# E3CoverNet: Progressive E(3)-aware 3D Backbone for Geometry-Sensitive Vision-Language Alignment

Yilong Guo
ylguo628@stu.suda.edu.cn
Soochow University
School of Computer Science & Technology
Suzhou, China

Jin-Hui Wu*
wujinhui@suda.edu.cn
Soochow University
School of Computer Science & Technology
Suzhou, China
Nanjing University
National Key Laboratory for Novel Software Technology
Nanjing, China

Hongzhe Liu
hzliu1019@stu.suda.edu.cn
Soochow University
School of Computer Science & Technology
Suzhou, China

Haixin Sun
linyu@imau.edu.cn
Inner Mongolia Agricultural University
School of Computer Science & Technology
Hohhot, China

Fanzhang Li*
lfzh@suda.edu.cn
Soochow University
School of Computer Science & Technology
Suzhou, China

## Abstract

3D vision-language understanding is central to multimedia research. Representative tasks such as visual grounding and text-to-shape retrieval require aligning 3D geometry with natural language that often depends on fine-grained geometric cues. Most existing 3D-language models build on generic backbones with limited geometric inductive bias, leaving them brittle to spatial transformations. Conversely, strictly invariant encoders risk collapsing the very spatial distinctions that geometry-sensitive queries depend on. We introduce Equivariant Covering Networks on E(3) (E3CoverNet), a 3D backbone designed to close this gap. Rather than directly modeling the full E(3) transformation space, E3CoverNet progressively builds equivariant representations along the subgroup chain $E(1) \subset E(2) \subset E(3)$. At each stage, a learned covering set discretizes the corresponding symmetry group, and a geometry-aware attention layer synchronously updates invariant point features and equivariant coordinates through a dual-path design. We plug E3CoverNet into standard 3D-language pipelines and evaluate on text-to-shape retrieval and 3D visual grounding benchmarks, where it outperforms competitive baselines across both tasks. A geometry-sensitive diagnostic further shows that E3CoverNet substantially improves robustness to spatial perturbations without sacrificing the semantic discrimination required for complex spatial queries.

*Corresponding authors.


*MM '26, Rio de Janeiro, Brazil*

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3836014

## CCS Concepts

• **Computing methodologies** → **Shape representations**.

## Keywords

3D Vision-Language Understanding; 3D Visual Grounding; Equivariant Neural Networks; Multimodal Alignment

**ACM Reference Format:**
Yilong Guo, Jin-Hui Wu, Hongzhe Liu, Haixin Sun, and Fanzhang Li. 2026. E3CoverNet: Progressive E(3)-aware 3D Backbone for Geometry-Sensitive Vision-Language Alignment. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3767308.3836014

## 1 Introduction

3D vision-language understanding has become a central problem in multimedia research, with applications in language-guided scene understanding, embodied interaction, robotic manipulation, and spatial reasoning. Representative tasks such as 3D visual grounding [1, 4] and text-to-shape retrieval [6, 28] require models to align raw unstructured 3D geometry with natural-language queries that often depend on fine-grained spatial cues. Compared with appearance-centric descriptions, many 3D-language expressions are inherently geometry-sensitive, involving relative position, orientation, shape configuration, and object-to-object relations [33], such as "the chair facing the window," "the lamp to the left of the sofa," or "an upside-down mug." For such cases, successful cross-modal alignment depends not only on semantic representation capacity, but also on whether the underlying 3D encoder provides a stable and informative geometric representation.

Although multimodal 3D understanding has advanced rapidly, driven in part by large-scale 3D-language pretraining [21, 26, 40],

3D large language models [7, 15, 39], and unified vision-language frameworks [17], 3D-language alignment remains fragile under geometric perturbations and coordinate variations. Many existing methods, including both task-specific pipelines and recent foundation-model-based approaches, still rely on generic 3D backbones whose geometric inductive biases are limited. As a result, learned representations may shift with viewpoint changes, object reorientation, or coordinate-frame transformations, weakening multimodal matching and grounding. Historically, equivariant architectures have shown that embedding transformation structure into 3D encoders can improve robustness and sample efficiency, particularly through groups such as SE(3) [8, 12, 29, 32], with continued advances in equivariant point-cloud modeling [11, 20]. Yet existing equivariant designs remain poorly matched to the needs of geometry-sensitive vision-language understanding: they are typically built for geometric perception tasks rather than cross-modal alignment, and directly modeling high-dimensional transformation spaces can introduce practical optimization and efficiency difficulties [12]. More fundamentally, a purely invariant encoder discards the spatial distinctions(left versus right, facing toward versus away)that geometry-sensitive language queries depend on, creating a tension between geometric robustness and semantic expressiveness that existing designs have not explicitly addressed in the multimodal setting.

These observations raise two questions. Can a more structured geometric prior improve multimodal alignment in 3D-language tasks, especially on queries that depend strongly on spatial structure? And can such a prior be introduced progressively, rather than through monolithic modeling of the full 3D transformation space? In this work, we address both questions by studying geometry-aware representation learning through the Euclidean group E(3), which includes translations, rotations, and reflections. We do not assume that all language semantics are invariant to every transformation in E(3); instead, we use E(3)-aware modeling as a structured inductive bias for learning geometry-sensitive 3D features and empirically examine its value for multimodal spatial understanding.

To this end, we propose Equivariant Covering Networks on E(3) (E3CoverNet), a geometry-sensitive 3D backbone for multimodal understanding. Rather than parameterizing the full E(3) transformation space in a single step, E3CoverNet adopts a progressive construction guided by the subgroup chain

$$\mathrm{E}(1) \subset \mathrm{E}(2) \subset \mathrm{E}(3),$$

so that geometric structure is introduced from low-dimensional to high-dimensional transformation spaces. Building on Lie group cover learning [18], our design incrementally expands the representation and injects transformation-aware geometric priors into the 3D encoder. This progressive construction offers a practical structured route to capturing increasingly rich geometric relationships while avoiding a monolithic treatment of E(3). It also allows geometric priors to be incorporated without assuming all language semantics are invariant to every spatial transformation.

