbf{f}_1^T, \ldots, \mathbf{f}_{N_T}^T\} \in \mathbb{R}^{N_T \times d}$ denote the fused features of the point cloud, image, and text, respectively. We employ two parallel SoftAlign modules to aggregate auxiliary information into the point cloud domain. Specifically, we use the point cloud features $\mathbf{F''_P}$ as the Query (Q) and the auxiliary modalities as Keys (K) and Values (V)

$$F_{PI}, A_{PI} = \text{CrossAttn}(Q = \mathbf{F''_P}, K = V = \mathbf{F'_I}) \tag{9}$$

$$F_{PT}, A_{PT} = \text{CrossAttn}(Q = \mathbf{F''_P}, K = V = \mathbf{F'_T}) \tag{10}$$

Here, F_{PI} and F_{PT} represent point-wise features enriched by visual and textual contexts, while A_{PI} and A_{PT} are the attention weights representing the attention degree between each point cloud token and the auxiliary tokens. For each point token f_i^P, its visual prototype $\hat{\mathbf{v}}_i$ and textual prototype $\hat{\mathbf{t}}_i$ are calculated as

$$\hat{\mathbf{v}}_i = \sum_{j=1}^{N_I} (A_{PI})_{ij} \mathbf{f}_j^I, \quad \hat{\mathbf{t}}_i = \sum_{k=1}^{N_T} (A_{PT})_{ik} \mathbf{f}_k^T \tag{11}$$

This step effectively projects the discrete feature spaces of images and texts into the geometric space of the point cloud, facilitating the construction of effective geometric triplets $(\mathbf{f}_i^P, \hat{\mathbf{v}}_i, \hat{\mathbf{t}}_i)$ for volume calculation. Since not all triplets contribute meaningfully to the loss computation, we introduce a Top-k filtering strategy to select the K most critical triplets. We utilize the peak values of attention weights to measure the confidence of modal alignment. When a point cloud token attends to specific regions of the image and text with high weights, it implies not only more accurate semantic alignment but also higher multi-modal consistency; consequently, such tokens are considered highly reliable. Based on this, we define a reliability score s_i for each point cloud token

$$s_i = \max_j((A_{PI})_{ij} \cdot \max_k((A_{PT})_{ik} \tag{12}$$

We select the index set $\mathcal{K}$ of the top-k points with the highest scores s_i.

