SpecBridge: Spectral Structure Alignment and Transitive Bridging for 3D-2D-Text Pre-Training

Dong Wang[1,2], **Jie Jiang**[1,2,*], **Weidong Min**[3], **Lixin Zhan**[1,2], **Xinpeng Zhao**[1,2] and **Ze Zhang**[4]
[1]College of Systems Engineering, National University of Defense Technology
[2]Laboratory for Big Data and Decision, National University of Defense Technology
[3]School of Mathematics and Computer Science, Nanchang University
[4]Institute of Medical Artificial Intelligence, South China Hospital, Medical School, Shenzhen University
{dongwang25, jiejiang, zhaoxinpeng}@nudt.edu.cn, minweidong@ncu.edu.cn,
zhanlixin98@outlook.com, zhangze21@mails.ucas.ac.cn

Abstract

Open-vocabulary 3D understanding aims to align 3D representations with a unified vision-language semantic space. However, existing methods suffer from the challenge of structural asymmetry caused by sparse observations and holistic geometries. Additionally, the inherent semantic chasm between discrete coordinates and abstract natural language hinders cross-modal mapping. This work introduces SpecBridge, a 3D-2D-Text pre-training framework that leverages CLIP priors as a foundational bridge to connect three modalities by synergizing spectral graph theory with transitive semantic learning. The method consists of two main sub-modules. First, Spectral Eigen-Modal Alignment is presented to construct cross-modal correlations within the intrinsic eigen-spectral space via Laplacian eigendecomposition. By aligning low-frequency geometric harmonics with CLIP-guided observation inference, it maintains structural consistency between 3D geometries and 2D projections. Second, a Transitive Spectral-Semantic Alignment method is developed to establish a 3D→2D→Text propagation chain across the CLIP priors bridge. It distills dense CLIP priors into 3D representations through transitive distillation, effectively mitigating the semantic chasm between geometry and language. Extensive experiments confirm that the proposed SpecBridge demonstrates state-of-the-art performance by overcoming challenges triggered by modality discrepancies.

1 Introduction

3D understanding tasks [Jain *et al.*, 2025; Zha *et al.*, 2025] aim to decipher geometric shapes and semantic characteristics within physical environments, serving as a critical foundation for autonomous driving [Yang *et al.*, 2025a], virtual reality [Jiang *et al.*, 2024], and embodied agents [Yang *et al.*, 2025b]. Currently, the field is expanding beyond traditional closed-set classification toward open-vocabulary learning. Previous methods [Xue *et al.*, 2024; Tang *et al.*, 2025] primarily leverage the generalization capabilities of 2D Vision-Language Models (e.g., CLIP [Radford *et al.*, 2021]) via large-scale 3D-2D-Text triplet pre-training to project 3D representations into a pre-aligned feature space. However, these methods insufficiently explore the extensive knowledge priors of CLIP to resolve the intrinsic structural discrepancies and semantic heterogeneities between modalities. Therefore, the lack of deep cross-modal alignment limits the performance of open-vocabulary 3D understanding tasks.

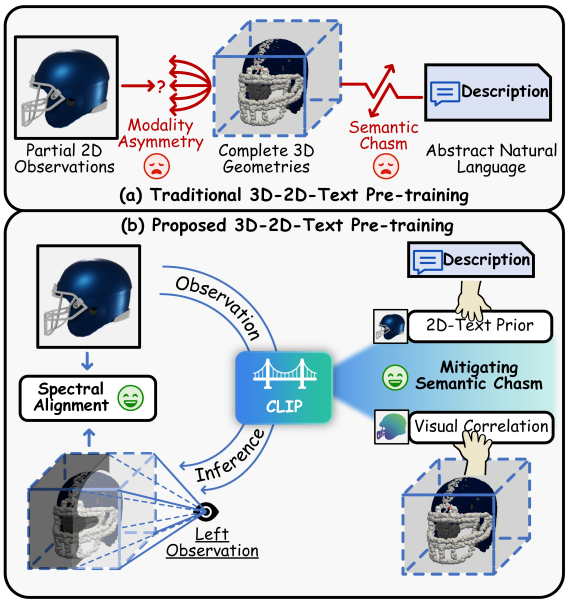

Figure 1: Comparison of (a) traditional and (b) proposed 3D-2D-Text pre-training paradigms.