The resulting E3CoverNet backbone improves the overall stability and quality of 3D representations under geometric perturbations while preserving sensitivity to spatial structure, rather than collapsing geometric variation into indiscriminate invariance. In summary, our main contributions are as follows:

- We propose E3CoverNet, a geometry-sensitive 3D backbone that injects structured E(3)-aware geometric priors, bridging unstructured point clouds and spatially complex language queries for multimodal alignment.
- We develop a progressive construction $\mathrm{E}(1) \subset \mathrm{E}(2) \subset \mathrm{E}(3)$ via Lie group covering, enabling computationally tractable modeling from low-dimensional to full 3D transformations.
- We identify a key distinction between geometric robustness and absolute spatial invariance in 3D-language alignment. Controlled perturbation tests and query-type analysis show that E3CoverNet improves stability under spatial transformations while preserving the semantic sensitivity that geometry-sensitive expressions require.

## 2 Related Work

To contextualize our proposed approach, this section reviews relevant literature across three highly intertwined domains: the evolution of 3D vision-language alignment methodologies (Section 2.1 and 2.2), and the development of geometry-aware 3D backbones (Section 2.3) which inspire our equivariant encoder design.

### 2.1 Visual Grounding and Vision-Language Alignment in 3D

3D visual grounding aims to localize the object referred to by a natural-language expression in a 3D scene, and has become a representative task for multimodal spatial understanding. Benchmarks such as ReferIt3D [1], including SR3D and NR3D, and ScanRefer [4], all built upon the ScanNet dataset [10], have enabled systematic study of grounding under complex scene geometry and free-form language descriptions. Recent surveys show that this area has rapidly evolved from early object-centric matching pipelines to more sophisticated frameworks that incorporate view reasoning, relation modeling, and stronger cross-modal alignment mechanisms.

Recent methods improve grounding through transformer-based relation modeling [16, 42], multi-view reasoning [14], geometric-semantic interaction [3, 34], and context-aware fusion [13]. In parallel, large-scale pretraining [17, 21, 40] and unified 3D vision-language frameworks [38, 44] have substantially raised the performance ceiling. Yet their underlying 3D encoders typically remain generic architectures without explicit geometric inductive bias. More recently, 3D large language models [7, 15, 39] have further raised performance ceilings by using LLM-based reasoning, but their 3D encoders are likewise inherited from standard point-cloud backbones without geometry-aware design.

Our work is complementary: instead of designing new cross-modal fusion strategies, scaling pretraining, or augmenting language reasoning capacity, we study how explicitly structured geometric modeling in the underlying 3D encoder can improve alignment quality, especially for geometry-sensitive queries.

### 2.2 Text-to-Shape Retrieval and Point-Language Representation Learning

Text-to-shape retrieval studies cross-modal matching between language and 3D shapes. Recent methods adopt contrastive learning and multimodal pretraining to improve alignment at scale [19, 21, 26–28, 40], improving retrieval quality. However, the underlying

3D representations still rely on generic encoders, leaving the role of structured geometric priors underexplored.

Rather than focusing on contrastive objectives or trimodal supervision, we investigate how a geometry-sensitive backbone can strengthen retrieval from the encoder side.

### 2.3 Geometry-aware 3D Backbones and Equivariant Representation Learning

Geometric deep learning [2] has shown that equivariance provides an effective inductive bias for structured representation learning [2, 8]. In 3D vision, prior work has explored geometry-aware architectures through point-based local aggregation [23, 24], local reference frames [25], spherical or tensor field constructions [9, 32], equivariant graph neural networks [12, 29], and point-cloud-specific equivariant operators [11]. In particular, SE(3)-equivariant and rotation-aware architectures, including EPN [5], E2PN [43], tensor field networks [32], and related equivariant graph models [12, 29], demonstrate that embedding transformation structure into 3D encoders can improve robustness and generalization. Recent work continues to advance equivariant point-cloud modeling: Vector Neurons [11] extend SO(3)-equivariance to standard point-cloud architectures through a lightweight neuron-level design, PointNeXt [25] revisits classical point-based methods with modern training strategies to establish strong non-equivariant baselines, and EquiformerV2 [20] pushes equivariant transformer architectures to new performance levels on molecular and 3D benchmarks.

However, these methods target geometric perception tasks rather than multimodal 3D-language understanding, and directly modeling rich transformation spaces in a single step poses practical optimization challenges. Our work differs by studying geometry-aware modeling specifically for vision-language alignment through a progressive subgroup chain $\mathrm{E}(1) \subset \mathrm{E}(2) \subset \mathrm{E}(3)$.

## 3 Method

In this section, we introduce the architecture and computational mechanism of E3CoverNet. Following a brief overview (Section 3.1), we detail how the aforementioned subgroup chain is operationalized to ensure computationally tractable transformation modeling (Section 3.2). We then formulate the stage-wise attention layer that powers this progressive expansion by synchronously updating invariant features and equivariant coordinates (Section 3.3), culminating in the multimodal adaptation modules for text-to-shape retrieval and visual grounding (Section 3.4).

### 3.1 Overview

Given a point cloud $P = \{\mathbf{p}_i\}_{i=1}^{N}$ and a language query $L$, our goal is to learn a 3D representation that supports geometry-sensitive cross-modal alignment. As illustrated in Figure 1, E3CoverNet follows a progressive design guided by the subgroup chain $\mathrm{E}(1) \subset \mathrm{E}(2) \subset \mathrm{E}(3)$ (where $\mathrm{E}(n) = \mathbb{R}^n \rtimes \mathrm{O}(n)$), gradually expanding from lower-dimensional geometric structure toward full $\mathrm{E}(3)$-aware modeling. At each stage, a geometry-aware attention layer jointly updates point features and coordinates with explicit awareness of spatial transformations. The method consists of three components: a progressive backbone construction (§3.2), a geometry-aware attention layer (§3.3), and a task-agnostic multimodal adaptation (§3.4).

### 3.2 Progressive $\mathrm{E}(3)$-aware Backbone

Directly modeling a rich transformation space such as $\mathrm{E}(3)$ can be particularly difficult in practice due to both optimization complexity and the large number of coupled geometric degrees of freedom involved. To address this, E3CoverNet adopts a progressive construction that expands geometric support stage by stage.

Let $C_n = \{g_i^{(n)}\}_{i=1}^{K_n} \subset \mathrm{E}(n)$ denote a learned covering set at stage $n$, where $K_n$ is the number of covering elements. Rather than directly parameterizing the continuous transformation space of $\mathrm{E}(n)$, we use these $K_n$ elements as a finite set of learnable geometric anchors that discretize the local support of the stage-specific transformation space. Each stage thus captures representative geometric variation through a tractable covering of $\mathrm{E}(n)$ while retaining the transformation-aware structure needed for downstream alignment. To connect adjacent stages, we use the natural embedding

$$\iota_{n\to n+1} : \mathrm{E}(n) \hookrightarrow \mathrm{E}(n+1), \tag{1}$$

which lifts the learned geometric support from a lower-dimensional group into the next stage of the hierarchy.