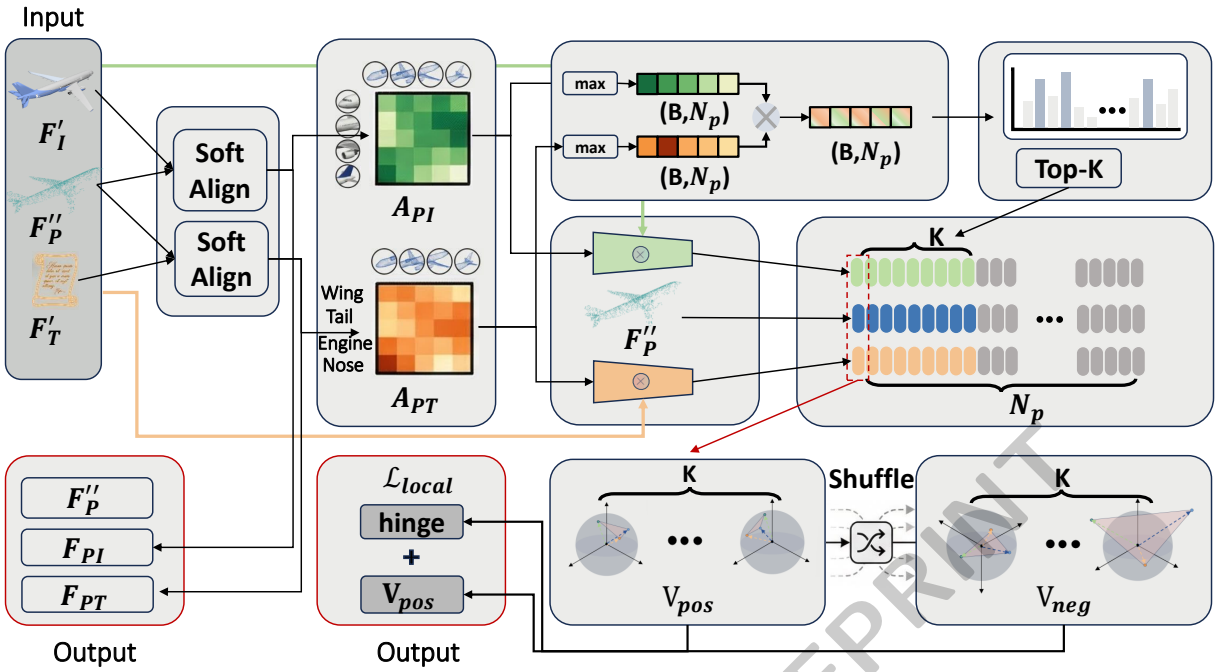

Figure 4: Pipeline of the Local GRAM Alignment module. First, SoftAlign maps image and text features into the point cloud space via cross-attention. Next, Top-K filtering selects only the most reliable matches based on attention confidence to exclude noise. Finally, a Hinge Ranking Loss minimizes the Gramian volume of positive triplets (V_{pos}) while maximizing that of shuffled negatives (V_{neg}), enforcing compact local geometric alignment.

Local GRAM. For the selected point $i \in \mathcal{K}$, we utilize the Gram determinant to calculate the positive sample volume $V_{pos}^{(i)}$ of the aligned triplet $(\mathbf{f}_i^P, \hat{\mathbf{v}}_i, \hat{\mathbf{t}}_i)$. Minimizing $V_{pos}^{(i)}$ alone might lead the model to concentrate attention meaninglessly to achieve a trivially small volume. To mitigate this, we adopt a Hinge Ranking Loss. We construct negative sample triplets by shuffling the aligned triplets to obtain $V_{neg}^{(i)}$, forcing the volume formed by matched tri-modal local features (V_{pos}) to be as small as possible, and the volume of mismatched features (V_{neg}) to be as large as possible. The local loss is defined as

$$\mathcal{L}_{local} = \frac{1}{k} \sum_{i \in \mathcal{K}} \left[\max \left(0, \mu + V_{pos}^{(i)} - V_{neg}^{(i)} \right) + V_{pos}^{(i)} \right] \quad (13)$$

where μ is a margin hyperparameter.

Global Confidence Gating. Complementing the fine-grained local alignment, we employ a Global Gramian Gating mechanism to handle overall semantic conflicts. As shown in Figure 2, We extract global features $(\mathbf{g}^I, \mathbf{g}^T)$ for images and text via CLIP, and generate the global point cloud feature $\mathbf{g}^P$ via average pooling of the point cloud tokens F_P. We estimate the global confidence $\gamma \in (0, 1]$ by calculating the Gramian volume V_{global} of the global features $(\mathbf{g}^P, \mathbf{g}^I, \mathbf{g}^T)$. A smaller volume indicates higher global consistency, yielding a higher confidence score

$$\gamma = \exp\left(-\frac{V_{global}}{\tau} \right) \quad (14)$$

Next, we generate a local feature gate. We concatenate the point features with the aggregated auxiliary contexts to form $\mathbf{F}_{cat} = [F_P''; F_{PI}; F_{PT}]$. This combined representation is passed through a gating network (MLP) to produce point-wise gates $\boldsymbol{\alpha} \in [0, 1]^{N_P \times d}$

$$\boldsymbol{\alpha} = \sigma(\text{MLP}(\mathbf{F}_{cat})) \quad (15)$$

Finally, we leverage both the learned local gate and the global confidence scalar to modulate feature injection. The updated point features $\hat{F}^P$ are computed via a residual connection

$$\hat{F}^P = F_P'' + (\gamma \cdot \boldsymbol{\alpha}) \odot (F_{PI} + F_{PT}) \quad (16)$$

3.4 Coarse-to-Fine Decoding

We employ a hierarchical coarse-to-fine decoder that takes only the fused point tokens $\hat{F}_P$ as input, without additional image or text skip features. The decoding process starts with a coarse decoder, which regresses a sparse structural scaffold from the latent embeddings using a transposed-convolution module and a residual MLP. The coarse output is then fed into cascaded upsampling layers to progressively densify point coordinates and recover fine-grained surface details until the target resolution is reached. Existing multi-modal completion networks often inject auxiliary visual features into the decoder. In contrast, our decoder relies solely on the fused point tokens produced by the proposed fusion and alignment modules. This design highlights the effectiveness of Gramian-guided fusion, suggesting that the fused point features encode sufficient cross-modal semantics for reconstruction and reduce the need for extra modality-specific skip features.