The main challenge faced by existing 3D understanding methods is the inherent modality asymmetry between sparse, partial 2D observations and holistic 3D geometries, as shown on the left of Figure 1(a). Existing methods [Xue *et al.*, 2023; Liu *et al.*, 2023] commonly enforce rigid global alignment over one-to-one 2D-3D pairs, which induces optimization

*Corresponding author.

ambiguity and hinders the model from learning viewpoint-invariant representations. Although some advancements [Gao *et al.*, 2024] attempt to mitigate this through many-to-one joint learning, they remain constrained by linear aggregation in Euclidean feature space, essentially performing superficial appearance matching. Such learned representations remain sensitive to viewpoint changes and local noise, resulting in inaccuracy when capturing the geometric essence. Our motivation is to achieve robust, viewpoint-invariant geometric alignment by incorporating cross-modal correlations with CLIP-guided observation inference in the intrinsic eigen-spectral space, as illustrated on the left side of Figure 1(b).

The second challenge is that the profound semantic chasm between discrete 3D coordinates and abstract natural language exacerbates the complexity of modal interaction, as shown in Figure 1(a) right. Most works [Lee *et al.*, 2025; Zhai *et al.*, 2025] commonly lean toward expanding the scale of training data rather than refining the structural interaction mechanism. However, the supervised information obtained solely from coarse global concatenation is still poor, as it fails to extract fine-grained 2D-Text correlations inherent in CLIP. When the model is trained with such imprecise interaction patterns, the semantic mapping between 3D and language remains suboptimal. Inspired by the transitive nature of vision-language priors, this paper further utilizes a semantic propagation chain to mitigate the chasm between modalities, as illustrated on the right side of Figure 1(b).

To effectively tackle the above-mentioned challenges, the main contributions of the proposed SpecBridge framework are summarized as follows:

(i) Spectral Eigen-Modal Alignment (SEMA) is designed to introduce cross-modal alignment in the intrinsic eigen-spectral space. By combining Laplacian eigendecomposition with CLIP-guided observation inference, SEMA constructs low-frequency geometric harmonics to ensure structural consistency between 3D geometries and sparse 2D observations while effectively decoupling extrinsic viewpoint noise. (ii) Transitive Spectral-Semantic Alignment (TSSA) establishes a 3D→2D→Text semantic propagation chain to mitigate the semantic chasm. By exploiting the spectral affinity between 3D and 2D tokens, TSSA distills fine-grained CLIP priors into 3D representations through transitive soft supervision, achieving robust knowledge transfer throughout the pre-training process. (iii) Extensive experiments demonstrate that the proposed method can effectively integrate spectral geometric priors and linguistic knowledge into 3D models. SpecBridge achieves state-of-the-art performance on zero-shot classification and retrieval tasks, while providing interpretable alignment through fine-grained visualizations.

2 Related Work

2.1 Open-vocabulary 3D Understanding

Existing 3D understanding tasks focus on evolving toward open-vocabulary representation learning by aligning 3D shapes with 2D vision and language descriptions. As open-vocabulary representation learning is data-hungry, previous approaches including ULIP-2 [Xue *et al.*, 2024] and GPT4Point [Qi *et al.*, 2024b] employ automated annotation engines to build large-scale 3D-2D-Text datasets for establishing a diverse data foundation. By comparison, methods including LLaVA-3D [Zhu *et al.*, 2025] and ShapeLLM [Qi *et al.*, 2024a] are dedicated to representation learning paradigms, which capitalize on pre-trained 2D vision-language priors to empower large language models with 3D understanding capacity. However, these learned representations still suffer from the challenge of a significant gap with respect to the inherent complexity of 3D geometries. Recent methods such as Text4Point [Huang *et al.*, 2024] and Lidar-CLIP [Hess *et al.*, 2024] capitalize on the visual similarities between 2D and 3D modalities, utilizing 2D images as an intermediary to mitigate the difficulty of cross-modal learning. Nevertheless, these aforementioned methods rely on coarse global alignment and thus fail to preserve fine-grained 3D geometric structural information. To overcome these challenges, this paper proposes SpecBridge, which establishes cross-modal correlations in the intrinsic eigen-spectral space for more robust fine-grained cross-modal alignment.