Given an intermediate representation $f_n(x)$ at stage $n$, the representation at stage $n+1$ is constructed as

$$f_{n+1}(x) = \iota_{n\to n+1}(f_n(x)) + \delta_{n+1}(x), \tag{2}$$

where $\delta_{n+1}(x)$ captures the additional geometric variation introduced at the higher-dimensional stage. Stage identity is thus reflected not only in feature depth but also in the dimensional expansion of the geometric support: lower-dimensional priors are preserved through $\iota_{n\to n+1}$, while new transformation-aware variation is incrementally introduced through $\delta_{n+1}(x)$. By carrying forward lower-dimensional geometric priors at each transition, E3CoverNet progressively expands toward full 3D Euclidean modeling without discarding previously learned structure.

In practice, the backbone contains three stages corresponding to $\mathrm{E}(1)$, $\mathrm{E}(2)$, and $\mathrm{E}(3)$, with covering-set sizes $K_1$=3, $K_2$=8, and $K_3$=10 respectively. The first stage captures simple low-dimensional geometric variation, the second stage enriches planar structural modeling, and the final stage produces full 3D geometry-aware features. This staged design provides a structured alternative to monolithic transformation modeling and makes it easier to inject geometric bias into the encoder progressively.

### 3.3 Geometry-aware Attention Layer

Each stage of E3CoverNet is implemented with a geometry-aware attention layer that jointly updates invariant feature channels and point coordinates. Let the input feature matrix be $H \in \mathbb{R}^{N\times d_h}$ and the point coordinates be $Z = \{\mathbf{z}_i\}_{i=1}^{N}$ with $\mathbf{z}_i \in \mathbb{R}^3$.

For points $(i, j)$, we compute the relative displacement

$$\mathbf{r}_{ij} = \mathbf{z}_j - \mathbf{z}_i, \tag{3}$$

and the Euclidean distance

$$d_{ij} = \|\mathbf{r}_{ij}\|_2. \tag{4}$$

The scalar distance is encoded by a radial basis function $\phi(d_{ij})$ [30], which provides geometry-aware but transformation-stable inputs to the invariant feature stream described below.

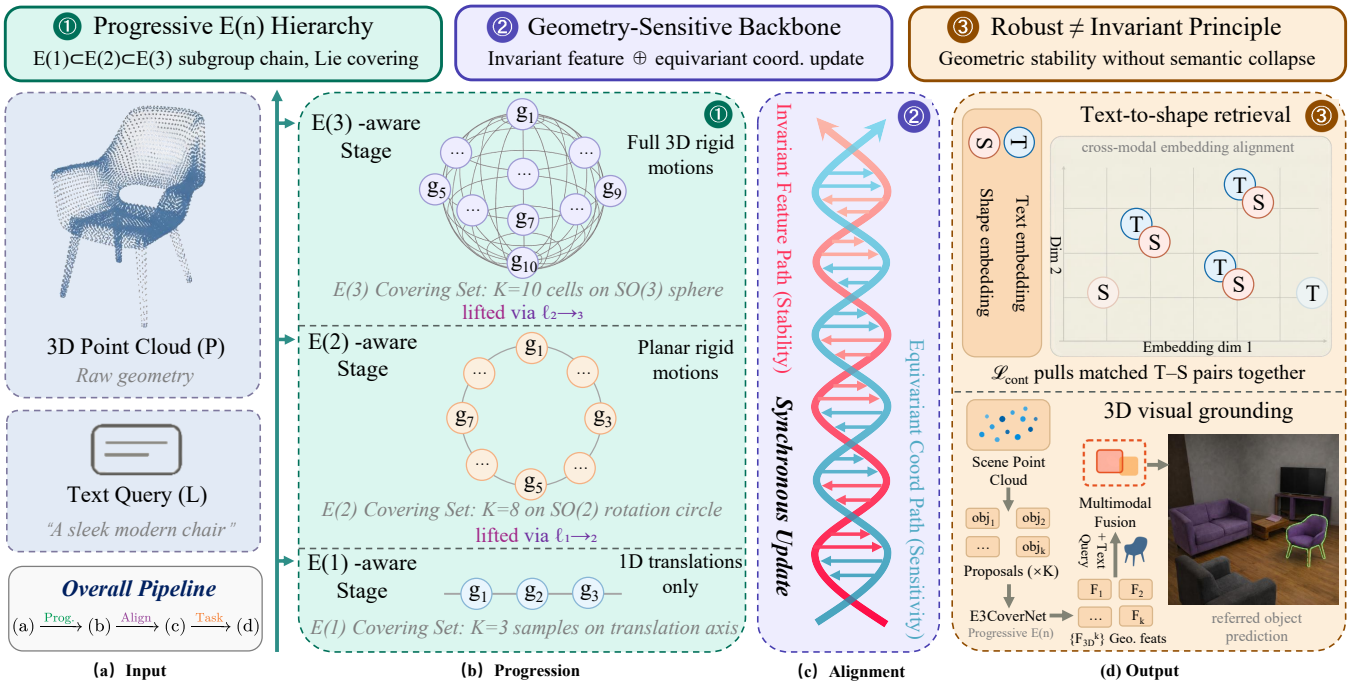


**Figure 1: E3CoverNet progressively builds geometry-sensitive 3D representations via an $\mathrm{E}(n)$ subgroup hierarchy for vision-language tasks. (a) The model takes a point cloud $P$ and a text query $L$ as inputs. (b) Following the Progressive $\mathrm{E}(n)$ Hierarchy (①), the backbone traverses three stages guided by $\mathrm{E}(1) \subset \mathrm{E}(2) \subset \mathrm{E}(3)$. Each stage constructs a *covering set* that discretises the corresponding symmetry group: $K$=3 samples on the translation axis, $K$=8 cells on the $\mathrm{SO}(2)$ circle, and $K$=10 cells on the $\mathrm{O}(3)$ sphere (the rotational component of $\mathrm{E}(3)$, including reflections), progressively expanding from 1D translations to full 3D Euclidean transformations. (c) Within the Geometry-Sensitive Backbone (②), an *invariant feature path* and an *equivariant coordinate path* are updated synchronously at every stage (dual-helix structure). (d) The resulting representation is applied to text-to-shape retrieval via contrastive alignment ($\mathcal{L}_{\text{cont}}$, top) and 3D visual grounding via multimodal fusion (bottom). By decoupling stability from sensitivity (③: Robust ≠ Invariant), E3CoverNet ensures geometric quality without semantic collapse.**

*Invariant Feature Update.* We first construct pairwise messages based on neighboring features and distance encoding:

$$m_{ij} = \mathrm{MLP}_m(h_j, \phi(d_{ij})), \tag{5}$$

$$\alpha_{ij} = \mathrm{softmax}_j\Big(\mathrm{MLP}_{attn}(h_i, h_j, \phi(d_{ij}))\Big). \tag{6}$$

The feature update is then given by

$$H' = H + \mathrm{Linear}\left(\sum_j \alpha_{ij} \cdot m_{ij}\right). \tag{7}$$

Since this branch depends on invariant geometric quantities such as inter-point pairwise distances and radial encodings, it provides stable feature interaction under Euclidean transformations.