Methods	Avg	Airplane	Cabinet	Car	Chair	Lamp	Sofa	Table	Watercraft
ViPC (CVPR 2021)	3.308	1.760	4.558	3.183	2.476	2.867	4.481	4.990	2.197
CSDN (TVCG 2024)	2.570	1.251	3.670	2.977	2.835	2.554	3.240	2.575	1.742
XMFnet (NeurIPS 2022)	1.443	0.572	1.980	1.754	1.403	1.810	1.702	1.386	0.945
CDPNet (AAAI 2024)	1.374	0.671	1.902	1.857	1.504	1.024	1.694	1.437	0.904
CMNet (TCSVT 2025)	1.270	0.606	1.845	1.759	1.337	0.878	1.556	1.306	0.869
EGIINet (ECCV 2024)	1.211	0.534	1.921	1.655	1.204	0.776	1.552	1.227	0.802
IAET (IJCAI 2025)	1.090	0.503	1.397	1.648	1.196	0.668	1.401	1.200	0.704
QGGA (Ours)	**1.007**	**0.499**	**1.332**	**1.502**	**1.037**	**0.616**	**1.261**	**1.180**	**0.629**

Table 1: Comparison of CD (lower is better) on ShapeNet-ViPC (with our auto-generated ViPC-Desc captions for QGGA).

Methods	Avg	Airplane	Cabinet	Car	Chair	Lamp	Sofa	Table	Watercraft
ViPC (CVPR 2021)	0.591	0.803	0.451	0.512	0.529	0.706	0.434	0.594	0.730
CSDN (TVCG 2024)	0.695	0.862	0.548	0.560	0.669	0.761	0.557	0.729	0.782
XMFnet (NeurIPS 2022)	0.796	0.961	0.662	0.691	0.809	0.792	0.723	0.830	0.901
CDPNet (AAAI 2024)	0.814	0.955	0.697	0.692	0.799	0.890	0.738	0.829	0.914
CMNet (TCSVT 2025)	0.827	0.965	0.700	0.702	0.824	0.903	0.759	0.848	0.917
EGIINet (ECCV 2024)	0.836	0.969	0.693	0.723	0.847	0.919	0.756	0.857	0.927
IAET (IJCAI 2025)	0.860	0.980	0.782	0.725	0.850	0.934	0.795	0.867	0.950
QGGA (Ours)	**0.878**	**0.981**	**0.801**	**0.752**	**0.884**	**0.946**	**0.828**	**0.868**	**0.962**

Table 2: Comparison of F-Score (higher is better) on ShapeNet-ViPC (with our auto-generated ViPC-Desc captions for QGGA).

3.5 Loss Functions

We adopt Chamfer Distance (CD) as the primary reconstruction objective to measure the geometric discrepancy between two point sets. Let P be the predicted point cloud and G be the ground truth point cloud. The CD loss is defined as:

$$\mathcal{L}_{\text{CD}}(P,G) = \frac{1}{|P|}\sum_{p\in P}\min_{g\in G}\|p-g\|_2^2 + \frac{1}{|G|}\sum_{g\in G}\min_{p\in P}\|g-p\|_2^2 \tag{17}$$

Since our decoder employs a "Coarse-to-Fine" multi-stage strategy comprising four resolution levels, the total objective function aggregates structural supervision from all stages and incorporates the Gramian alignment regularization term proposed in this paper:

$$\mathcal{L}_{\text{total}} = \sum_{k=0}^{3}\mathcal{L}_{\text{CD}}(P_k, G_k) + \lambda\cdot\mathcal{L}_{\text{local}} \tag{18}$$

where P_k and G_k denote the predicted and ground truth point clouds at stage k, respectively. $\mathcal{L}_{\text{local}}$ is the reliability-aware Gramian alignment loss defined in Section 3.3, and λ is a balancing hyperparameter.