2.2 Homogeneous Modalities Alignment

Homogeneous modality alignment aims to establish correlations between data with different modalities but similar information structures, such as the relationship between 2D and 3D modalities. Methods such as NCLR [Zhang *et al.*, 2025b] and MM-Point [Yu and Song, 2024] enforce the distillation of dense semantic priors from 2D foundation models into 3D encoders via knowledge distillation. In contrast, other approaches including Plain-PCQA [Chai *et al.*, 2024] and MM-PCQA+ [Zhang *et al.*, 2023] are inspired by the similar structural information across 2D and 3D modalities, and leverage sequences of 2D views to facilitate texture-aware feature extraction. Specialized tasks in I2PNet [Wang *et al.*, 2025] and HandDiff [Cheng *et al.*, 2024] further explore dense point-to-pixel correlations. However, Euclidean alignment remains susceptible to viewpoint variations and neglects the intrinsic manifold topology. Inspired by spectral analysis, SEMA is developed to capture viewpoint-invariant geometric harmonics in the intrinsic eigen-spectral space, thus ensuring structural consistency beyond superficial appearance matching.

2.3 Heterogeneous Modalities Alignment

Heterogeneous modality alignment attempts to associate data with similar semantics but vastly different structures, such as discrete 3D coordinates and abstract natural language. Recent works focus on mitigating the semantic chasm by integrating 3D encoders into Large Language Models. Typical methods, including PointLLM-V2 [Xu *et al.*, 2025] and MeshLLM [Fang *et al.*, 2025], fuse hierarchical 3D tokens with linguistic embeddings to enable context-aware reasoning. To circumvent the scarcity of paired data, GreenPLM [Tang *et al.*, 2025] leverages massive text data to compensate for limited 3D samples. However, Beacon3D [Huang *et al.*, 2025] reveals that direct 3D-vision-language alignment provides limited meaningful grounding coherence due to the significant modality chasm. While ImOV3D [Yang *et al.*, 2024] has explored an auxiliary strategy that employs 2D modalities as an intermediary, it lacks a structured semantic propagation mechanism, making it difficult to achieve accurate mapping