*Equivariant Coordinate Update.* To preserve geometric structure in the coordinate stream, we update coordinates using relative displacement vectors modulated by invariant scalar gates:

$$w_{ij}^h = \mathrm{MLP}_w^h(h_j, \phi(d_{ij})), \tag{8}$$

$$\Delta \mathbf{z}_i = \sum_{h=1}^{H_a} \sum_j \alpha_{ij}^h \cdot w_{ij}^h \cdot \mathbf{r}_{ij}, \tag{9}$$

$$Z' = Z + \Delta Z, \tag{10}$$

where $H_a$ is the total number of attention heads. Because the coordinate update is entirely formed from relative displacement vectors and scalar coefficients, the coordinate stream is designed to transform consistently under Euclidean transformations.

*Stage-Wise Instantiation.* The same layer template is used across stages, while the stage identity determines how geometric priors are organized and expanded through the subgroup hierarchy. The subgroup chain controls the progressive expansion of the backbone, and the geometry-aware attention layer serves as the computational unit that realizes this construction in practice.

Compared with directly parameterizing a large transformation space in a single block, this design is easier to optimize and naturally aligned with staged geometric modeling. If fully connected neighborhoods are used, the complexity of a backbone layer is

$$O\left(N^2 \sum_{n=1}^{3} d_n\right), \tag{11}$$

where $d_n$ denotes the stage-dependent geometric dimensionality. In practice, complexity can be reduced by using sparse neighborhoods.

*Unified Forward Procedure.* To clearly consolidate the overall progressive backbone construction (§3.2) and the geometry-aware

**Algorithm 1** Forward Pass of E3CoverNet Backbone

**Input:** Covering sizes ($K_1$=3, $K_2$=8, $K_3$=10); point cloud $P = \{\mathbf{p}_i\}_{i=1}^{N}$
**Output:** Geometry-sensitive 3D feature matrix $H^{(3)}$
1: $Z^{(0)} \leftarrow P$; $H^{(0)} \leftarrow \text{Embed}(P)$
2: **for** stage $n = 1, 2, 3$ **do**
3: *% Covering-set construction (§3.2)*
4: Obtain learned covering set $C_n = \{g_i^{(n)}\}_{i=1}^{K_n} \subset \mathrm{E}(n)$
5: **if** $n > 1$ **then**
6: Lift representation via $\iota_{n-1\to n}$ and add stage-specific variation $\delta_n$ *// Eq. (2)*
7: **end if**
8: *% Geometry-aware attention (§3.3)*
9: Compute pairwise $\mathbf{r}_{ij}$, $d_{ij}$, $\phi(d_{ij})$ from $Z^{(n-1)}$ *// Eqs. (3)–(4)*
10: *% Invariant feature path*
11: Compute messages $m_{ij}$ and attention weights $\alpha_{ij}$ *// Eqs. (5)–(6)*
12: $H^{(n)} \leftarrow H^{(n-1)} + \text{Linear}\big(\sum_j \alpha_{ij} \cdot m_{ij}\big)$ *// Eq. (7)*
13: *% Equivariant coordinate path*
14: Compute scalar gates $w_{ij}^h$ *// Eq. (8)*
15: $Z^{(n)} \leftarrow Z^{(n-1)} + \sum_h \sum_j \alpha_{ij}^h \cdot w_{ij}^h \cdot \mathbf{r}_{ij}$ *// Eqs. (9)–(10)*
16: **end for**
17: **return** $H^{(3)}$ *// backbone output for downstream tasks*

attention layer described above, Algorithm 1 summarizes the complete forward pass of the E3CoverNet backbone.

As illustrated in Figure 1, Algorithm 1 realizes a three-stage forward recursion rather than a monolithic E(3) block. At each stage, the covering set $C_n$ discretizes the stage-specific symmetry group, and the inter-stage lifting $\iota_{n-1\to n}$ (line 6) ensures that lower-dimensional geometric priors are preserved rather than discarded. Within each stage, the invariant feature path (lines 10–11) and the equivariant coordinate path (lines 13–14) are updated synchronously, realizing the dual-helix structure depicted in Figure 1 (c). Because all scalar gates and attention weights are derived from invariant quantities (pairwise distances), the coordinate stream maintains transformation-consistent behavior throughout the hierarchy, embodying the "Robust ≠ Invariant" principle central to our design.

## 3.4 Multimodal Adaptation

E3CoverNet serves as a general-purpose plug-in 3D backbone for standard vision-language pipelines.

*Text-to-Shape Retrieval.* The backbone output $H^{(3)}$ is aggregated via global average pooling to produce a shape embedding $F_{\text{3D}} \in \mathbb{R}^D$, aligned with the text embedding $F_{\text{text}} \in \mathbb{R}^D$ from a frozen language encoder via a symmetric InfoNCE loss [22, 26]:

$$\mathcal{L}_{\text{cont}} = -\frac{1}{2}\left(\mathbb{E}\left[\log \frac{e^{s_{ii}/\tau}}{\sum_j e^{s_{ij}/\tau}}\right] + \mathbb{E}\left[\log \frac{e^{s_{ii}/\tau}}{\sum_j e^{s_{ji}/\tau}}\right]\right), \tag{12}$$

where $s_{ij} = F_{\text{3D},i}^{\top} F_{\text{text},j}$ and $\tau$ is a learnable temperature.

*3D Visual Grounding.* For each candidate object $k$, E3CoverNet produces a feature $F_{\text{3D}}^{(k)}$. The set $\{F_{\text{3D}}^{(k)}\}_{k=1}^{K}$ is fused with language tokens through a cross-attention module, where text queries attend over geometry-sensitive 3D features for referred-object prediction.

