

From Alignment to Fusion in 3D Vision-Language

Xueqi Qiu, Xingyu Miao, Jingjing Deng, Haoran Duan, Yang Long, and Ling Shao, *Fellow, IEEE*

Abstract—Unified 3D vision-language systems must combine complementary geometry, scale, and appearance cues while supporting tasks from instance segmentation to language-guided reasoning. Existing methods often process point clouds, voxel grids, and multi-view images independently; directly combining these heterogeneous representations may leave substantial feature discrepancy unresolved, while subsequent unconstrained adaptation may distort their internal geometry. We propose an “align-then-fuse” framework that first applies triple pairwise cosine alignment to establish segment-level correspondence across the three representations and then retrieves task-conditioned features with a prompt-guided query decoder. Before fusion, representation-specific query features are transformed by learnable mappings constrained to the special orthogonal group. These mappings preserve inner products and Euclidean distances within each representation, permitting controlled representation-specific re-parameterisation without arbitrarily distorting its internal geometry. The transformed features are subsequently combined through Adaptive Fusion under downstream task supervision. Experiments cover eight datasets for instance segmentation, visual grounding, question answering, and dense captioning. Compared with PQ3D, the model improves average precision by 3.2 points on ScanNet200 and grounding accuracy by 2.9, 10.6, 4.6, and 4.1 points on ScanRefer, Nr3D, Sr3D, and Multi3DRefer, respectively, while also improving performance on ScanQA, SQA3D, and Scan2Cap. Ablations further support the complementary roles of alignment and orthogonal re-parameterisation and the effectiveness of Adaptive Fusion.

Index Terms—3D vision-language learning, multimodal alignment, multimodal fusion, orthogonal transformation.

I. INTRODUCTION

In real-world scenarios, achieving a joint understanding of 3D environments and natural language is essential for constructing agents with practical operational abilities [1]–[7]. In recent years, with the continuous development of 3D vision-language tasks and their related datasets, significant progress has been made in areas ranging from fundamental 3D semantic and instance segmentation [8], [9] to visual grounding [10]–[14], 3D question answering [15]–[17], 3D dense captioning [18], and open-vocabulary 3D understanding [19]–[22].

Previous research often employed task-specific models built on a single type of 3D representation [15], [23]–[27], while more recent studies increasingly incorporate multiple scene representations within shared 3D vision-language frameworks

[7], [28], [29]. Earlier multimodal 3D methods demonstrated the benefit of combining image and 3D information through spatial projection, feature augmentation, or cross-modal interaction [30]–[35]. More recent approaches address representation heterogeneity more explicitly. Inst3D-LMM integrates multi-view semantics with matched 3D instances [36], while DSPNet selects question-relevant views before combining image and point-cloud features [37]. CrossOver maps heterogeneous scene modalities into a shared embedding space [38], whereas CUA-O3D and Proxy3D reduce feature heterogeneity through uncertainty-aware knowledge aggregation [39] and semantic-aware proxy construction [40], respectively. These developments establish increasingly effective connections among heterogeneous scene representations, but the underlying features are still generated by representation-specific encoding processes before being jointly used for prediction. Consequently, incorporating these heterogeneous representations into a shared prediction architecture does not by itself ensure their latent compatibility before joint prediction.

To address this issue, we propose an “align-then-fuse” framework for multi-representation 3D vision-language understanding. Because features produced by different encoders can remain misaligned before joint prediction, we first introduce Triple Alignment to enforce pairwise cosine correspondence among spatially matched point-cloud, voxel, and multi-view features before prompt-guided decoding. Alignment improves cross-representation correspondence, while the aligned features still retain representation-specific parameterizations before fusion. However, unconstrained transformations may arbitrarily alter the geometry already encoded within each feature space. We therefore constrain each representation-specific mapping to the special orthogonal group $SO(D)$. The resulting transformations preserve pairwise Euclidean distances and inner products within each representation while permitting learnable re-parameterisation. Finally, the transformed features are combined through Adaptive Fusion and jointly optimized under downstream task supervision. Experiments on various benchmarks spanning instance segmentation, visual grounding, question answering, and dense captioning demonstrate broad improvements over the displayed baselines and support the effectiveness of explicitly aligning heterogeneous representations before their controlled adaptation and fusion.

In summary, our contributions are fourfold:

- We investigate two closely related challenges in unified 3D vision-language learning: representation heterogeneity before joint modeling and latent-basis discrepancy during cross-representation fusion.
- To address representation heterogeneity, we introduce Triple Alignment to establish explicit pairwise correspondence among spatially matched point-cloud, voxel, and multi-view features.

Xueqi Qiu and Xingyu Miao contributed equally to this work.

Xueqi Qiu, Xingyu Miao and Yang Long are with the Department of Computer Science, Durham University, UK (e-mail: xueqi.qiu@durham.ac.uk; xingyu.miao@durham.ac.uk; yang.long@durham.ac.uk).

Jingjing Deng is with the School of Engineering Mathematics and Technology, University of Bristol, Bristol, UK (e-mail: jingjing.deng@bristol.ac.uk).

Haoran Duan is with the Department of Automation, Tsinghua University, Beijing, China (e-mail: haoran.duan@ieee.org).

Ling Shao is with the UCAS-Terminus AI Lab, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: ling.shao@ieee.org).

Yang Long and Haoran Duan are the corresponding authors.

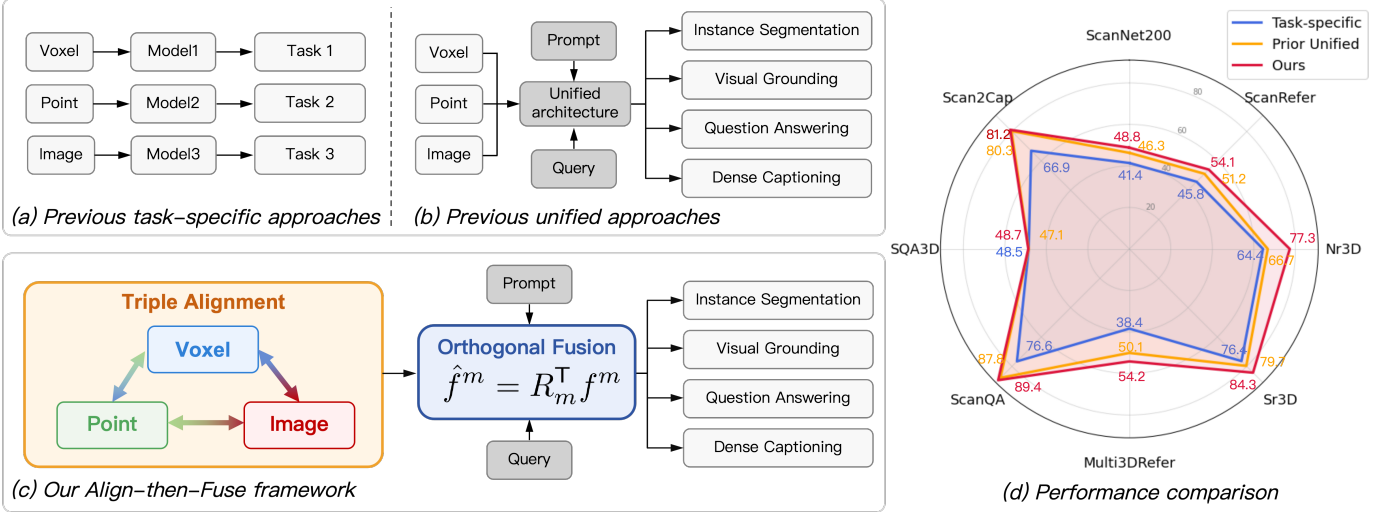


Fig. 1. **Overview and performance of our unified 3D vision-language framework.** We illustrate the progression from previous task-specific and unified approaches in (a) and (b) to our align-then-fuse framework in (c), which integrates voxel, point-cloud, and image representations through Triple Alignment and Orthogonal Fusion. The performance comparison across eight benchmarks is shown in (d).

- To address latent-basis discrepancy during fusion, we introduce representation-specific orthogonal reparameterisation based on $SO(D)$, allowing learnable adaptation while preserving intra-representation geometry, followed by Adaptive Fusion.
- We evaluate our approach on eight 3D vision-language datasets: ScanNet200 [8] for instance segmentation; ScanRefer [12], ReferIt3D (Nr3D and Sr3D) [11], and Multi3DRefer [14] for visual localization; ScanQA [15] and SQA3D [16] for question answering; and Scan2Cap [18] for dense description. In extensive experiments, we demonstrate that our method outperforms the baseline.

II. RELATED WORK

A. Query-Based 3D Vision-Language Learning

3D vision-language learning has expanded from individual perception tasks to a broad range of language-conditioned scene understanding problems, including segmentation [8], [41], visual grounding [10]–[12], [14], [23], question answering [15], [16], and dense captioning [18]. Early methods were predominantly developed for individual tasks, with task-specific architectures and prediction heads tailored to different forms of 3D perception and reasoning.

Query-based prediction provides a natural route toward a more unified formulation. Originating from set-based object prediction [42]–[46], query representations have been extended to mask prediction [47], language-guided segmentation [48], [49], and multimodal prediction [50]–[54]. Rather than defining a separate prediction mechanism for each task, latent queries interact with visual representations and can be conditioned by different forms of prompts, providing a shared interface between input guidance and downstream outputs.

This paradigm has subsequently been adopted in 3D scene understanding. OpenMask3D [22], OpenScene [55], PLA [56], and Mosaic3D [41] associate 3D regions or proposals with

vision-language representations for open-vocabulary recognition and segmentation. More recent methods extend query-based interaction beyond category-level prediction toward language-conditioned scene understanding. LL3DA [29] introduces language-guided interaction with 3D scene features, while PQ3D [28] employs a shared set of promptable queries to accommodate multiple 3D vision-language tasks within a common prediction framework. These developments mark a transition from task-specific 3D vision-language pipelines toward query-based formulations in which diverse prediction objectives can be expressed through a shared interaction mechanism.