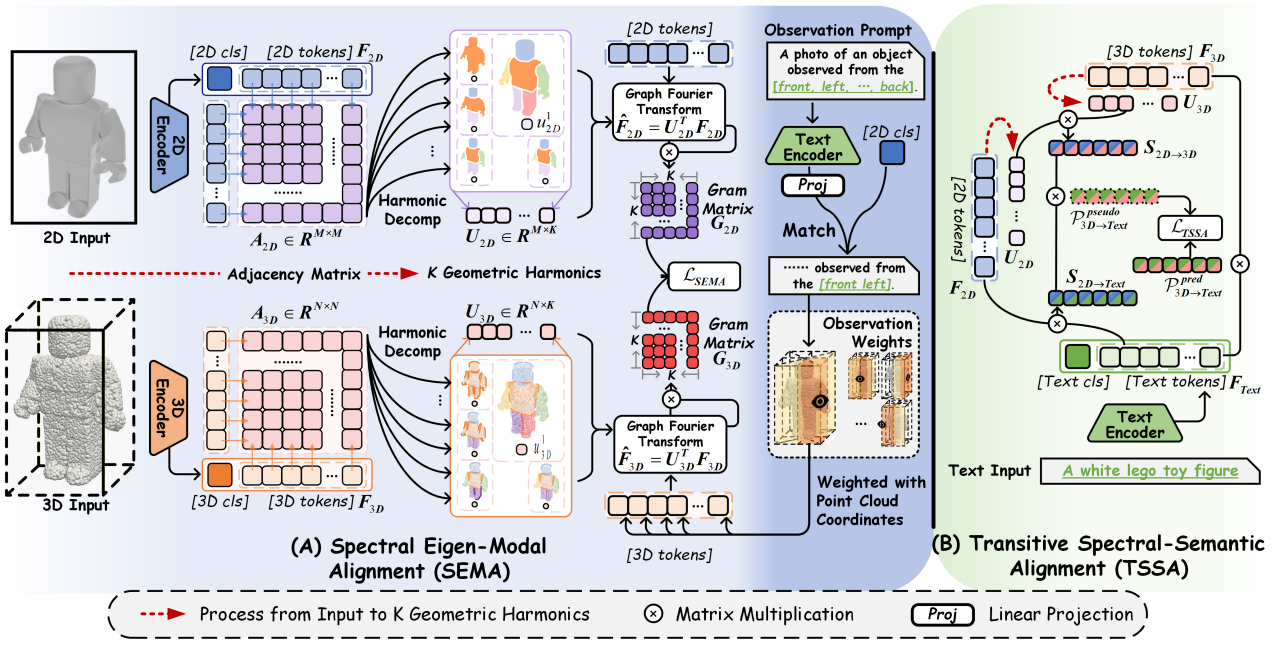

Figure 2: Overview of the SpecBridge framework. (A) SEMA extracts intrinsic geometric harmonics U_{2D}, U_{3D} from modality-specific adjacency matrices A_{2D}, A_{3D}. By projecting tokens onto these harmonics, where F_{3D} is modulated by CLIP-guided observation weights to address structural asymmetry, SEMA synchronizes cross-modal signatures through normalized spectral Gram matrices. (B) TSSA exploits a 2D structural intermediary $S_{2D\to 3D}$ to synthesize semantic prototypes $P^{\text{pseudo}}_{3D\to Text}$ from 2D-Text affinities $S_{2D\to Text}$. These provide dense supervision for the predictive distribution $P^{\text{pred}}_{3D\to Text}$, facilitating transitive linguistic knowledge propagation.

of fine-grained semantics between 3D and Text modalities. To address these issues, TSSA reformulates the 2D modality as a structural intermediary and establishes a 3D→2D→Text propagation chain, effectively avoiding the limitations of direct heterogeneous mapping.

3 Method

3.1 Framework Overview

This section designs a SpecBridge framework to unify geometric structures and linguistic semantics within a spectral representation space, as presented in Figure 2. The overall workflow of the proposed method consists of two main modules: spectral eigen-modal alignment (SEMA) and transitive spectral-semantic alignment (TSSA). To address the challenge of structural asymmetry, SEMA (Figure 2A) performs Laplacian eigendecomposition on modality-specific adjacency matrices A_{2D} and A_{3D} to extract geometric harmonics $U \in \{U_{2D}, U_{3D}\}$. Feature tokens are subsequently projected into the spectral domain by $\widehat{F} = U^\top F$, where F_{3D} are modulated by viewpoint-aware observation weights derived from CLIP-based matching. By aligning the resulting spectral Gram matrices G_{2D} and G_{3D} via the SEMA loss, the proposed method captures intrinsic structures and transcends superficial appearance matching. Furthermore, TSSA (Figure 2B) is developed to mitigate the semantic chasm by establishing a 3D→2D→Text propagation chain. A correspondence matrix $S_{2D\to 3D}$ is constructed to correlate heterogeneous modalities, enabling the generation of transitive semantic prototype as pseudo-label $P^{\text{pseudo}}_{3D\to Text}$ for token-level alignment with prediction $P^{\text{pred}}_{3D\to Text}$. Optimization by the TSSA loss ensures that fine-grained semantics are effectively distilled into 3D geometries through the 2D visual intermediary.

3.2 Spectral Eigen-Modal Alignment

Euclidean domain limitations and viewpoint sensitivity are transcended by SEMA through the construction of cross-modal correlations in the intrinsic eigen-spectral domain. Rather than matching superficial features, SEMA separates extrinsic noise by capturing inherent geometric signatures through three processes: (i) intra-modal structural correlation modeling, (ii) intrinsic geometric harmonic decomposition, and (iii) spectral coherence synchronization. The detailed process of SEMA is illustrated in Figure 2A.

Intra-modal Structural Correlation Modeling. To capture intrinsic relational dependencies, modality-specific topology graphs are constructed to represent the structural distribution of feature tokens. Given 2D visual tokens $F_{2D} = \{f_i^{2D}\}_{i=1}^{M} \in \mathbb{R}^{M\times d}$ and 3D geometric tokens $F_{3D} = \{f_j^{3D}\}_{j=1}^{N} \in \mathbb{R}^{N\times d}$, adjacency matrices $A_{2D} \in \mathbb{R}^{M\times M}$ and $A_{3D} \in \mathbb{R}^{N\times N}$ are computed. For modality $m \in \{2D, 3D\}$ with cardinality n_m, the entries of A_m are defined as:

$$(A_m)_{ij} = \begin{cases} \exp\left(-\dfrac{\|f_i^m - f_j^m\|_2^2}{2\tau^2}\right), & \text{if } i \neq j, \\ 0, & \text{if } i = j, \end{cases} \quad \text{s.t.} \quad i, j \in \{1, \ldots, n_m\}. \tag{1}$$

where $i, j \in \{1, \ldots, n_m\}$ and σ indicates the graph bandwidth determined by a self-tuning scheme [Zelnik-Manor and Perona, 2004] based on k-NN distances. By modeling the adjacency in the feature space, the semantic and geometric relationships between discrete tokens are effectively encoded into a symmetric correlation graph. This formulation ensures that subsequent spectral analysis reflects the underlying structural consistency of the feature space.