# 4 Experiments

We evaluate E3CoverNet as a geometry-sensitive 3D backbone for vision-language alignment. Our experiments examine whether E3CoverNet improves performance on standard 3D vision-language tasks, whether the gains remain competitive against representative task-specific methods, whether the learned representation is more robust under spatial perturbations and on geometry-sensitive queries, and whether the progressive subgroup construction contributes to the observed improvements.

## 4.1 Experimental Setup

*Tasks and Datasets.* We evaluate E3CoverNet on two 3D vision-language tasks: text-to-shape retrieval and 3D visual grounding.

For text-to-shape retrieval, we use the Text2Shape benchmark [6]. Given a natural-language description, the goal is to retrieve the most relevant 3D shape from a candidate set.

For 3D visual grounding, we use the ReferIt3D benchmark [1], including both SR3D and NR3D. SR3D contains template-based spatial descriptions, while NR3D contains free-form human utterances. These datasets are suitable for evaluating geometry-sensitive vision-language alignment because many queries depend on relative position, orientation, or spatial configuration.

*Comparison Protocol.* We consider two complementary settings: a *task-level* comparison against representative methods reported for each benchmark, and a *backbone-level* comparison in which only the 3D encoder is swapped within a unified multimodal pipeline (same language encoder, fusion module, loss, and schedule). The task-level setting positions E3CoverNet relative to full systems; the backbone-level setting isolates the encoder's contribution.

*Implementation Details.* For all tasks, input point clouds are normalized before feature extraction. All experiments are conducted on NVIDIA RTX 4070. The language encoder is kept identical across all compared backbone variants so that the comparison focuses on the contribution of the 3D representation. During training, we avoid aggressive transformation-heavy augmentation, since our goal is to test whether geometric robustness is provided by the model architecture itself rather than by augmentation.

For text-to-shape retrieval, E3CoverNet uses only text and 3D shape inputs. In particular, unlike trimodal retrieval methods such as TriCoLo, our model does not use an additional image modality. For 3D visual grounding, E3CoverNet is evaluated as a plug-in 3D encoder within a fixed grounding pipeline, without introducing task-specific modules beyond the backbone replacement. All results are averaged over multiple runs with the same evaluation protocol.

*Evaluation Metrics.* For text-to-shape retrieval, we report Recall @$K$ (R@1, R@5, and R@10). For 3D visual grounding, we report grounding accuracy on both the SR3D and NR3D splits following the standard evaluation protocols established in prior work [1].

**Table 1: Comparison with representative methods on Text2Shape retrieval. Methods marked with † use large-scale pretraining or image/multi-view supervision beyond the text-shape pair. E3CoverNet uses only text-3D inputs (bimodal). For a controlled backbone-only comparison, see Table 3.**

| Method | R@1 | R@5 | R@10 |
|---|---|---|---|
| *Task-specific methods* | | | |
| Text2Shape baseline [6] | 6.3 | 19.1 | 29.4 |
| Parts2Words [31] | 9.5 | 26.3 | 37.1 |
| TriCoLo† [28] | 13.2 | 34.1 | 45.8 |
| *Pretraining-based methods* | | | |
| ULIP† [40] | 15.6 | 38.4 | 50.2 |
| OpenShape† [21] | 18.3 | 43.5 | 55.8 |
| **E3CoverNet (Ours)** | 14.1 | 35.4 | 47.0 |

**Table 2: Comparison with representative methods on ReferIt3D grounding. ‡ denotes a unified pre-trained 3D vision-language system with large-scale pretraining data. E3CoverNet is evaluated as a plug-in backbone replacement without additional pretraining. For a controlled backbone-only comparison, see Table 3.**

| Method | SR3D Acc. | NR3D Acc. |
|---|---|---|
| *Task-specific grounding methods* | | |
| ReferIt3DNet [1] | 40.8 | 35.6 |
| ViewRefer [14] | 54.6 | 42.6 |
| G3-LQ [34] | 56.3 | 44.2 |
| DOrA [36] | 57.5 | 45.0 |
| CFA [13] | 58.4 | 45.9 |
| *Pre-trained unified framework* | | |
| 3D-VisTA‡ [44] | 64.2 | 51.4 |
| **E3CoverNet (Ours)** | 58.9 | 46.3 |

## 4.2 Main Results

*4.2.1 Comparison with Task-Specific Methods.* We first compare the performance of E3CoverNet with several representative and competitive task-specific methods on both benchmarks.

*Text-to-Shape Retrieval.* Table 1 reports the retrieval results on Text2Shape. E3CoverNet achieves the strongest performance among task-specific methods using only text and 3D inputs. Pretraining-based methods (ULIP, OpenShape) benefit from additional image supervision and massive paired data; integrating E3CoverNet into such frameworks is a natural future direction.

*3D Visual Grounding.* Table 2 reports the grounding results on ReferIt3D. Among methods without large-scale pretraining, E3Cover Net achieves the best performance on SR3D and NR3D. The gap to 3D-VisTA‡ reflects the difference between backbone-level improvement and system-level pretraining, a complementary direction.

*4.2.2 Comparison under a Unified Backbone-Swap Setting.* To isolate the contribution of the 3D encoder, we next compare E3CoverNet with both standard point-based and geometry-aware equivariant backbones in the same multimodal pipeline. In this setting, only the 3D backbone is changed, while the language encoder, fusion module, loss, and training schedule remain fixed.

Table 3 reports the backbone-swap comparison. With all other components fixed, E3CoverNet consistently outperforms both standard point-based encoders, including the recent PTv3 [37], and geometry-aware backbones (EPN, E2PN, EquiformerV2), providing direct evidence that the progressive E(3)-aware construction contributes meaningfully to multimodal alignment across both retrieval and grounding tasks. In particular, E3CoverNet outperforms EquiformerV2 [20], the strongest monolithic equivariant transformer in our comparison, at substantially lower computational cost (§4.6), confirming that progressive subgroup construction is more effective and efficient than directly modeling high-dimensional transformation spaces for vision-language alignment.