B. 3D Multimodal Fusion

Point clouds, voxel grids, and multi-view images provide complementary descriptions of a 3D scene. Point clouds preserve detailed geometric observations, voxel grids impose regular spatial structure, and multi-view images provide dense appearance information. These representations, however, arise from different sampling processes and are typically processed by independently designed encoders, resulting in representation-specific feature spaces with no inherent guarantee of latent compatibility.

Earlier multimodal 3D methods project image features into a shared spatial representation for intermediate fusion [30], [31], or augment point features with image-derived semantic and appearance cues [32], [33]. Subsequent approaches introduce cross-modal attention and deformable interaction to support more flexible information exchange between image and 3D representations [34], [35], [57]–[59]. These methods demonstrate the effectiveness of combining complementary representations, although their interaction mechanisms are typically developed around a particular representation pair or downstream setting.

More recent studies address representation heterogeneity more explicitly. Inst3D-LMM [36] injects multi-view semantics



Fig. 2. **Feature distribution discrepancy across heterogeneous 3D representations.** Feature distributions of voxel, multi-view (MV), and point-cloud (PC) representations are visualized for four representative scenes. Despite describing the same physical scenes, the three representations exhibit distinct distributional characteristics, illustrating the representation discrepancy induced by heterogeneous encoding processes.

into matched 3D instances and models their spatial relations, whereas DSPNet [37] selects question-relevant views before integrating image and point-cloud features. CrossOver [38] maps multiple, potentially incomplete scene modalities into a shared embedding space, while CUA-O3D [39] and Proxy3D [40] reduce feature heterogeneity through uncertainty-aware knowledge aggregation and semantic-aware proxy construction, respectively. These approaches provide effective mechanisms for connecting heterogeneous scene observations, but the resulting features are still produced by representation-specific encoding processes before being combined for prediction. Consequently, their latent compatibility is not guaranteed by the shared prediction architecture itself. This motivates an explicit treatment of feature compatibility before joint prediction, together with a controlled mechanism for adapting and combining heterogeneous representations without unnecessarily altering the structure already encoded within each feature space.

III. PRELIMINARIES

Task Prompt Encoding. We formulate different 3D vision-language tasks using unified task prompts of two types: textual prompts, including object categories, referring expressions, and questions; and numerical prompts, including 3D bounding boxes and 3D locations. Textual prompts are encoded using a pretrained CLIP model [60], while numerical prompts are projected to the same hidden dimension through learnable fully connected layers. The resulting prompt features provide task-specific conditioning for downstream prediction.

Multi-Representation 3D Encoding. To encode a 3D scene, three representations are considered: point clouds, voxel grids, and multi-view images. For point clouds, the entire set of 3D points is partitioned into segments. In each segment, 1,024 points are sampled and normalized to a unit sphere. These points are processed through a pretrained PointNet++ backbone [61], [62] to obtain point features $\mathcal{P} = \{p_1, p_2, \dots, p_M\} \in \mathbb{R}^{M \times D}$, where M denotes the number of segments. Voxel features are extracted by discretizing the 3D scene into voxels.

The voxel grid is processed with a sparse convolutional U-Net that includes both downsampling and upsampling stages to capture hierarchical details. Features from each voxel are pooled to their corresponding segments and projected into the hidden dimension to form $\mathcal{V} = \{v_1, v_2, \dots, v_M\} \in \mathbb{R}^{M \times D}$. Multi-view image features are derived by computing per-pixel embeddings using a pretrained segmentation model (e.g., OpenSeg [63]). The 2D pixels are back-projected into the 3D space, and features from multiple views are averaged to form image features $\mathcal{I} = \{i_1, i_2, \dots, i_M\} \in \mathbb{R}^{M \times D}$.

Prompt-Guided Query Decoding (Multi-Task Heads). The fused scene representation and task prompt are processed by a Transformer-style query decoder with N_q learnable instance queries,

$$Q_i \in \mathbb{R}^{N_q \times D}, \quad (1)$$

whose positional embeddings are initialized from farthest-point-sampled 3D anchors [64]. At each decoder layer, the queries aggregate scene information through cross-attention, interact with the task prompt for task conditioning, and are further refined by spatial self-attention. The resulting prompt-conditioned queries are finally fed to task-specific prediction heads.

IV. METHOD

A. Problem Formulation

- **Modal heterogeneity.** Prior work [28], [29] employs prompt-based queries to extract features from point clouds, multi-scale voxels, and multi-view images. These representations are typically processed in parallel and are either selected according to downstream tasks or mapped into a shared feature space. However, as shown in Figure 2, different representations exhibit distinct feature distributions even for the same scene, revealing a persistent discrepancy among heterogeneous representations. This motivates explicit feature alignment before their joint use.
- **Latent-basis discrepancy.** Alignment alone does not require independently encoded representations to share

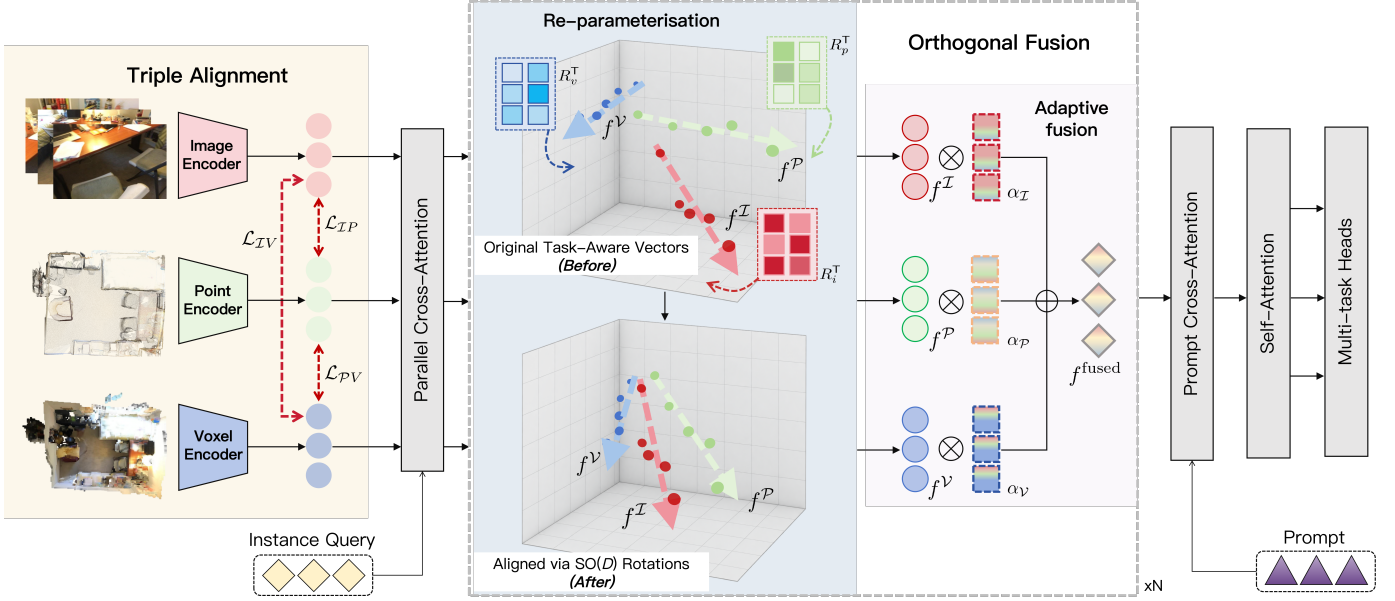


Fig. 3. **Architecture of the proposed “align-then-fuse” framework.** Point-cloud, voxel, and multi-view image features are first coordinated through Triple Alignment to establish cross-representation correspondence. The resulting representation-specific features are then adapted by orthogonal re-parameterisation, which preserves their intra-representation geometry, and combined through Adaptive Fusion. The fused representation is subsequently conditioned on the task prompt and processed by the multi-task prediction heads for downstream 3D vision-language tasks.

the same latent parameterization. After alignment, representation-specific adaptation can therefore still be beneficial to retain the flexibility of individual representations before fusion. We refer to this remaining mismatch between independently parameterized feature spaces as latent-basis discrepancy. Applying unconstrained transformations at this stage, however, may arbitrarily distort their encoded feature geometry. We therefore seek a constrained re-parameterisation that can adapt each representation-specific feature space while preserving its internal geometric structure.

To address these challenges, we adopt an “align-then-fuse” strategy. We first explicitly align the features of point-cloud, voxel, and multi-view image representations. The resulting features are then adapted through representation-specific orthogonal re-parameterisation, which permits learnable transformation while preserving pairwise geometry within each representation. The adapted features are subsequently combined through Adaptive Fusion, yielding a fused scene representation that is passed to the multi-task prediction heads for downstream prediction. The overall formulation therefore separates cross-representation correspondence from representation-specific adaptation, allowing heterogeneous 3D features to be coordinated without unconstrained distortion before fusion.

B. Alignment for Representation Heterogeneity

Features extracted from multi-view images, point clouds, and voxel grids are generated by heterogeneous encoders and may exhibit substantial representation discrepancy. We therefore introduce Triple Alignment, a triple pairwise cosine alignment objective that explicitly encourages corresponding

scene-segment features to occupy compatible directions before task-conditioned decoding.

Given multi-view image features $\mathcal{I} \in \mathbb{R}^{M \times D}$, point-cloud features $\mathcal{P} \in \mathbb{R}^{M \times D}$, and voxel features $\mathcal{V} \in \mathbb{R}^{M \times D}$, we formulate Triple Alignment as a segment-level triple pairwise cosine alignment objective. We first define the cosine distance between two feature vectors $x, y \in \mathbb{R}^D$ as

$$\mathcal{L}(x, y) = 1 - \frac{x^\top y}{\|x\|_2 \|y\|_2}, \quad (2)$$

where $x^\top y$ denotes the inner product and $\|\cdot\|_2$ denotes the Euclidean norm.