Intrinsic Geometric Harmonic Decomposition. To extract the fundamental structural components from the constructed topologies, a spectral analysis is performed on the adjacency matrices. For adjacency matrix A_m, a diagonal degree matrix $D_m \in \mathbb{R}^{n_m \times n_m}$ is defined to quantify the connectivity density of each token:

$$(D_m)_{ij} = \begin{cases} \sum_{j=1}^{n_m} A_{ij}, & i = j, \\ 0, & i \neq j, \end{cases} \quad i \in \{1, \ldots, n_m\}. \quad (2)$$

To ensure spectral stability and scale-invariance when aligning M 2D modality patches with N 3D modality tokens, the symmetric normalized Laplacian $L_m \in \mathbb{R}^{n_m \times n_m}$ is instantiated as core structural descriptors:

$$L_m = I - D_m^{-1/2} A_m D_m^{-1/2}. \quad (3)$$

where I indicates an identity matrix. The intrinsic geometric properties of each modality-specific manifold are characterized through the eigendecomposition of Laplacian matrix:

$$L_m U_m = U_m \Lambda_m, \quad \text{s.t.} \quad U_m^\top U_m = I_K. \quad (4)$$

where $U_m = [u_1^m, u_2^m, \ldots, u_K^m] \in \mathbb{R}^{n_m \times K}$ denotes the first K eigenvectors, termed geometric harmonics, and $\Lambda_m = \mathrm{diag}(\lambda_1^m, \lambda_2^m, \ldots, \lambda_K^m)$ is a diagonal matrix of corresponding eigenvalues. Truncating the spectrum to K components allows SEMA to filter high-frequency noise and provide robust spectral bases for subsequent cross-modal synchronization.

Spectral Coherence Synchronization. This process establishes structural alignment by projecting spatial tokens into intrinsic eigen-spectral domain. For the 2D modality, spectral coefficients are obtained by Graph Fourier Transform $\widehat{F}_{2D} = U_{2D}^\top F_{2D}$. Due to the modality asymmetry, not all 3D point clouds can be effectively observed from the corresponding 2D image viewpoint. A series of observation prompts "A photo of an object observed from the [Orientation]" are introduced for 2D image observation inference based on CLIP priors. Notably, since CLIP focuses more on category descriptions rather than viewpoint information, a lightweight linear projection layer is pre-learned to filter out the viewpoint-related information in text representations. More details are provided in the supplementary materials. Furthermore, based on observation inference, weight adjustment is performed on the tokens corresponding to 3D point cloud clusters under the current 2D observation view with decay from the visual center to the unobservable region. The 3D spectral coefficients $\widehat{F}_{3D}$ integrate this spatial weighting into the spectral projection:

$$\begin{aligned} \widehat{F}_{3D} &= U_{3D}^\top (\omega \odot F_{3D}), \\ \text{s.t.} \quad w_i &= \max\left(\eta, 1 - \exp(\delta_i - \bar{\delta})\right). \end{aligned} \quad (5)$$

where $\omega = [w_1, \ldots, w_N]^\top$ represents the observation weight vector, δ_i denotes the distance of the i-th cluster from the visual center, $\bar{\delta}$ indicates the average distance, and η defines the minimum weight threshold to preserve essential geometric information even in unobservable regions. To characterize patterns independently of cardinality, a normalized harmonic descriptor $G_m \in \mathbb{R}^{K \times K}$ is formulated by computing the Frobenius-normalized spectral Gram matrix:

$$G_m = \frac{\widehat{F}_m \widehat{F}_m^\top}{\|\widehat{F}_m \widehat{F}_m^\top\|_F}. \quad (6)$$

where G_m encodes the correlations of all distinct frequency modes [Yi *et al.*, 2017] and reveals the intrinsic structural vibration profile of the modality. The SEMA loss is subsequently defined as the squared Frobenius distance between the normalized spectral Gram matrices of the 3D and 2D modalities:

$$\mathcal{L}_{\mathrm{SEMA}} = \|G_{3D} - G_{2D}\|_F^2. \quad (7)$$

Minimizing $\mathcal{L}_{\mathrm{SEMA}}$ enforces structural coherence between the 3D latent representations of the intrinsic 2D visual manifold.

3.3 Transitive Spectral-Semantic Alignment

Existing methods remain plagued by the challenge of the semantic gap between 3D geometry and abstract language description. However, the fine-grained semantic connections between 3D and 2D, as well as between 2D and text, which are respectively realized by SEMA and CLIP, provide an opportunity to overcome this challenge. TSSA thus leverages the 2D modality as an intermediary pipeline to construct a 2D-3D-Text semantic propagation chain, minimizing this chasm. This transitive alignment is executed in two phases: (i) spectral transitive semantic prototype synthesis, and (ii) transitive semantic consistency optimization.