**Table 3: Controlled backbone-swap comparison under a unified multimodal framework. Only the 3D encoder is changed; the language encoder, fusion module, loss function, and training schedule are kept identical across all rows. Absolute scores are lower than those in Tables 1–2 because the unified framework strips task-specific optimisations to ensure a fair backbone-only comparison.**

| 3D Backbone | Retrieval | | | Grounding | |
|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | SR3D | NR3D |
| *Standard point-cloud backbones* | | | | | |
| PointNet++ [24] | 8.8 | 24.5 | 34.2 | 48.5 | 38.4 |
| DGCNN [35] | 9.6 | 26.1 | 36.5 | 50.1 | 41.5 |
| PointTransformer [41] | 11.2 | 29.2 | 40.1 | 52.8 | 41.2 |
| PointNeXt [25] | 11.7 | 30.5 | 39.6 | 53.5 | 42.0 |
| PTv3 [37] | 12.0 | 31.2 | 42.0 | 54.8 | 42.8 |
| *Geometry-aware / equivariant backbones* | | | | | |
| EPN [5] | 12.1 | 31.4 | 42.8 | 54.6 | 42.9 |
| E2PN [43] | 12.6 | 32.6 | 44.1 | 55.4 | 43.7 |
| EquiformerV2 [20] | 13.0 | 33.4 | 44.8 | 56.2 | 44.1 |
| **E3CoverNet (Ours)** | **13.4** | **34.0** | **45.6** | **56.9** | **44.8** |

## 4.3 Robustness under Spatial Perturbations

We evaluate the geometric stability of the learned representation through controlled perturbation tests at inference time. For text-to-shape retrieval, we apply rotations, reflections, and coordinate-frame perturbations to the input shapes without retraining. For 3D visual grounding, we perturb candidate object point clouds while keeping the grounding pipeline unchanged.

We do not interpret these tests as assuming that every language query is invariant to every geometric transformation. In particular, expressions involving left/right, front/back, or orientation-specific references may not preserve their intended semantics under arbitrary spatial transformations. Our goal is instead to examine whether E3CoverNet produces more stable geometry-sensitive 3D features under controlled spatial shifts of the input geometry.

Figure 2 illustrates the rotation-based slice of this study on SR3D, comparing different backbones as the perturbation angle increases. E3CoverNet degrades more gracefully than competing backbones as perturbation strength increases, indicating that the proposed

backbone learns a more stable geometric representation rather than relying mainly on canonical input orientation. This controlled stress test supports our central claim that E3CoverNet improves geometric robustness without collapsing orientation-sensitive spatial semantics into indiscriminate invariance.

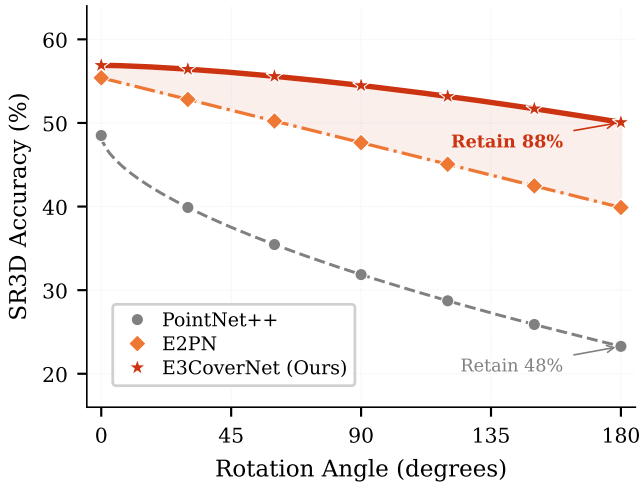


**Figure 2: Robustness under rotation on SR3D (backbone-swap setting). E3CoverNet still retains 88% of its initial accuracy at 180° (vs. 72% for E2PN, 48% for PointNet++). Shading clearly denotes the robustness gain over E2PN. Crucially, E3CoverNet degrades gracefully rather than remaining entirely flat, preserving the spatial sensitivity required for complex spatial queries ("Robust ≠ Invariant", §4.3).**

*Equivariance vs. Data Augmentation.* To disentangle the contribution of architectural equivariance from data augmentation, we compare E3CoverNet (trained on canonical orientations only) against PTv3 with and without SO(3) augmentation. All rows are tested on unseen rotation angles. As shown in Table 4, PTv3 without augmentation collapses to 28.1% at 180°; adding full SO(3) augmentation recovers it to only 34.0% (62% retention). E3CoverNet, without any augmentation, retains 50.1% (88% retention), confirming that structural equivariance provides stronger rotation robustness than finite-sample augmentation.

**Table 4: Equivariance vs. data augmentation under unseen rotations (SR3D Acc. %). "Aug" denotes training-time SO(3) rotation augmentation.**

| Backbone | 0° | 90° | 180° |
|---|---|---|---|
| PTv3 (No Aug) | 54.5 | 38.0 | 28.1 |
| PTv3 (w/ SO(3) Aug) | 54.8 | 45.2 | 34.0 |
| **E3CoverNet (No Aug)** | **56.9** | **53.4** | **50.1** |

*Robustness under Partial Observation.* Under random point dropping at test time on NR3D, E3CoverNet degrades gracefully: 44.8%→43.0%→40.6%→35.7% at 0/20/40/60% drop rates, versus PointNeXt's 42.0%→37.5%→32.2%→23.8%. At 40% drop, E3CoverNet loses only 4.2 points versus PointNeXt's 9.8, as the redundant covering anchors maintain partial group coverage when local regions are missing.

## 4.4 Ablation Studies

*4.4.1 Effect of Progressive Group Hierarchy.* To study the role of the progressive subgroup construction, we compare several variants of the proposed backbone: a non-equivariant version, lower-stage subgroup variants, an SE(3)-aware version without the full E(3) construction, and the complete E3CoverNet model.

Table 5 reports the results. The progressive hierarchy consistently improves performance, showing that staged geometric expansion from E(1) to E(3) is more effective than removing geometric structure or introducing it in a weaker form. Under a controlled iso-parameter (5.1M) comparison, three equivariant variants—monolithic full E(3) ($K$=30), two-stage skip E(1)→E(3), and the proposed progressive chain—achieve Text2Shape R@1 of 11.5, 12.1, and 13.4, respectively, confirming that staged decomposition simplifies optimization over high-dimensional covering spaces.