For the m -th scene segment, let i_m , p_m , and v_m denote the corresponding image, point-cloud, and voxel features, respectively. Triple Alignment explicitly considers all three unordered representation pairs:

$$\begin{aligned} \mathcal{L}_{\mathcal{IP}} &= \frac{1}{M} \sum_{m=1}^M \mathcal{L}(i_m, p_m), \\ \mathcal{L}_{\mathcal{IV}} &= \frac{1}{M} \sum_{m=1}^M \mathcal{L}(i_m, v_m), \\ \mathcal{L}_{\mathcal{PV}} &= \frac{1}{M} \sum_{m=1}^M \mathcal{L}(p_m, v_m). \end{aligned} \quad (3)$$

The overall Triple Alignment objective is defined as

$$\mathcal{L}_{\text{align}} = w_{\mathcal{IP}} \mathcal{L}_{\mathcal{IP}} + w_{\mathcal{IV}} \mathcal{L}_{\mathcal{IV}} + w_{\mathcal{PV}} \mathcal{L}_{\mathcal{PV}}, \quad (4)$$

where $w_{\mathcal{IP}}$, $w_{\mathcal{IV}}$, and $w_{\mathcal{PV}}$ control the contributions of the image–point, image–voxel, and point–voxel alignment terms, respectively. Minimizing Equation (4) encourages corresponding segment features from the three representations to occupy compatible directions in the latent space before task-conditioned

decoding. Triple Alignment thereby establishes explicit pre-fusion correspondence and improves cross-representation comparability without requiring the three representation distributions to become identical.

C. Orthogonal Fusion with Re-parameterisation

Although Triple Alignment improves correspondence among the three representations, the queried features are still generated through independently parameterized representation streams. Directly applying unconstrained linear adapters before fusion may introduce arbitrary scaling, shearing, and anisotropic distortion, thereby altering the internal feature geometry of each stream. We therefore restrict the modality-specific adapters to the special orthogonal group $\text{SO}(D)$. This design does not assume that the residual cross-representation discrepancy is generated by a rotation. Instead, orthogonality serves as a geometry-preserving constraint that allows each representation to be adaptively re-parameterized without changing its internal pairwise structure.

Given the aligned scene representations, each query independently retrieves features from the image, point-cloud, and voxel streams through parallel cross-attention. This produces three query-conditioned vectors, $f^{\mathcal{I}}, f^{\mathcal{P}}, f^{\mathcal{V}} \in \mathbb{R}^D$, which retain their representation-specific parameterizations before cross-representation fusion. We apply orthogonal re-parameterisation to the three query-conditioned features and subsequently combine them through adaptive fusion. The fused query representation is then passed to the subsequent query decoding and task-specific prediction heads.

a) Group preliminaries: The set of real matrices satisfying $M^T M = I$ forms the orthogonal group $\text{O}(D)$. Its subset with $\det M = 1$ is the special orthogonal group $\text{SO}(D)$, whose elements are pure rotations. For any $R \in \text{SO}(D)$ and any $x, y \in \mathbb{R}^D$, the action of R preserves Euclidean distance and inner product:

$$\|Rx - Ry\|_2 = \|x - y\|_2 \quad \text{and} \quad \langle Rx, Ry \rangle = \langle x, y \rangle. \quad (5)$$

b) Why QR decomposition: To obtain a differentiable orthogonality-constrained adapter, we project an unconstrained square matrix onto an element of $\text{SO}(D)$. The thin QR decomposition $A = QR$ meets this requirement because its orthogonal factor Q already lies in $\text{O}(D)$ and can be converted to a proper rotation with at most one column-sign flip. We adopt QR decomposition as a differentiable and computationally convenient way to obtain the orthogonal factor [65], with efficient implementations available in modern deep-learning frameworks. Most importantly, the orthogonality of Q preserves Euclidean distances and inner products within each representation stream, preventing the adapter from arbitrarily distorting its internal geometry before fusion.

c) Isometric re-parameterisation: For each modality $m \in \mathcal{I}, \mathcal{P}, \mathcal{V}$ we keep a learnable square matrix $A_m \in \mathbb{R}^{D \times D}$ and compute its thin QR decomposition $A_m = O_m U_m$ at every forward pass, where O_m is orthogonal and U_m is upper-triangular. We form the rotation:

$$R_m = \begin{cases} O_m, & \det O_m = 1 \\ \text{diag}(1, \dots, 1, -1) O_m, & \det O_m = -1 \end{cases}, \quad (6)$$

where $R_m \in \text{SO}(D)$.

The modality feature is then rotated:

$$\hat{f}^m = R_m^T f^m. \quad (7)$$

By (5), applying R_m^T preserves distances and inner products between features within the same representation. The mapping therefore performs modality-specific controlled adaptation without changing the internal geometry of that representation stream.

d) Lemma: For any $R \in \text{SO}(D)$ and $x, y \in \mathbb{R}^D$, Equation (5) holds.

Proof. Let $R \in \text{SO}(D)$ and $x, y \in \mathbb{R}^D$. We want to show that

$$\|Rx - Ry\|_2 = \|x - y\|_2 \quad \text{and} \quad \langle Rx, Ry \rangle = \langle x, y \rangle. \quad (8)$$

First, consider the squared Euclidean norm of the difference:

$$\|Rx - Ry\|_2^2 = (Rx - Ry)^T (Rx - Ry). \quad (9)$$

Since matrix transpose distributes over addition and scalar multiplication,

$$(Rx - Ry)^T = x^T R^T - y^T R^T, \quad (10)$$

so that

$$\|Rx - Ry\|_2^2 = (x^T R^T - y^T R^T)(Rx - Ry). \quad (11)$$

By associativity of matrix multiplication,

$$\|Rx - Ry\|_2^2 = x^T R^T R x - x^T R^T R y - y^T R^T R x + y^T R^T R y. \quad (12)$$

Because R is orthogonal, $R^T R = I$. Substituting this identity gives

$$\|Rx - Ry\|_2^2 = x^T x - x^T y - y^T x + y^T y = \|x - y\|_2^2. \quad (13)$$

Taking square roots on both sides yields the first equality, $\|Rx - Ry\|_2 = \|x - y\|_2$.

Next, we verify preservation of the inner product. By definition,

$$\langle Rx, Ry \rangle = (Rx)^T (Ry) = x^T R^T R y. \quad (14)$$

Again using $R^T R = I$, this reduces to

$$\langle Rx, Ry \rangle = x^T y = \langle x, y \rangle. \quad (15)$$

This completes the proof that applying the same element of $\text{SO}(D)$ to two feature vectors preserves their Euclidean distance and inner product.

Since Triple Alignment is applied before orthogonal re-parameterisation, corresponding features from the three representations enter the transformation stage from an aligned latent space. Although the representation-specific transformations R_m are learned independently, their outputs are jointly fused and optimized under the same downstream task objective. This allows each representation to retain its own geometry-preserving adaptation while coordinating its contribution to the final prediction. Triple Alignment thus provides explicit pre-fusion correspondence, while the fused task objective provides implicit task-driven coordination among the transformed representations.