Spectral Transitive Semantic Prototype Synthesis. To implement an efficient semantic propagation mechanism, it is critical to define high-quality 3D-Text semantic prototypes as learning objectives. The fine-grained 2D-Text correlation can be directly derived from the CLIP-prior-based cross-modal token correlation $S_{2D \to Text} = F_{2D} F_{Text}^\top$, where $F_{Text} \in \mathbb{R}^{C \times d}$ denotes the text tokens. However, since there is no pre-aligned prior between the 2D and 3D encoders, to circumvent potential viewpoint noise in homogeneous modalities, the fine-grained 2D-3D correlation $S_{2D \to 3D}$ is characterized by cross-modal geometric harmonics:

$$S_{2D \to 3D} = U_{3D} U_{2D}^\top. \quad (8)$$

The synthesized semantic prototypes, denoted as $P_{3D \to Text}^{\mathrm{pseudo}} \in \mathbb{R}^{N \times C}$, are computed by projecting $S_{2D \to Text}$ and $S_{2D \to 3D}$ token-wise affinities through the spectral topology followed by:

$$P_{3D \to Text}^{\mathrm{pseudo}} = \mathrm{Softmax}\left(S_{2D \to 3D}^\top \cdot S_{2D \to Text}\right). \quad (9)$$

Through the intermediary role of the 2D modality, $P_{3D \to Text}^{\mathrm{pseudo}}$ achieves fine-grained soft matching between het-

erogeneous modalities. This process ensures that the resulting prototypes inherit the rich, zero-shot priors of the foundation model while remaining strictly constrained by the intrinsic manifold structure of the 3D data, resulting in a robust, structurally-consistent supervisory signal.

Transitive Semantic Consistency Optimization. This process distills knowledge within the synthesized prototypes into 3D encoder through consistency-driven optimization. The predictive 3D-Text matching $P^{\text{pred}}_{3D\to Text} \in \mathbb{R}^{N\times C}$ is derived by computing the normalized similarity between the 3D geometric features F_{3D} and the text embeddings F_{Text} of the linguistic tokens:

$$P^{\text{pred}}_{3D\to Text} = \text{Softmax}\left(F_{3D} \cdot F^{\top}_{Text}\right). \tag{10}$$

To inherit the dense semantic priors from the foundation model, the TSSA loss $\mathcal{L}_{\text{TSSA}}$ is defined as the Kullback-Leibler (KL) divergence between $P^{\text{pred}}_{3D\to Text}$ and $P^{\text{pseudo}}_{3D\to Text}$:

$$\mathcal{L}_{\text{TSSA}} = \frac{1}{N}\sum_{i=1}^{N} \text{KL}\left(P^{\text{pseudo}}_{3D\to Text,i} \,\|\, P^{\text{pred}}_{3D\to Text,i}\right). \tag{11}$$

$\mathcal{L}_{\text{TSSA}}$ projects 3D geometric features into linguistic space of CLIP. By aligning in fine-grained soft distribution, the correlation between heterogeneous modal tokens is revealed, enabling the 3D encoder to achieve robust open-vocabulary generalization through cross-modal spectral alignment.

4 Experiments

4.1 Experiment Setting

Datasets. SpecBridge is pre-trained on a consolidated ensemble of four large-scale 3D datasets including Objaverse [Deitke *et al.*, 2023], ShapeNet [Chang *et al.*, 2015], 3D-FUTURE [Fu *et al.*, 2021], and ABO [Collins *et al.*, 2022], following the OpenShape protocol to ensure the learning of diverse geometric representations of 875,665 shapes. Performance evaluation utilizes two downstream benchmarks characterized by distinct data properties. ModelNet40 [Wu *et al.*, 2015] comprises 12,311 synthetic CAD models, and Objaverse-LVIS [Deitke *et al.*, 2023] is a long-tailed distributed subset of Objaverse including 1,156 semantic categories.