4 Experiments

4.1 Datasets

ShapeNet-ViPC. We evaluate our method on the widely used ShapeNet-ViPC benchmark [Zhang *et al.*, 2021]. Derived from ShapeNet, this dataset contains 38,328 objects across 8 categories. Each object is rendered from 24 different viewpoints to generate partial views, totaling 919,872 samples. Following the official split, we use 80% of the data for training and 20% for testing.

4.2 Experimental Settings

Implementation Details. All experiments are conducted using the PyTorch framework on an NVIDIA H100 GPU. We use the Adam optimizer with an initial learning rate of 1×10^{-3} and weight decay of 1×10^{-4}. The batch size is set to 64. The learning rate decays by a factor of 0.7 at fixed milestones. Following the ViPC benchmark [Zhang *et al.*, 2021], we set the input image height H and width W to 224. The input partial point cloud contains 2,048 points, which are downsampled via Farthest Point Sampling (FPS) to $N_P = 256$ center point tokens before being fed into the encoder. The feature embedding dimension C for all modalities is uniformly set to 256. We employ a pre-trained and frozen CLIP (ViT-B/16) model as the visual and text encoder [Radford *et al.*, 2021], with the maximum text token sequence length set to 77. The core multi-modal fusion module consists of 4 Transformer blocks, each containing 8 attention heads. The decoder progressively restores the point count to 2,048 via three upsampling stages (factor of 2) for final evaluation. For the proposed Gramian alignment, based on ablation results, we set the balance weight $\lambda = 0.05$, the margin $\mu = 0.07$, the $\tau = 0.1$ and the top-k ratio to 25%.

Metrics. Following standard protocols, we employ L2 Chamfer Distance (CD) and F-Score as the primary evaluation metrics. CD measures the geometric fidelity between the predicted shape and the ground truth, while F-Score evaluates the recall and precision of the reconstruction.

4.3 Evaluation on ShapeNet-ViPC

We compare our method with some multi-modal completion methods, as shown in Table 1 and Table 2. Our method achieves superior performance in CD metrics across most