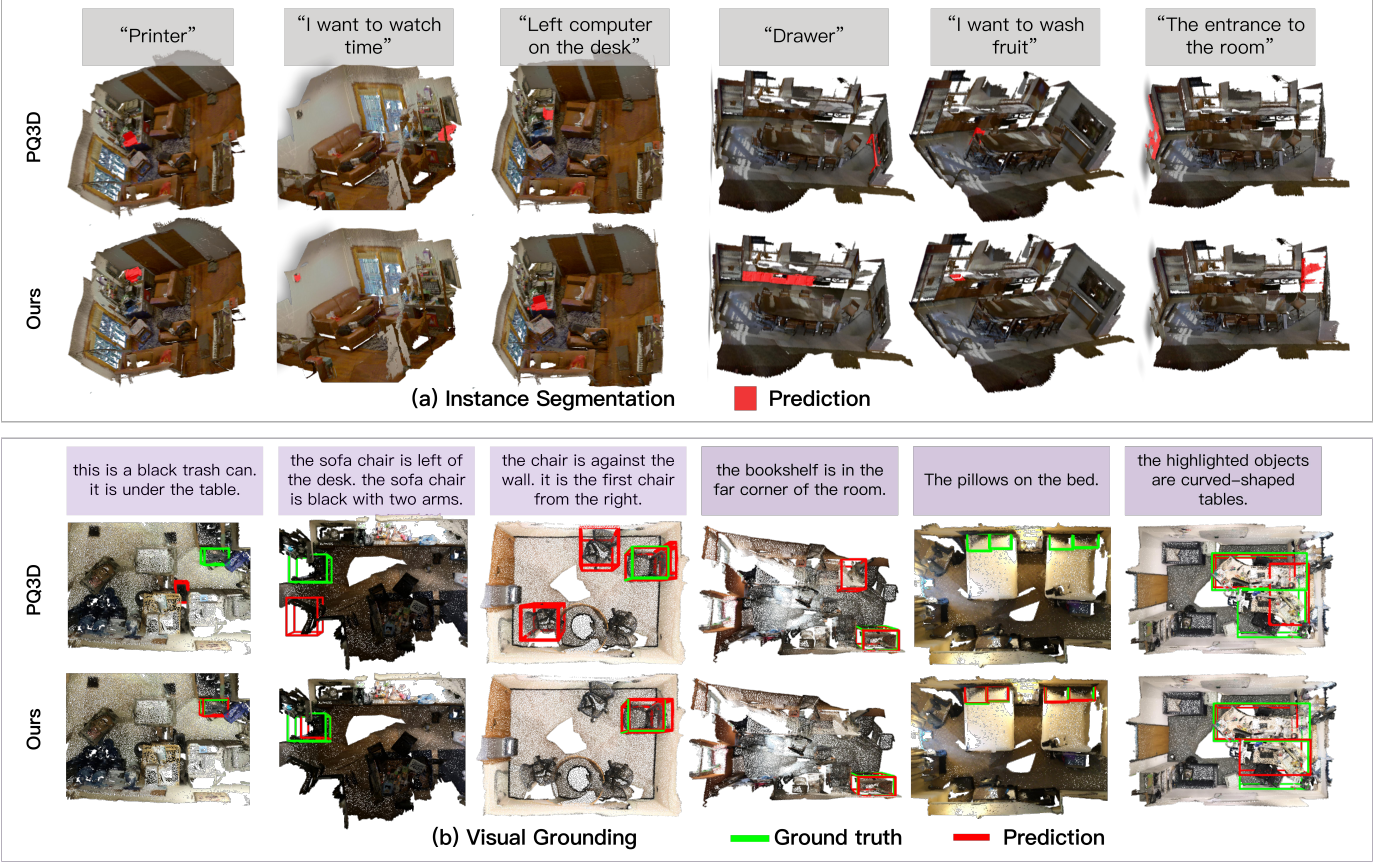


Fig. 4. **Qualitative results on (a) instance segmentation and (b) visual grounding.** The results demonstrate accurate instance-level scene understanding and reliable localization of language-referred objects across diverse 3D scenes.

e) Adaptive fusion: After orthogonal re-parameterisation, the three representation streams retain complementary scene information but need not contribute equally to the final prediction. A direct summation assigns equal importance to all representations, while an unconstrained weighted combination may alter the feature scale. We therefore employ a lightweight adaptive fusion mechanism to learn their relative contributions under downstream supervision. Specifically, we assign a learnable fusion weight to each of the three re-parameterized feature streams:

$$w = [w_{\mathcal{I}}, w_{\mathcal{P}}, w_{\mathcal{V}}] \in \mathbb{R}^3. \quad (16)$$

The corresponding fusion coefficients are obtained through softmax normalization:

$$\alpha_m = \frac{\exp(w_m)}{\sum_{n \in \{\mathcal{I}, \mathcal{P}, \mathcal{V}\}} \exp(w_n)}, \quad m \in \{\mathcal{I}, \mathcal{P}, \mathcal{V}\}, \quad (17)$$

such that

$$\alpha_m \geq 0, \quad \sum_{m \in \{\mathcal{I}, \mathcal{P}, \mathcal{V}\}} \alpha_m = 1. \quad (18)$$

The transformed representation features are then fused as

$$f^{\text{fused}} = \sum_{m \in \{\mathcal{I}, \mathcal{P}, \mathcal{V}\}} \alpha_m \hat{f}^m. \quad (19)$$

Because the fusion coefficients are non-negative and sum to one, Equation (19) forms a convex combination of the transformed features. Together with the norm-preserving property of orthogonal re-parameterisation, the fused feature satisfies

$$\begin{aligned} \|f^{\text{fused}}\|_2 &\leq \sum_m \alpha_m \|\hat{f}^m\|_2 \\ &= \sum_m \alpha_m \|f^m\|_2 \\ &\leq \max_m \|f^m\|_2. \end{aligned} \quad (20)$$

Thus, the fusion operation allows the relative contributions of the three representations to be learned without introducing unbounded feature-scale amplification through the fusion coefficients.

The fusion weights are optimized jointly with the representation-specific transformations and downstream prediction objectives. Here, “adaptive” refers to the learned adjustment of representation contributions during end-to-end training rather than input-dependent or query-dependent gating. The proposed modules are optimized jointly with the downstream task objectives. Specifically, the overall training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}, \quad (21)$$

where $\mathcal{L}_{\text{task}}$ denotes the aggregate task-specific prediction loss and λ_{align} controls the contribution of Triple Alignment.

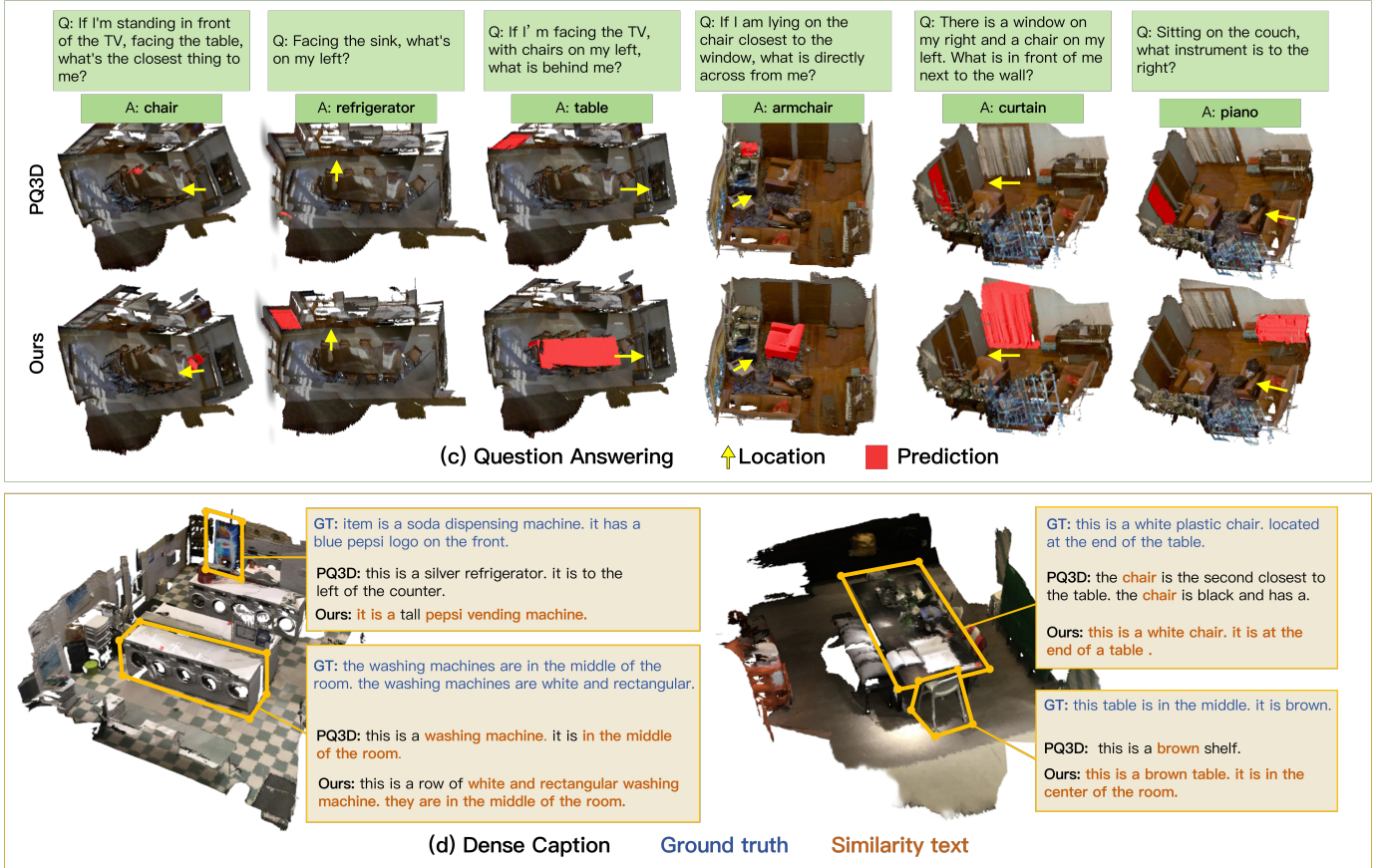


Fig. 5. **Qualitative results on (c) question answering and (d) dense captioning.** Representative predictions from our model are shown for different 3D scenes.

TABLE I
INSTANCE SEGMENTATION RESULTS ON THE SCANNet200 VALIDATION SET.

Method	ScanNet200					
	AP	AP ₅₀	AP ₂₅	head	common	tail
Isbnet [66]	24.5	32.7	37.6	38.6	20.5	12.5
Mask3D [27]	26.9	36.2	41.4	39.8	21.7	17.9
PQ3D [28]	27.0	38.9	46.3	35.8	24.2	20.0
Ours	30.2	41.6	48.8	52.7	26.7	21.4

V. EXPERIMENT

A. Datasets, Metrics, and Implementation

We evaluate models on eight public benchmarks spanning four task families. ScanNet200 [8] supplies instance masks, evaluated with AP, AP₅₀, AP₂₅, and head/common/tail category AP. ScanRefer [12], Nr3D [11], Sr3D [11], and Multi3DRefer [14] evaluate language-guided localization with their standard accuracy splits. ScanQA is evaluated with exact match, BLEU-1, METEOR, and CIDEr; SQA3D [16] uses exact-match answer accuracy; and Scan2Cap [18] uses CIDEr, BLEU-4, METEOR, and ROUGE under IoU@0.5. We follow the benchmark protocols used by PQ3D [28].

The training procedure consists of two stages. In the first

stage, we train the model with instance segmentation alone on ScanNet200 for 800 epochs, using a classification head on instance queries instead of prompting object categories, and we guide the first 200 epochs with ground truth mask guidance [44] to speed convergence. In the second stage, we continue training on the full training set with all objectives for 50 epochs. We optimize with AdamW at a learning rate of 1×10^{-4} , batch size of 16, $\beta_1 = 0.9$, and $\beta_2 = 0.98$, applying a cosine decay learning rate scheduler. We follow the data augmentation strategy of PQ3D [28], including geometric transformations and color augmentation during training. We set the number of queries to 120 to balance inference speed and the number of objects in the scene. All training is run on 4 NVIDIA A100 GPUs.