Implementation Details. SpecBridge is implemented using the PyTorch environment and trained on four NVIDIA RTX A6000 GPUs for 300 epochs. The framework utilizes Point-BERT [Yu *et al.*, 2022] as the 3D backbone encoder to extract geometric features. To ensure stable multi-modal semantic priors throughout the optimization process, a frozen OpenCLIP ViT-bigG-14 [Cherti *et al.*, 2023] model is employed as the vision-language foundation. Optimization is conducted via the AdamW [Loshchilov and Hutter, 2019] optimizer with a global batch size of 1280, incorporating a linear warmup phase during the initial 20 epochs to reach a peak learning rate of 5×10^{-3}, followed by a cosine decay schedule to 3×10^{-3}.

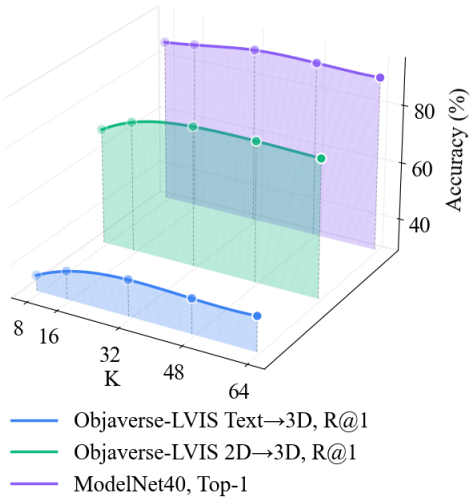

Figure 3: Ablation on the Harmonic Order K. Top-1 zero-shot classification accuracy on ModelNet40 and R@1 cross-modal retrieval on Objaverse-LVIS across different K.

#	SEMA	TSSA	ModelNet40		Objaverse-LVIS	
			Top-1	Top-5	Text→3D	2D→3D
1	✗	✗	85.1	97.1	30.2	66.8
2	◯	✗	86.4	98.4	32.6	73.7
3	✓	✗	88.6	99.0	33.9	75.9
4	✗	◯	86.1	98.3	38.4	69.3
5	✗	✓	87.4	98.9	41.1	71.1
6	✓	✓	**90.2**	**99.2**	**42.7**	**78.7**

Table 1: Effect of Different Modules. Symbols ✗ and ✓ denote module exclusion and full integration. SEMA ◯ represents direct computation of $\widehat{F}_{3D}$ omitting CLIP-guided observation inference. TSSA ◯ computes $S_{2D\to 3D}$ via direct token-level feature similarity instead of the spectral transitive bridge $U_{3D}U^{\top}_{2D}$.

4.2 Ablation Study

Ablation on Harmonic Order K. The harmonic order K determines the granularity of geometric features within the intrinsic feature spectral domain, with the experiment result shown in Figure 3. Due to the truncated spectral bases fail to encapsulate the fundamental low-frequency harmonics of complex 3D manifolds, lower values of K perform suboptimal accuracy. Conversely, increasing K beyond 32 leads to stagnant or even degraded performance, indicating that high-frequency eigenfunctions introduce extraneous noise and make it difficult to maintain stable internal topology. Therefore, $K = 32$ is selected as the optimal trade-off, which ensures sufficient structural details while preserving robustness against local perturbations and viewpoint variations.

Effect of Different Modules. The synergistic effect between the SEMA and TSSA modules in SpecBridge is further verified, with the experimental results presented in Table 1. For 3D-2D visual learning, the complete SEMA module in line 3 significantly outperforms the baseline in line 1 and the

Methods	Reference	Training Data	Objaverse-LVIS			ModelNet40		
			Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
OpenShape [Liu *et al.*, 2023]	*NeurIPS '23*	ShapeNet	10.8	20.2	25.0	70.3	86.9	91.3
ULIP-2 [Xue *et al.*, 2024]	*CVPR '24*		16.4	–	34.3	75.2	–	95.0
MixCon3D [Gao *et al.*, 2024]	*CVPR '24*		22.3	37.5	44.3	72.6	87.1	91.3
GoNet [Zhang *et al.*, 2025a]	*PR '25*		–	–	–	59.8	–	80.1
SpecBridge	*This Paper*		**25.6**	**39.9**	**48.9**	**77.2**	**90.7**	**95.6**
OpenShape [Liu *et al.*, 2023]	*NeurIPS '23*	Ensemble (no LVIS)	39.1	60.8	68.9	85.3	96.2	97.4
ULIP-2* [Xue *et al.*, 2024]	*CVPR '24*		46.3	–	75.0	84.0	–	97.2
Uni3D [Zhou *et al.*, 2024]	*ICLR '24*		47.2	69.8	76.1	86.8	97.4	98.4
MixCon3D [Gao *et al.*, 2024]	*CVPR '24*		47.5	69.0	76.2	87.3	96.8	98.1
CLIP-GS [Jiao *et al.*, 2025]	*ICCV '25*		48.5	70.3	77.5	86.7	97.6	98.6
PointCache [Sun *et al.*, 2025]	*CVPR '25*		47.3	69.0	76.3	87.1	97.1	98.4
SpecBridge	*This Paper*		**50.1**	**72.1**	**80.7**	**90.2**	**98.1**	**99.2**
OpenShape [Liu *et al.*, 2023]	*NeurIPS '23*	Ensemble	46.8	69.1	77.0	84.4	96.5	98.0
ULIP-2* [Xue *et al.*, 2024]	*CVPR '24*		50.6	–	79.1	84.7	–	97.1
Uni3D [Zhou *et al.*, 2024]	*ICLR '24*		53.5	74.9	82.0	87.3	96.4	99.2
MixCon3D [Gao *et al.*, 2024]	*CVPR '24*		52.5	74.5	81.2	86.8	96.9	98.3
CLIP-GS [Jiao *et al.*, 2025]	*ICCV '25*		53.5	76.1	82.0	87.0	97.9	98.8
PointCache [Sun *et al.*, 2025]	*CVPR '25*		55.2	76.0	82.5	89.2	98.2	99.4
SpecBridge	*This Paper*		**56.6**	**77.4**	**85.9**	**90.6**	**98.4**	**99.5**