**Table 5: Ablation study on the subgroup hierarchy.**

| Model Variant | R@1 | R@5 | R@10 |
|---|---|---|---|
| Vanilla (No covering) | 10.6±0.3 | 28.9±0.4 | 39.5±0.5 |
| E(1) only | 11.3±0.3 | 30.2±0.4 | 41.4±0.4 |
| E(1) ⊂ E(2) | 12.4±0.2 | 32.5±0.3 | 43.7±0.4 |
| SE(3)-aware variant | 12.8±0.2 | 33.1±0.3 | 44.3±0.3 |
| **Full E(3) (E3CoverNet)** | **13.4**±0.2 | **34.0**±0.3 | **45.6**±0.3 |

*4.4.2 Effect of Geometry-aware Attention.* We further evaluate whether the gains come from the geometry-aware attention mechanism itself or only from the progressive hierarchy. We compare the full model with variants that remove coordinate updates, replace geometry-aware attention with standard self-attention, or remove the coupling between invariant feature interaction and equivariant coordinate propagation. The results are reported in Table 6.

**Table 6: Ablation on the geometry-aware attention layer.**

| Model Variant | SR3D Acc. | NR3D Acc. |
|---|---|---|
| Standard attention replacement | 54.7±0.3 | 42.8±0.4 |
| No coordinate update | 55.5±0.3 | 43.4±0.3 |
| No coupled geometric branch | 55.9±0.2 | 43.8±0.3 |
| **Full E3CoverNet** | **56.9**±0.2 | **44.8**±0.3 |

## 4.5 Analysis on Query Types

To better understand where the gains come from, we divide test queries into two groups: appearance-driven queries, which mainly rely on category or attribute cues, and geometry-sensitive queries, which involve spatial relations, directional language, or relative configuration. The split is constructed using rule-based annotations derived from relation and orientation keywords, and the detailed keyword list is provided in the supplementary material.

Table 7 shows that the improvement of E3CoverNet over E2PN is substantially larger on geometry-sensitive queries (+4.0%) than on appearance-driven ones (+0.4%), under the task-level grounding evaluation. This pattern is consistent with the motivation of our method: E3CoverNet is designed to strengthen 3D representations when multimodal alignment depends strongly on spatial configuration, orientation, or relative position. The smaller but still positive gain on appearance-driven queries suggests that the proposed backbone improves geometric sensitivity without sacrificing general representation quality. The larger gains on geometry-sensitive queries further support our claim that the progressive construction of E3CoverNet preserves spatial semantic distinctions that standard invariant backbones are more likely to blur or collapse.

*Failure Cases.* Error analysis reveals two recurring failure modes. First, in heavily occluded scenes where point-cloud segments are severely incomplete, the local pairwise distances $\phi(d_{ij})$ become unreliable. This disproportionately affects the E(3) stage, as the covering anchors in $C_3$ depend on stable local geometric contexts to estimate full 3D rigid motions, occasionally destabilising the equivariant coordinate updates. Second, queries involving multi-hop spatial reasoning (e.g., "the box on the desk next to the chair facing the window") remain challenging. While E3CoverNet provides robust low-level spatial features, resolving such compositional logic requires explicit sequential reasoning over multiple reference objects, a capacity bounded by the downstream fusion module rather than by geometric encoding. These observations suggest that improving geometric robustness alongside compositional relational reasoning is a natural direction for future work.

**Table 7: SR3D grounding accuracy breakdown by query type in the task-level grounding pipeline (Table 2 setting). Queries are split into 55% appearance-driven and 45% geometry-sensitive subsets based on spatial-relation and orientation keywords (see supplementary for the full keyword list). We compare E3CoverNet with E2PN under the same evaluation.**

| Query Type | E2PN | E3CoverNet |
|---|---|---|
| Appearance-driven (55%) | 58.2 | 58.6 |
| Geometry-sensitive (45%) | 55.3 | **59.3** |
| **Overall** | 56.9 | 58.9 |

## 4.6 Efficiency Analysis

Since E3CoverNet introduces additional geometric structure into the 3D backbone, it is worth quantifying the associated computational cost. We compare model size, floating-point operations, and inference time across backbone variants. As shown in Table 8, E3CoverNet incurs moderate computational overhead compared with E2PN while remaining well within real-time requirements. EquiformerV2 requires roughly 3× the parameters and 2× the inference latency of E3CoverNet yet achieves lower accuracy across all metrics (Table 3), indicating that the progressive subgroup construction attains a more favorable accuracy–efficiency trade-off than monolithic high-dimensional equivariant modeling. E3CoverNet also surpasses PTv3 with under half the parameters, suggesting that geometric priors can compensate for encoder scale. Regarding $K_3$ sensitivity, NR3D accuracy varies ≤0.6% across $K_3 \in \{6, 8, 10, 12\}$, indicating that once the covering set is moderately dense, learned anchors redistribute without requiring careful tuning.

**Table 8: Efficiency comparison of different 3D backbones.**

| Backbone | Params | FLOPs | Latency |
|---|---|---|---|
| *Non-equivariant backbones* | | | |
| PointNet++ | 1.5M | 3.8G | 12.3ms |
| DGCNN | 1.8M | 4.6G | 15.7ms |
| PointTransformer | 7.8M | 12.5G | 28.6ms |
| PointNeXt | 3.9M | 6.8G | 17.5ms |
| PTv3 | 11.2M | 15.6G | 31.4ms |
| *Equivariant backbones* | | | |
| EPN | 3.5M | 7.4G | 20.3ms |
| E2PN | 4.2M | 8.6G | 22.1ms |
| EquiformerV2 | 14.8M | 24.3G | 52.6ms |
| **E3CoverNet** | **5.1M** | **9.8G** | **24.7ms** |

E3CoverNet is also compatible with pretraining-heavy pipelines: replacing only the 3D encoder in ULIP [40] (frozen text/image encoders, same objective, 20 epochs) raises Text2Shape R@1 from 15.63 to 16.87, confirming that geometric priors and data-driven pretraining are complementary. Preliminary evaluation on ScanRefer likewise shows a gain (41.7% vs. PTv3's 39.8% Acc@0.5 under the same backbone-swap protocol).

## 5 Conclusion

We presented E3CoverNet, a progressive E(3)-aware 3D backbone that introduces structured geometric priors into vision-language alignment through a subgroup chain $\mathrm{E}(1) \subset \mathrm{E}(2) \subset \mathrm{E}(3)$. Experiments on text-to-shape retrieval and 3D visual grounding show that E3CoverNet improves multimodal alignment across both task-level and backbone-level evaluations. Perturbation robustness testing and query analysis further jointly confirm that improving geometric robustness does not require, and should not entail, collapsing all spatial structure into absolute invariance. Instead, structured E(3)-aware modeling can stabilize 3D representations under nuisance spatial shifts while preserving the semantic sensitivity needed for orientation-dependent language expressions. The main remaining limitations are smaller gains on appearance-dominated queries and the assumption of a single fixed language encoder. The current framework assumes rigid E(3) transformations; extending to deformable or articulated objects will require representations beyond the present group-theoretic scope, while multi-hop compositional spatial reasoning demands stronger downstream reasoning capacity. Combining geometry-sensitive backbones with large-scale multimodal pretraining and extending them to non-rigid and embodied settings is a natural next step.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (62176172,61672364) and the National Key Research and Development Program of China (No.2018YFA0701700 No.2018YFA0701701).