B. Qualitative Results

Figures 4 and 5 present qualitative results across the four evaluated task families. In Figure 4, for instance segmentation, our model responds to both explicit category prompts, such as “Printer” and “Drawer”, and more functional descriptions, such as “I want to wash fruit”, demonstrating consistent language-conditioned scene understanding. In visual grounding, the advantage is more evident for expressions that require spatial disambiguation among nearby objects. For example, for “the black trash can under the table” and “the first chair from the

TABLE II

GROUNDING PERFORMANCE (%) ON 3D VISUAL GROUNDING BENCHMARKS. SCANREFER AND MULTI3DREFER ARE EVALUATED UNDER AN IOU THRESHOLD OF 0.5, WHILE NR3D AND SR3D FOLLOW THE STANDARD EVALUATION PROTOCOL USING GROUND-TRUTH OBJECT MASKS. FOR MULTI3DREFER, ZT, ST, AND MT DENOTE THE ZERO-TARGET, SINGLE-TARGET, AND MULTI-TARGET SETTINGS, RESPECTIVELY.

Method	ScanRefer			Nr3D			Sr3D			Multi3DRefer			
	Unique	Multiple	Avg.	Easy	Hard	Avg.	Easy	Hard	Avg.	ZT	ST	MT	Avg.
ViL3DRel [23]	68.6	30.7	37.7	70.2	57.4	64.4	74.9	67.9	72.8	-	-	-	-
3DJCG [67]	64.3	30.8	37.3	-	-	-	-	-	-	66.9	16.7	26.2	26.6
UniT3D [68]	73.1	31.1	39.1	-	-	-	-	-	-	-	-	-	-
M3DRef-CLIP [14]	77.2	36.8	44.7	55.6	43.4	49.4	-	-	-	39.4	30.6	37.9	38.4
3D-VisTA [7]	75.1	39.1	45.8	72.1	56.7	64.2	78.8	71.3	76.4	-	-	-	-
PQ3D [28]	<u>78.2</u>	<u>46.2</u>	<u>51.2</u>	<u>75.0</u>	<u>58.7</u>	<u>66.7</u>	<u>82.7</u>	<u>72.8</u>	<u>79.7</u>	57.7	43.6	40.9	50.1
Ours	80.8	49.2	54.1	83.9	71.0	77.3	84.5	84.0	84.3	<u>59.3</u>	48.2	46.5	54.2

TABLE III

ANSWER ACCURACY ON THE SCANQA AND SQA3D BENCHMARKS. FOR SCANQA, EACH ENTRY REPORTS THE RESULTS ON “TEST WITH OBJECT” AND “TEST WITHOUT OBJECT”, RESPECTIVELY. IN SQA3D, THE RESULTS ARE REPORTED USING EXACT MATCH (EM) ACCURACY BETWEEN THE PREDICTED AND GROUND-TRUTH ANSWERS.

Method	ScanQA				SQA3D						
	EM@1	BLEU-1	METEOR	CIDEr	What	Is	How	Can	Which	Other	Avg.
GPT-3 [69]	-	-	-	-	39.7	46.0	40.5	45.6	36.1	38.4	41.0
SQA3D [16]	23.5 / 20.9	31.6 / 30.7	13.6 / 12.6	67.3 / 60.2	31.6	<u>63.8</u>	46.0	<u>69.5</u>	43.9	45.3	46.6
3D-VisTA [7]	<u>27.0</u> / <u>23.0</u>	34.4 / 30.2	15.2 / 12.9	76.6 / 62.6	34.8	63.3	45.4	69.8	47.2	<u>48.1</u>	<u>48.5</u>
PQ3D [28]	26.1 / 20.0	<u>43.0</u> / <u>36.1</u>	<u>17.8</u> / <u>13.9</u>	<u>87.8</u> / <u>65.2</u>	<u>37.1</u>	61.3	44.5	60.9	47.0	45.1	47.1
Ours	28.2 / 23.5	46.4 / 38.0	19.6 / 16.5	89.4 / 67.3	36.5	63.9	<u>45.5</u>	67.4	<u>47.1</u>	50.8	48.7

TABLE IV

CAPTIONING PERFORMANCE ON THE SCAN2CAP BENCHMARK UNDER THE IOU@0.5 CRITERION.

Method	CIDEr	Scan2Cap		
		BLEU-4	METEOR	ROUGE
Scan2Cap [18]	35.2	22.4	21.4	43.5
3DJCG [67]	47.7	31.5	24.3	51.8
3D-VisTA [7]	66.9	34.0	27.1	54.3
PQ3D [28]	<u>80.3</u>	<u>36.0</u>	<u>29.1</u>	<u>57.9</u>
Ours	81.2	39.4	30.4	63.5

right”, our predictions align more closely with the ground-truth regions than those of PQ3D, suggesting better use of relational cues beyond object semantics. Similar behavior is observed in question answering in Figure 5, where egocentric questions involving directions such as left, behind, and across require reasoning with respect to the specified viewpoint; our predictions more consistently identify the corresponding answer-relevant objects. The dense-captioning examples further reveal differences in semantic and spatial understanding. PQ3D confuses the Pepsi vending machine with a refrigerator and the brown table with a shelf, whereas our model correctly identifies the object categories while retaining relevant attributes and spatial relations. For the washing-machine example, our description additionally captures its white, rectangular appearance and its location in the middle of the room. These examples qualitatively support the model’s ability to associate language with object identity, attributes, and spatial structure across different 3D vision-language tasks.

TABLE V

ABLATION STUDY: TRIPLE ALIGNMENT (TA) AND ORTHOGONAL FUSION (OF) FOR SCANNet200 INSTANCE SEGMENTATION.

TA	OF	AP	AP ₅₀	AP ₂₅
-	-	27.0	38.9	46.3
✓	-	30.0	41.1	47.2
-	✓	29.8	41.3	47.6
✓	✓	30.2	41.6	48.8

TABLE VI

ABLATION STUDY: TRIPLE ALIGNMENT (TA) AND ORTHOGONAL FUSION (OF) FOR VISUAL GROUNDING. “TA” IS SHORT FOR TRIPLE ALIGNMENT, “OF” IS SHORT FOR ORTHOGONAL FUSION.

TA	OF	ScanRefer	Nr3D	Sr3D	Multi3DRefer
-	-	51.2	66.7	79.7	50.1
✓	-	<u>53.3</u>	71.8	80.5	<u>53.1</u>
-	✓	52.8	<u>75.4</u>	83.7	51.9
✓	✓	54.1	77.3	84.3	54.2

C. Quantitative Results

a) *Instance Segmentation:* As shown in Table I, our method achieves the best performance among the compared methods on ScanNet200 across all evaluation metrics. Compared with representative methods such as ISBNNet [66], Mask3D [27], and PQ3D [28], our approach improves the overall Average Precision (AP) from 27.0 to 30.2, and further increases AP₅₀ and AP₂₅ from 38.9 and 46.3 to 41.6 and 48.8, respectively. In terms of long-tailed class performance,

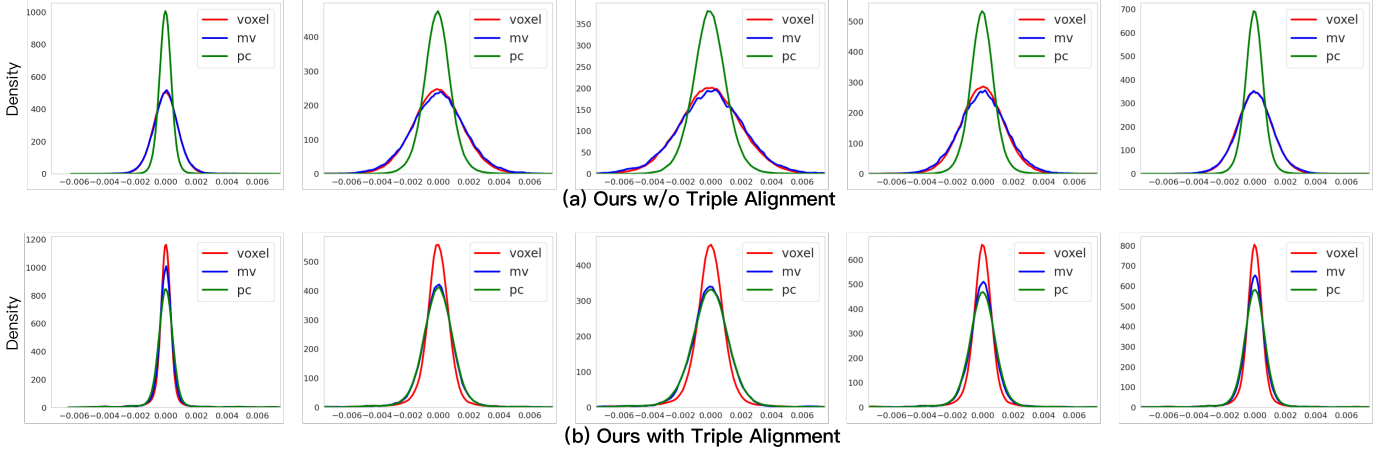


Fig. 6. **Ablation study of Triple Alignment on feature distributions.** Density curves of voxel, multi-view (MV), and point-cloud (PC) features are shown for representative ScanNet200 scenes, comparing (a) our model without Triple Alignment and (b) the full model. Triple Alignment leads to more consistent feature distributions across the three representations.

TABLE VII
ABLATION STUDY: TRIPLE ALIGNMENT (TA) AND ORTHOGONAL FUSION (OF) FOR QUESTION ANSWERING ON SCANQA.