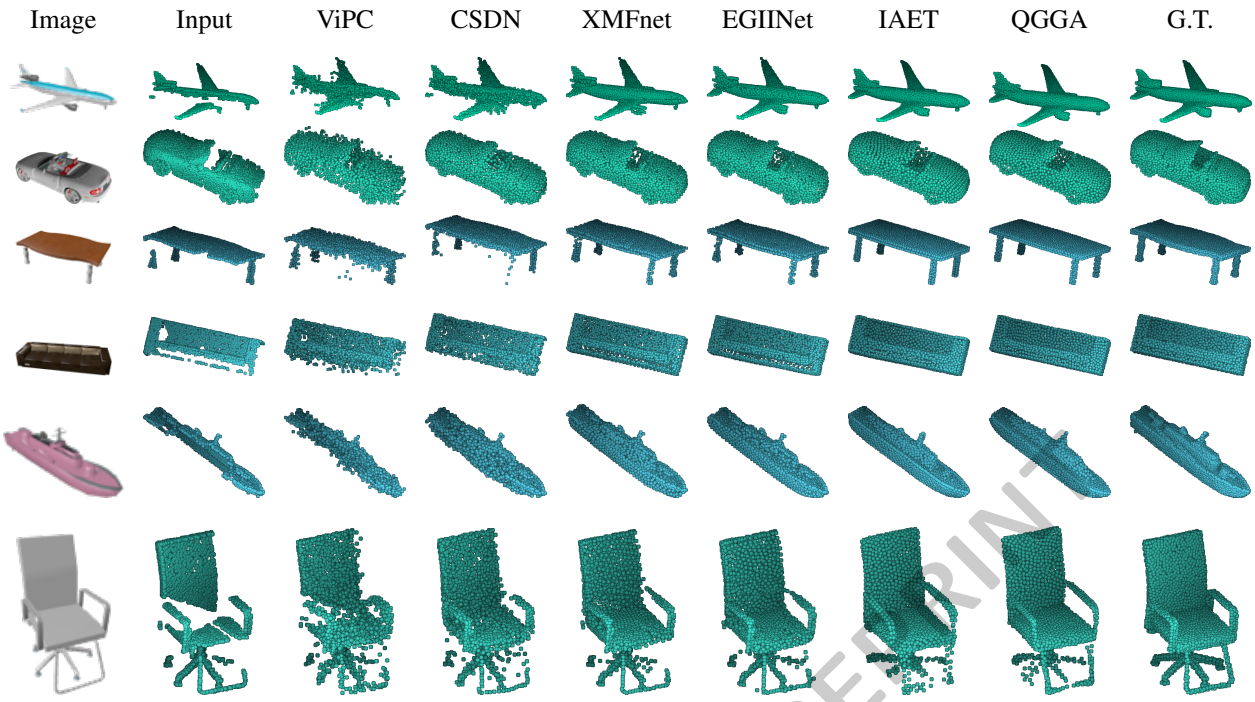

Figure 5: Visualizations comparison on ShapeNet-ViPC dataset.

categories. Notably, compared to the strong baseline IAET [Fang *et al.*, 2025], our method significantly reduces the average CD, validating the effectiveness of Gramian-guided alignment in mitigating semantic ambiguity. Figure 5 visualizes reconstruction results on challenging samples. Due to modal misalignment, baseline methods often produce noisy artifacts or fail to recover fine structures. In contrast, our method generates clean, high-fidelity shapes while preserving fine-grained local details.

4.4 Ablation Study

To validate the effectiveness of each proposed contribution, we conducted ablation studies on the cabinet category of the ShapeNet-ViPC dataset, as shown in Table 3. First, replacing QDA with standard attention (w/o QDA) degraded the performance to 1.522 CD and 0.753 F-Score, demonstrating the importance of query-dependent gating for controlling cross-modal information injection. Second, removing the local Gramian alignment loss (w/o L-GRAM) further increased the CD to 1.571 and decreased the F-Score to 0.744, showing that reconstruction supervision alone is insufficient for reliable fine-grained cross-modal alignment. Furthermore, excluding global confidence gating (w/o Gram-Gate) raised the CD to 1.808 and lowered the F-Score to 0.723, confirming that global reliability estimation helps suppress inconsistent auxiliary information. Overall, the full QGGA model achieves the best performance on the cabinet category, demonstrating the effectiveness of combining query-dependent gated fusion, local Gramian alignment, and global confidence gating.

Model	CD ↓	F-Score ↑
w/o QDA	1.522	0.753
w/o L-GRAM	1.571	0.744
w/o Gram-Gate	1.808	0.723
QGGA (Ours)	**1.332**	**0.801**

Table 3: Ablation study on the cabinet category of ShapeNet-ViPC with generated text descriptions.

5 Conclusion

In this paper, we addressed the problem of multi-modal point cloud completion, leveraging image appearance cues and textual semantic priors to assist geometric recovery under incomplete point cloud observations. To tackle the heterogeneity in representation forms and granularity across modalities, as well as unstable cross-modal correspondences, we proposed QGGA. For fusion, we employed interleaved cross-modal interaction and introduced query-dependent gating to achieve controllable information injection at channel levels. For alignment, we constructed soft correspondences based on Cross-Attention weights, utilized Gramian volumes to constrain local consistency, and incorporated global consistency confidence to adaptively regulate fusion intensity. The fused and aligned features are fed into a coarse-to-fine decoder to progressively generate the complete point cloud. Finally, we established the ShapeNet-ViPC-Desc multi-modal benchmark with text descriptions, providing unified data support and an evaluation foundation for future research.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U25A202453, 62536006, 62376142, 62406180).

References

[Aiello *et al.*, 2022] 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, pages 37349–37362, New Orleans, USA, November 2022.

[Cicchetti *et al.*, 2025] Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 42128–42149, Singapore, April 2025.

[Du *et al.*, 2024a] 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, pages 1635–1643, Vancouver, Canada, February 2024.

[Du *et al.*, 2024b] 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 2024.