## References


[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. 2020. ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes. In *European Conference on Computer Vision*. Springer, 422–440.

[2] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. 2021. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv preprint arXiv:2104.13478* (2021).

[3] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14455–14465.

[4] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. 2020. ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language. In *European Conference on Computer Vision*. Springer, 202–221.

[5] Haiwei Chen, Shichen Liu, Weikai Chen, Hao Li, and Randall Hill. 2021. Equivariant Point Network for 3D Point Cloud Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14514–14523.

[6] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. 2018. Text2Shape: Generating Shapes from Natural Language by Learning Joint Embeddings. In *Asian Conference on Computer Vision*. Springer, 100–116.

[7] Sijin Chen et al. 2024. LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26428–26438.

[8] Taco Cohen and Max Welling. 2016. Group Equivariant Convolutional Networks. In *International Conference on Machine Learning*. PMLR, 2990–2999.

[9] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. 2018. Spherical CNNs. *arXiv preprint arXiv:1801.10130* (2018).

[10] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5828–5839.

[11] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. 2021. Vector Neurons: A General Framework for SO(3)-Equivariant Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12200–12209.

[12] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. 2020. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. *Advances in Neural Information Processing Systems* 33 (2020), 1970–1981.

[13] Peng Guo, Hongyuan Zhu, Hancheng Ye, Taihao Li, and Tao Chen. 2024. Revisiting 3D visual grounding with Context-aware Feature Aggregation. *Neurocomputing* 601 (2024), 128195.

[14] Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. 2023. ViewRefer: Grasp the Multi-View Knowledge for 3D Visual Grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15372–15383.

[15] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3D-LLM: Injecting the 3D World into Large Language Models. *Advances in Neural Information Processing Systems* 36 (2023), 20482–20494.

[16] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. 2022. Bottom Up Top Down Detection Transformers for Language Grounding in Images and Point Clouds. In *European Conference on Computer Vision*. Springer, 417–433.

[17] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. 2024. SceneVerse: Scaling 3D Vision-Language Learning for Grounded Scene Understanding. In *European Conference on Computer Vision*. Springer, 289–310.

[18] Fanzhang Li, Li Zhang, and Zhao Zhang. 2018. *Lie Group Machine Learning*. Walter de Gruyter GmbH & Co KG.

[19] Xiong Li, Yikang Yan, Zhenyu Wen, Qin Yuan, Fangda Guo, Zhen Hong, and Ye Yuan. 2025. Open3DSearch: Zero-Shot Precise Retrieval of 3D Shapes Using Text Descriptions. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6183–6192.

[20] Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. 2023. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. *arXiv preprint arXiv:2306.12059* (2023).

[21] Minghua Liu et al. 2023. OpenShape: Scaling Up 3D Shape Representation Towards Open-World Understanding. *Advances in Neural Information Processing Systems* 36 (2023), 44860–44879.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748* (2018).

[23] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 652–660.

[24] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *Advances in Neural Information Processing Systems* 30 (2017).

[25] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. 2022. PointNeXt: Revisiting PointNet++ with Improved Training and Scaling Strategies. *Advances in Neural Information Processing Systems* 35 (2022), 23192–23204.

[26] Alec Radford et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*. PmLR, 8748–8763.

[27] Junlong Ren, Hao Wu, Hui Xiong, and Hao Wang. 2025. SCA3D: Enhancing Cross-modal 3D Retrieval via 3D Shape and Caption Paired Data Augmentation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9550–9557.

[28] Yue Ruan, Han-Hung Lee, Yiming Zhang, Ke Zhang, and Angel X Chang. 2024. TriCoLo: Trimodal Contrastive Loss for Text to Shape Retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5815–5825.

[29] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. 2021. E(n) Equivariant Graph Neural Networks. In *International Conference on Machine Learning*. PMLR, 9323–9332.

[30] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. 2017. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems* 30 (2017).

[31] Chuan Tang, Xi Yang, Bojian Wu, Zhizhong Han, and Yi Chang. 2023. Parts2Words: Learning Joint Embedding of Point Clouds and Texts by Bidirectional Matching Between Parts and Words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6884–6893.

[32] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. 2018. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. *arXiv preprint arXiv:1802.08219* (2018).

[33] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. 2020. Learning 3D Semantic Scene Graphs from 3D Indoor Reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3961–3970.

[34] Yuan Wang, Yali Li, and Shengjin Wang. 2024. $G^3$-LQ: Marrying Hyperbolic Alignment with Explicit Semantic-Geometric Modeling for 3D Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13917–13926.

[35] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. 2019. Dynamic Graph CNN for Learning on Point Clouds. *ACM Transactions on Graphics (tog)* 38, 5 (2019), 1–12.

[36] Tung-Yu Wu, Sheng-Yu Huang, and Yu-Chiang Frank Wang. 2025. Data-Efficient 3D Visual Grounding via Order-Aware Referring. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 3107–3117.

[37] Xiaoyang Wu et al. 2024. Point Transformer V3: Simpler, Faster, Stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4840–4851.

[38] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. 2023. EDA: Explicit Text-Decoupling and Dense Alignment for 3D Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19231–19242.

[39] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2024. PointLLM: Empowering Large Language Models to Understand Point Clouds. In *European Conference on Computer Vision*. Springer, 131–147.

[40] Le Xue et al. 2023. ULIP: Learning a Unified Representation of Language, Images, and Point Clouds for 3D Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1179–1189.

[41] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. 2021. Point Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16259–16268.

[42] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 2021. 3DVG-Transformer: Relation Modeling for Visual Grounding on Point Clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2928–2937.

[43] Minghan Zhu, Maani Ghaffari, William A Clark, and Huei Peng. 2023. E2PN: Efficient SE(3)-Equivariant Point Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1223–1232.

[44] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 2023. 3D-VisTA: Pre-trained Transformer for 3D Vision and Text Alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2911–2921.