TA	OF	EM@1	BLEU-1	METEOR	CIDEr
-	-	26.1 / 20.0	43.0 / 36.1	17.8 / 13.9	87.8 / 65.2
✓	-	27.5 / 21.4	45.8 / 38.0	19.2 / 15.2	89.0 / 66.8
-	✓	27.9 / 21.8	46.4 / 38.5	18.8 / 14.5	88.5 / 67.1
✓	✓	28.2 / 23.5	46.4 / 38.0	19.6 / 16.5	89.4 / 67.3

TABLE VIII
ABLATION STUDY: TRIPLE ALIGNMENT (TA) AND ORTHOGONAL FUSION (OF) FOR ANSWER ACCURACY ON SQA3D.

TA	OF	What	Is	How	Can	Which	Other	Avg.
-	-	37.1	61.3	44.5	60.9	47.0	45.1	47.1
✓	-	36.6	62.1	45.2	63.7	46.1	46.2	50.8
-	✓	37.5	62.3	44.8	62.0	47.2	49.5	49.9
✓	✓	36.5	63.9	45.5	67.4	47.1	50.8	48.7

our method substantially improves the head-class AP to 52.7, surpassing PQ3D by 16.9 points, while also improving the common and tail categories from 24.2 and 20.0 to 26.7 and 21.4.

Unlike approaches that directly operate on individual or heterogeneous representations, our method explicitly establishes cross-representation correspondence before their joint use. Triple Alignment first aligns spatially corresponding features, after which representation-specific orthogonal reparameterisation enables controlled adaptation while preserving intra-representation geometry, followed by adaptive fusion under downstream supervision. The consistent improvements across head, common, and tail categories demonstrate the effectiveness of integrating complementary representations within this align-then-fuse framework.

b) *Visual Grounding*: As shown in Table II, we evaluate grounding accuracy on four benchmark datasets, including ScanRefer, Nr3D, Sr3D, and Multi3DRefer. Compared with PQ3D, our method improves the average accuracy by 2.9, 10.6,

TABLE IX
ABLATION STUDY: TRIPLE ALIGNMENT (TA) AND ORTHOGONAL FUSION (OF) FOR CAPTION PERFORMANCE ON SCAN2CAP.

TA	OF	CIDEr	BLEU-4	METEOR	ROUGE
-	-	80.3	36.0	29.1	57.9
✓	-	81.3	38.5	30.7	62.4
-	✓	80.7	38.9	29.8	62.5
✓	✓	<u>81.2</u>	39.4	<u>30.4</u>	63.5

and 4.6 points on ScanRefer, Nr3D, and Sr3D, respectively. On Multi3DRefer, the average accuracy increases from 50.1 to 54.2, with the ST and MT scores improving from 43.6 and 40.9 to 48.2 and 46.5, respectively. For the ZT setting, our method achieves 59.3, improving upon PQ3D’s 57.7 but remaining below the best result of 66.9 from 3DJCG. Overall, the consistent gains across the four grounding benchmarks indicate that explicitly aligning heterogeneous scene representations before their controlled adaptation and fusion is particularly beneficial for language-conditioned object localization.

c) *Question Answering*: As shown in Table III, our method achieves the best overall results on ScanQA across both exact-match and language-generation metrics. On the ‘test with object’ split, our model improves EM@1 from the previous best of 27.0 to 28.2, while increasing BLEU-1, METEOR, and CIDEr over PQ3D from 43.0, 17.8, and 87.8 to 46.4, 19.6, and 89.4, respectively. On the ‘test without object’ split, EM@1 increases from 23.0 to 23.5, while BLEU-1, METEOR, and CIDEr improve from 36.1, 13.9, and 65.2 to 38.0, 16.5, and 67.3. These improvements show that the benefit of combining aligned complementary representations extends beyond object localization to language-conditioned scene reasoning.

On SQA3D, our method achieves the highest average accuracy of 48.7, slightly exceeding the previous best result of 48.5. It also obtains the best performance for the ‘Is’ and ‘Other’ question categories, reaching 63.9 and 50.8, respectively, while remaining competitive across the other question types. Although the overall improvement is modest compared with

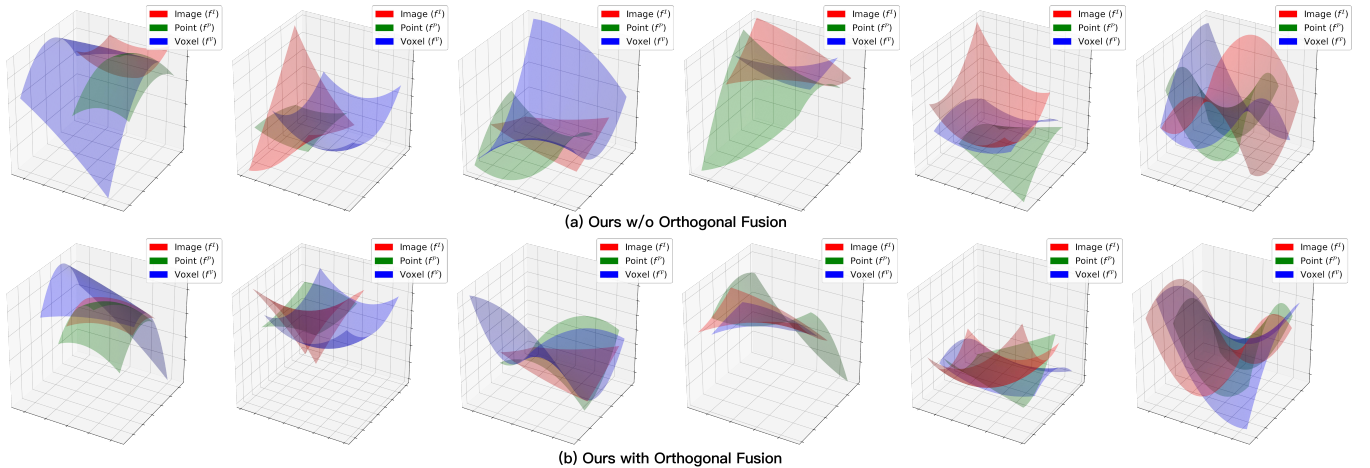


Fig. 7. **Ablation study on Orthogonal Fusion.** For representative examples, we fit bivariate polynomial surfaces to the feature distributions of the image, point-cloud, and voxel representations, comparing the variant without Orthogonal Fusion (top) and the full model (bottom). After orthogonal re-parameterisation with Equation (7), the fitted surfaces exhibit greater overlap, indicating more consistent coordination among the three representations before fusion.

TABLE X
ABLATION STUDY: FUSION METHODS FOR INSTANCE SEGMENTATION ON SCANNet200.

Methods	AP	AP ₅₀	AP ₂₅
AF	30.2	41.6	48.8
$\mathcal{I} \oplus \mathcal{P} \oplus \mathcal{V}$	28.2	39.2	45.4
$\mathcal{I} \otimes \mathcal{P} \otimes \mathcal{V}$	30.0	41.2	48.3
$\mathcal{I} \oplus \mathcal{P} \otimes \mathcal{V}$	28.5	40.6	47.1
$\mathcal{I} \otimes \mathcal{P} \oplus \mathcal{V}$	28.7	40.5	47.3

that observed on the grounding benchmarks, the results indicate that the proposed representation integration strategy remains effective for situated question answering without uniformly improving every question category.

d) *Dense Captioning*: As shown in Table IV, our method achieves the best performance among the compared approaches across all four Scan2Cap metrics. Compared with PQ3D, CIDEr increases from 80.3 to 81.2, BLEU-4 from 36.0 to 39.4, METEOR from 29.1 to 30.4, and ROUGE from 57.9 to 63.5. In particular, the larger gains on BLEU-4 and ROUGE indicate improved agreement between the generated descriptions and reference captions, while the improvements in CIDEr and METEOR further confirm the overall captioning quality. Together with the results on segmentation, grounding, and question answering, these results demonstrate that the proposed align-then-fuse formulation generalizes across both perception-oriented and language-generation tasks.

D. Ablation Study

a) Ablation on Triple Alignment and Orthogonal Fusion:

Figure 6 provides an illustration of the effect of introducing the Triple Alignment module on the feature distributions across different representations. In the original setting, observable discrepancies exist among the representations, as reflected by differences in peak position and distribution width. After introducing Triple Alignment, the density curves become more

TABLE XI
ABLATION STUDY: FUSION METHODS FOR VISUAL GROUNDING. “AF” REPRESENTS ADAPTIVE FUSION, $\oplus$ REPRESENTS ELEMENT-BY-ELEMENT ADDITION. $\otimes$ REPRESENTS FEATURE CONCATENATION AND IS MAPPED THROUGH A FULLY CONNECTED LAYER.

Methods	ScanRefer	Nr3D	Sr3D	Multi3DRefer
AF	54.1	77.3	84.3	54.2
$\mathcal{I} \oplus \mathcal{P} \oplus \mathcal{V}$	49.3	65.5	78.3	44.4
$\mathcal{I} \otimes \mathcal{P} \otimes \mathcal{V}$	53.6	75.2	81.9	55.1
$\mathcal{I} \oplus \mathcal{P} \otimes \mathcal{V}$	50.2	72.1	79.7	48.0
$\mathcal{I} \otimes \mathcal{P} \oplus \mathcal{V}$	50.4	71.8	80.2	47.6

TABLE XII
ABLATION STUDY: FUSION METHODS FOR ANSWER ACCURACY ON SCANQA.