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

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

[Fang *et al.*, 2026] 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, pages 3822–3830, Singapore, January 2026.

[Geiger *et al.*, 2012] 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)*, pages 3354–3361, Providence, USA, June 2012.

[Guo *et al.*, 2024] Ting Guo, Da Wang, Jiye Liang, Kaihan Zhang, and Jianchao Zeng. Spear: A spectral approach for zero-shot node classification. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37, pages 69000–69021, 2024.

[Guo *et al.*, 2026a] Ting Guo, Yangrui Fan, Caixia Cui, Jiye Liang, Jiao Zhao, and Da Wang. Topological anomaly quantification for semi-supervised graph anomaly detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Rio de Janeiro, Brazil, April 2026.

[Guo *et al.*, 2026b] Ting Guo, Dongyu Pei, Gangzhu Qiao, Kaixuan Yao, and Da Wang. PAGE: Progressive anomaly generation network for semi-supervised graph anomaly detection. In *Proceedings of the ACM Web Conference (WWW)*, page 959–967, Dubai, United Arab Emirates, April 2026.

[He *et al.*, 2026] 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, pages 4672–4680, Singapore, January 2026.

[Huang *et al.*, 2020] Zitian Huang, Yikuan Yu, Jiawen Xu, Feng Ni, and Xinyi Le. PF-Net: Point fractal network for 3D point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7662–7670, Virtual, June 2020.

[Li *et al.*, 2023] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M. Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. VoxFormer: Sparse voxel Transformer for camera-based 3D semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9087–9098, Vancouver, Canada, June 2023.

[Liang *et al.*, 2023] 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.

[Liang *et al.*, 2025a] 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, pages 5146–5154, Philadelphia, USA, February 2025.

[Liang *et al.*, 2025b] 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)*, page 7133–7141, Dublin, Ireland, October 2025.

[Liang *et al.*, 2026] Hanzhe Liang, Jie Zhou, Can Gao, Bingyang Guo, Jinbao Wang, and Linlin Shen. A lightweight 3d anomaly detection method with rotationally invariant features. *Pattern Recognition*, 174:112924, 2026.

[Qi *et al.*, 2017] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5105–5114, Long Beach, USA, December 2017.

[Radford *et al.*, 2021] 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, pages 8748–8763, Virtual, July 2021.

[Song *et al.*, 2017] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1746–1754, Honolulu, USA, July 2017.

[Tchapmi *et al.*, 2019] Lyne P Tchapmi, Vineet Kosaraju, Hamid Rezatofighi, Ian Reid, and Silvio Savarese. TopNet: Structural point cloud decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 383–392, Long Beach, USA, June 2019.

[Tesema *et al.*, 2024] 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.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 6000–6010, Long Beach, USA, December 2017.

[Xiang *et al.*, 2023] Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han. Snowflake point deconvolution for point cloud completion and generation with skip-Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(5):6320–6338, May 2023.

[Xie *et al.*, 2020] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Jiageng Mao, Shengping Zhang, and Wenxiu Sun. GRNet: Gridding residual network for dense point cloud completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 365–381, Glasgow, UK, August 2020.

[Xiu *et al.*, 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)*, pages 27435–27444, Honolulu, USA, October 2025.

[Xu *et al.*, 2024] 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)*, pages 414–432, Milan, Italy, September 2024.

[Yan *et al.*, 2025] Hongyu Yan, Zijun Li, Kunming Luo, Li Lu, and Ping Tan. SymmCompletion: High-fidelity and high-consistency point cloud completion with symmetry guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 9094–9102, Philadelphia, USA, February 2025.

[Yu *et al.*, 2021] 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)*, pages 12478–12487, Montreal, Canada, October 2021.

[Yu *et al.*, 2023] 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.

[Yuan *et al.*, 2018] 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)*, pages 728–737, Verona, Italy, September 2018.

[Zhang *et al.*, 2021] 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)*, pages 15885–15894, Virtual, June 2021.

[Zhou *et al.*, 2026] 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, pages 13871–13879, Singapore, January 2026.

[Zhu *et al.*, 2023] 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)*, pages 14508–14518, Paris, France, October 2023.

[Zhu *et al.*, 2024] 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.