Methods	EM@1	BLEU-1	METEOR	CIDEr
AF	28.2 / 23.5	46.4 / 38.0	19.6 / 16.5	89.4 / 67.3
$\mathcal{I} \oplus \mathcal{P} \oplus \mathcal{V}$	25.0 / 19.9	42.2 / 34.1	17.8 / 13.9	82.0 / 60.8
$\mathcal{I} \otimes \mathcal{P} \otimes \mathcal{V}$	28.0 / 23.5	45.8 / 37.5	19.2 / 16.0	89.2 / 67.3
$\mathcal{I} \oplus \mathcal{P} \otimes \mathcal{V}$	26.3 / 21.5	43.4 / 36.3	18.0 / 14.8	86.4 / 64.2
$\mathcal{I} \otimes \mathcal{P} \oplus \mathcal{V}$	26.8 / <u>21.8</u>	44.0 / 36.5	18.3 / 15.2	87.5 / <u>65.1</u>

consistent in both peak location and distribution width, indicating reduced representation discrepancy in the latent space. This observation is consistent with the role of Triple Alignment in establishing correspondence among spatially corresponding scene features. As shown in Table VI, introducing Triple Alignment alone improves the visual grounding accuracy over the baseline by 2.1, 5.1, 0.8, and 3.0 points on ScanRefer, Nr3D, Sr3D, and Multi3DRefer, respectively. When Orthogonal Fusion is already enabled, further introducing Triple Alignment still yields improvements of 1.3, 1.9, 0.6, and 2.3 points on the four benchmarks. These results demonstrate that explicitly establishing segment-level correspondence before fusion consistently benefits language-conditioned object localization.

Figure 7 qualitatively illustrates the effect of Orthogonal Fusion on multi-representation feature coordination. We fit

TABLE XIII
ABLATION STUDY: FUSION METHODS FOR ANSWER ACCURACY ON SQA3D.

Methods	What	Is	How	Can	Which	Other	Avg.
AF	36.5	63.9	45.5	67.4	47.1	50.8	48.7
$\mathcal{I} \oplus \mathcal{P} \oplus \mathcal{V}$	34.7	61.2	43.9	60.3	45.5	46.8	42.3
$\mathcal{I} \otimes \mathcal{P} \otimes \mathcal{V}$	<u>36.2</u>	63.9	45.5	67.4	<u>47.2</u>	50.8	48.2
$\mathcal{I} \oplus \mathcal{P} \otimes \mathcal{V}$	35.6	<u>62.5</u>	<u>44.5</u>	62.4	46.7	48.2	46.5
$\mathcal{I} \otimes \mathcal{P} \oplus \mathcal{V}$	35.5	<u>62.2</u>	<u>44.2</u>	<u>63.8</u>	46.5	<u>49.3</u>	47.4

TABLE XIV
ABLATION STUDY: FUSION METHODS FOR CAPTIONING PERFORMANCE ON SCAN2CAP.

Methods	CIDEr	BLEU-4	METEOR	ROUGE
AF	81.2	39.4	30.4	63.5
$\mathcal{I} \oplus \mathcal{P} \oplus \mathcal{V}$	78.3	35.7	28.1	55.9
$\mathcal{I} \otimes \mathcal{P} \otimes \mathcal{V}$	<u>81.6</u>	<u>39.0</u>	<u>30.2</u>	<u>63.8</u>
$\mathcal{I} \oplus \mathcal{P} \otimes \mathcal{V}$	79.5	37.5	29.5	60.4
$\mathcal{I} \otimes \mathcal{P} \oplus \mathcal{V}$	79.8	38.5	29.7	61.3

bivariate polynomial surfaces to features extracted from the image, point-cloud, and voxel representations for representative examples, comparing the variant without Orthogonal Fusion with the full model. After orthogonal re-parameterisation, the fitted surfaces generally exhibit greater overlap, providing a qualitative indication that the representation-specific features are more consistently coordinated before fusion. Importantly, this effect does not imply that the three representations are forced into an identical feature distribution. Instead, the orthogonal transformations provide representation-specific adaptation while preserving pairwise distances and inner products within each representation.

The quantitative results further support this observation. As shown in Table V, introducing Orthogonal Fusion alone improves AP, AP₅₀, and AP₂₅ on ScanNet200 by 2.8, 2.4, and 1.3 points, respectively. Similar improvements are observed for visual grounding: compared with the baseline, Orthogonal Fusion alone improves performance by 1.6, 8.7, 4.0, and 1.8 points on ScanRefer, Nr3D, Sr3D, and Multi3DRefer, respectively. On ScanQA (Table VII), it improves BLEU-1 by 3.4/2.4 points and CIDEr by 0.7/1.9 points on the test sets with/without object annotations. The results on SQA3D (Table VIII) and Scan2Cap (Table IX) further show that both components provide complementary benefits across question answering and captioning, although neither consistently dominates every individual metric. Together, these results suggest that geometry-preserving re-parameterisation provides effective representation-specific adaptation before cross-representation fusion.

b) Ablation for Fusion Methods: As shown in Table X to XIV, we systematically compare our adaptive fusion with alternative multi-representation fusion strategies based on element-wise addition ($\oplus$) and feature concatenation followed by projection ($\otimes$).

Our Adaptive Fusion achieves the strongest overall performance across the evaluated tasks. It obtains the best results on all three ScanNet200 metrics and on three of the four visual-

TABLE XV
MODEL COMPLEXITY COMPARISON UNDER PROMPT-CONDITIONED INFERENCE.

Model	Parameters	Single Forward (s)	Forward FLOPs
PQ3D	247,875,873 (247.88M)	0.1340	25,808,556,116 (25.81 GFLOPs)
Ours	254,953,773 (254.95M)	<u>0.1648</u>	26,941,755,488 (26.94 GFLOPs)

grounding benchmarks, while full concatenation performs best on Multi3DRefer with 55.1. A similar pattern is observed for the language tasks: Adaptive Fusion achieves the best or tied-best results across the ScanQA metrics and the highest average accuracy of 48.7 on SQA3D. On Scan2Cap, it achieves the best BLEU-4 and METEOR scores of 39.4 and 30.4, whereas full concatenation obtains slightly higher CIDEr and ROUGE scores of 81.6 and 63.8.

Overall, full concatenation generally performs better than direct element-wise addition, suggesting that simply combining the representations with equal contribution is insufficient. In comparison, our Adaptive Fusion assigns normalized learnable coefficients to the transformed representations and optimizes their relative contributions jointly under downstream task supervision. Its consistently strong performance across different task families supports the use of learned representation coordination after alignment and orthogonal re-parameterisation, while the few metrics favoring concatenation also indicate that no single fusion rule uniformly dominates every evaluation setting.

E. Model Complexity Comparison.

We further compare the computational complexity of our method with PQ3D under prompt-conditioned inference. As shown in Table XV, PQ3D contains 247.88M parameters and requires 25.81 GFLOPs for a single forward pass, whereas our model contains 254.95M parameters and requires 26.94 GFLOPs. This corresponds to an increase of approximately 7.08M parameters (2.9%) and 1.13 GFLOPs (4.4%), respectively. The inference time increases from 0.1340 s to 0.1648 s per forward pass, corresponding to an additional 0.0308 s (23.0%). The additional inference cost mainly arises from the representation-specific orthogonal re-parameterisation and Adaptive Fusion operations. Overall, our method introduces a modest increase in model complexity while consistently improving performance across the evaluated 3D vision-language tasks.

VI. DISCUSSION AND LIMITATIONS

The results support a specific proposition: alignment and fusion solve different parts of heterogeneous representation integration. Triple Alignment improves cross-representation correspondence before task decoding, whereas Orthogonal Fusion restricts the subsequent re-parameterisation so that it does not alter pairwise distances or inner products within each representation. Their combination is consistently strongest in the visual-grounding ablation, although the ablations across the remaining tasks show that neither component dominates every individual metric. These findings should be interpreted within the tested indoor-scene benchmarks and pretrained feature pipelines.

The method relies on pretrained, representation-specific encoders and therefore does not establish fully end-to-end cross-representation learning. Orthogonal re-parameterisation and Adaptive Fusion introduce additional inference cost: relative to PQ3D, the measured model has 254.95M rather than 247.88M parameters, 26.94 rather than 25.81 GFLOPs, and a 0.1648-s rather than 0.1340-s forward pass. Evaluation is limited to the eight reported datasets, which largely share indoor-scene structure; robustness to outdoor scenes, missing representations, and sensor corruption remains untested. These limitations motivate more efficient representation adaptation and fusion, as well as explicit evaluation under missing-representation settings.

VII. CONCLUSION

We present an “align-then-fuse” framework for 3D vision-language understanding that integrates point clouds, voxel grids, and multi-view images within a unified architecture. Our Triple Alignment establishes segment-level correspondence across heterogeneous representations before task decoding, while representation-specific orthogonal re-parameterisation enables controlled adaptation by preserving the pairwise geometry within each representation. The resulting features are then combined through Adaptive Fusion under downstream task supervision. Extensive experiments on eight datasets show that our model achieves competitive or state-of-the-art performance across tasks ranging from low-level segmentation to high-level reasoning.

REFERENCES

- [1] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian *et al.*, “Do as i can, not as i say: Grounding language in robotic affordances,” in *Conference on Robot Learning (CoRL)*, 2022.
- [2] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suen-derhauf, “Sayplan: Grounding large language models using 3d scene graphs for scalable task planning,” in *7th Annual Conference on Robot Learning*, 2023.
- [3] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, “A survey of embodied ai: From simulators to research tasks,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 2, pp. 230–244, 2022.
- [4] J. Huang, S. Yong, X. Ma, X. Linghu, P. Li, Y. Wang, Q. Li, S.-C. Zhu, B. Jia, and S. Huang, “An embodied generalist agent in 3d world,” in *International Conference on Machine Learning (ICML)*, 2024.
- [5] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, “Language models as zero-shot planners: Extracting actionable knowledge for embodied agents,” in *International Conference on Machine Learning (ICML)*. PMLR, 2022, pp. 9118–9147.
- [6] A. Padmakumar, J. Thomason, A. Shrivastava, P. Lange, A. Narayan-Chen, S. Gella, R. Piramuthu, G. Tur, and D. Hakkani-Tur, “Teach: Task-driven embodied agents that chat,” in *AAAI Conference on Artificial Intelligence (AAAI)*, 2022.
- [7] Z. Zhu, X. Ma, Y. Chen, Z. Deng, S. Huang, and Q. Li, “3d-vista: Pre-trained transformer for 3d vision and text alignment,” in *International Conference on Computer Vision (ICCV)*, 2023, pp. 2911–2921.
- [8] D. Rozenberszki, O. Litany, and A. Dai, “Language-grounded indoor 3d semantic segmentation in the wild,” in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 125–141.
- [9] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma *et al.*, “The replica dataset: A digital replica of indoor spaces,” *arXiv preprint arXiv:1906.05797*, 2019.
- [10] A. Abdelreheem, K. Olszewski, H.-Y. Lee, P. Wonka, and P. Achlioptas, “Scanents3d: Exploiting phrase-to-3d-object correspondences for improved visio-linguistic models in 3d scenes,” in *Proceedings of Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 3524–3534.
- [11] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas, “Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes,” in *European Conference on Computer Vision (ECCV)*, 2020.
- [12] D. Z. Chen, A. X. Chang, and M. Nießner, “Scanrefer: 3d object localization in rgb-d scans using natural language,” in *European Conference on Computer Vision (ECCV)*, 2020.
- [13] B. Jia, Y. Chen, H. Yu, Y. Wang, X. Niu, T. Liu, Q. Li, and S. Huang, “Sceneverse: Scaling 3d vision-language learning for grounded scene understanding,” *European Conference on Computer Vision (ECCV)*, 2024.
- [14] Y. Zhang, Z. Gong, and A. X. Chang, “Multi3drefer: Grounding text description to multiple 3d objects,” in *International Conference on Computer Vision (ICCV)*, 2023, pp. 15 225–15 236.
- [15] D. Azuma, T. Miyayoshi, S. Kurita, and M. Kawanabe, “Scanqa: 3d question answering for spatial scene understanding,” in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [16] X. Ma, S. Yong, Z. Zheng, Q. Li, Y. Liang, S.-C. Zhu, and S. Huang, “Sqa3d: Situated question answering in 3d scenes,” *International Conference on Learning Representations (ICLR)*, 2023.
- [17] L. Zhao, D. Cai, J. Zhang, L. Sheng, D. Xu, R. Zheng, Y. Zhao, L. Wang, and X. Fan, “Towards explainable 3d grounded visual question answering: A new benchmark and strong baseline,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.
- [18] Z. Chen, A. Gholami, M. Nießner, and A. X. Chang, “Scan2cap: Context-aware dense captioning in rgb-d scans,” in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [19] J. Guo, X. Ma, Y. Fan, H. Liu, and Q. Li, “Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.
- [20] H. Ha and S. Song, “Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models,” in *Conference on Robot Learning (CoRL)*, 2022.
- [21] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, “Lerf: Language embedded radiance fields,” in *International Conference on Computer Vision (ICCV)*, 2023, pp. 19 729–19 739.
- [22] A. Takmaz, E. Fedele, R. W. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, “OpenMask3D: Open-Vocabulary 3D Instance Segmentation,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [23] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, “Language conditioned spatial relation reasoning for 3d object grounding,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [24] Z. Guo, Y. Tang, R. Zhang, D. Wang, Z. Wang, B. Zhao, and X. Li, “Viewrefer: Grasp the multi-view knowledge for 3d visual grounding,” in *International Conference on Computer Vision (ICCV)*, 2023, pp. 15 372–15 383.
- [25] D. He, Y. Zhao, J. Luo, T. Hui, S. Huang, A. Zhang, and S. Liu, “Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.
- [26] J. Luo, J. Fu, X. Kong, C. Gao, H. Ren, H. Shen, H. Xia, and S. Liu, “3d-sps: Single-stage 3d visual grounding via referred point progressive selection,” in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [27] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, “Mask3d: Mask transformer for 3d semantic instance segmentation,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 8216–8223.
- [28] Z. Zhu, Z. Zhang, X. Ma, X. Niu, Y. Chen, B. Jia, Z. Deng, S. Huang, and Q. Li, “Unifying 3d vision-language understanding via promptable queries,” in *European Conference on Computer Vision*. Springer, 2024, pp. 188–206.
- [29] S. Chen, X. Chen, C. Zhang, M. Li, G. Yu, H. Fei, H. Zhu, J. Fan, and T. Chen, “LI3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning,” in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [30] M. Liang, B. Yang, S. Wang, and R. Urtasun, “Deep continuous fusion for multi-sensor 3d object detection,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 641–656.
- [31] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, “Multi-task multi-sensor fusion for 3d object detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7345–7353.
- [32] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, “Pointpainting: Sequential fusion for 3d object detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.

- [33] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 794–11 803.
- [34] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1090–1099.
- [35] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, "Deformable feature aggregation for dynamic multi-modal 3d object detection," in *European conference on computer vision*. Springer, 2022, pp. 628–644.
- [36] H. Yu, W. Li, S. Wang, J. Chen, and J. Zhu, "Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 14 147–14 157.
- [37] J. Luo, Y. Liu, W. Chen, Z. Li, Y. Wang, G. Li, and L. Lin, "Dspnet: Dual-vision scene perception for robust 3d question answering," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 14 169–14 178.
- [38] S. D. Sarkar, O. Miksik, M. Pollefeys, D. Barath, and I. Armeni, "Crossover: 3d scene cross-modal alignment," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 8985–8994.
- [39] J. Li, C. Saltori, F. Poiesi, and N. Sebe, "Cross-modal and uncertainty-aware agglomeration for open-vocabulary 3d scene understanding," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 19 390–19 400.
- [40] J. Jiang, H. Sun, D. Gudovskiy, Y. Nakata, T. Okuno, K. Keutzer, and W. Zheng, "Proxy3d: Efficient 3d representations for vision-language models via semantic clustering and alignment," *arXiv preprint arXiv:2605.08064*, 2026.
- [41] J. Lee, C. Park, J. Choe, Y.-C. F. Wang, J. Kautz, M. Cho, and C. Choy, "Mosaic3d: Foundation dataset and model for open-vocabulary 3d segmentation," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 14 089–14 101.
- [42] B. Dong, F. Zeng, T. Wang, X. Zhang, and Y. Wei, "Solq: Segmenting objects by learning queries," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 21 898–21 909, 2021.
- [43] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu, "Instances as queries," in *International Conference on Computer Vision (ICCV)*, 2021, pp. 6910–6919.
- [44] J. He, P. Li, Y. Geng, and X. Xie, "Fastinst: A simple query-based model for real-time instance segmentation," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 23 663–23 672.
- [45] J. Hu, L. Cao, Y. Lu, S. Zhang, Y. Wang, K. Li, F. Huang, L. Shao, and R. Ji, "Istr: End-to-end instance segmentation with transformers," *arXiv preprint arXiv:2105.00637*, 2021.
- [46] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=gZ9hCDWe6ke>
- [47] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [48] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vision-language transformer and query generation for referring segmentation," in *International Conference on Computer Vision (ICCV)*, 2021, pp. 16 321–16 330.
- [49] J. Wu, Y. Jiang, P. Sun, Z. Yuan, and P. Luo, "Language as queries for referring video object segmentation," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4974–4984.
- [50] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer *et al.*, "Perceiver io: A general architecture for structured inputs & outputs," *arXiv preprint arXiv:2107.14795*, 2021.
- [51] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning (ICML)*, 2022.
- [52] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PmlR, 2023, pp. 19 730–19 742.
- [53] Z. Yang, L. Jiang, Y. Sun, B. Schiele, and J. Jia, "A unified query-based paradigm for point cloud understanding," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 8541–8551.
- [54] X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Behl, J. Wang, L. Yuan *et al.*, "Generalized decoding for pixel, image, and language," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 15 116–15 127.
- [55] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser *et al.*, "Openscene: 3d scene understanding with open vocabularies," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 815–824.
- [56] R. Ding, J. Yang, C. Xue, W. Zhang, S. Bai, and X. Qi, "Pla: Language-driven open-vocabulary 3d scene understanding," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7010–7019.
- [57] Z. Yang, J. Chen, Z. Miao, W. Li, X. Zhu, and L. Zhang, "Deepinteraction: 3d object detection via modality interaction," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1992–2005, 2022.
- [58] J. Yan, Y. Liu, J. Sun, F. Jia, S. Li, T. Wang, and X. Zhang, "Cross modal transformer: Towards fast and robust 3d object detection," 2023. [Online]. Available: <https://arxiv.org/abs/2301.01283>
- [59] Y. Xie, C. Xu, M.-J. Rakotosaona, P. Rim, F. Tombari, K. Keutzer, M. Tomizuka, and W. Zhan, "Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 591–17 602.
- [60] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 8748–8763.
- [61] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 652–660.
- [62] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [63] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 540–557.
- [64] Y. Li, S. Si, G. Li, C.-J. Hsieh, and S. Bengio, "Learnable fourier features for multi-dimensional spatial positional encoding," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 15 816–15 829, 2021.
- [65] D. A. Roberts and L. R. Roberts, "Qr and lq decomposition matrix backpropagation algorithms for square, wide, and deep-real or complex-matrices and their software implementation," in *Proceedings of the 2024 International Conference on Algorithmic Differentiation (AD)*. SIAM, 2026, pp. 28–47.
- [66] T. D. Ngo, B.-S. Hua, and K. Nguyen, "Isbnnet: a 3d point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 550–13 559.
- [67] D. Cai, L. Zhao, J. Zhang, L. Sheng, and D. Xu, "3djeg: A unified framework for joint dense captioning and visual grounding on 3d point clouds," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [68] Z. Chen, R. Hu, X. Chen, M. Nießner, and A. X. Chang, "Unit3d: A unified transformer for 3d dense captioning and visual grounding," in *International Conference on Computer Vision (ICCV)*, 2023, pp. 18 109–18 119.
- [69] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020. [Online]. Available: <https://arxiv.org/abs/2005.14165>