Table 2: Zero-shot classification results on Objaverse-LVIS and ModelNet40. Performance is evaluated using average Top-1, Top-3, and Top-5 accuracy. Ensemble indicates a consolidated training set consisting of ShapeNet, Objaverse, ABO, and 3D-FUTURE. Methods marked with an asterisk (*) indicate training configurations using only Objaverse and ShapeNet for pre-training.

Method	2D→3D		3D→2D	
	R@1	R@5	R@1	R@5
OpenShape [Liu *et al.*, 2023]	61.6	86.4	53.8	80.7
ULIP-2 [Xue *et al.*, 2024]	5.6	15.3	25.0	50.0
Uni3D [Zhou *et al.*, 2024]	65.1	88.4	49.3	75.7
MixCon3D [Gao *et al.*, 2024]	64.6	86.3	44.3	68.6
CLIP-GS [Jiao *et al.*, 2025]	75.6	93.9	56.9	82.3
SpecBridge	**76.5**	**95.8**	**74.5**	**92.9**
Method	**Text→3D**		**3D→Text**	
	R@1	R@5	R@1	R@5
OpenShape [Liu *et al.*, 2023]	24.4	52.7	22.6	49.6
ULIP-2 [Xue *et al.*, 2024]	4.5	14.7	5.3	16.8
Uni3D [Zhou *et al.*, 2024]	27.8	57.0	23.1	49.5
MixCon3D [Gao *et al.*, 2024]	33.6	66.2	31.2	63.8
CLIP-GS [Jiao *et al.*, 2025]	36.8	68.1	30.0	59.1
SpecBridge	**49.7**	**77.2**	**44.8**	**73.8**

Table 3: Cross-modal retrieval on Objaverse-LVIS.

incomplete learning setup without CLIP-guided observation inference in line 2, especially on the 2D→3D retrieval task. This indicates that SEMA effectively explores CLIP priors to disentangle extrinsic observation noise and achieve more stable cross-modal structural correlations. In terms of 3D-Text learning, the spectral semantic propagation chain setup in line 5 maintains a substantial performance margin over the direct token-wise semantic propagation strategy in line 4 on the Text→3D retrieval task. This demonstrates that geometric harmonics preserve more effective structural consistency during the distillation of CLIP semantic priors. In line 6, the complete SpecBridge integrating both SEMA and TSSA achieves excellent performance across all benchmarks, which illustrates that SEMA and TSSA can synergize to enable superior 3D-2D-Text representation learning.

4.3 Main Results

Zero-shot Classification. Table 2 presents the performance comparison results with existing methods. Pioneering frameworks like OpenShape rely on rigid global alignment that hinders structural understanding for 3D shapes. While methods such as ULIP-2 and Uni3D leverage large-scale data expansion, they are still constrained by the performance bottleneck of Euclidean appearance matching. SpecBridge overcomes these limitations by performing alignment learning on low-frequency geometric harmonics to capture the intrinsic geometric essence of 3D shapes. Under the Ensemble (no LVIS) training setup, it achieves Top-1 accuracy on ModelNet that is 4.9%, 6.2%, and 3.4% higher than that of OpenShape, ULIP-2, and Uni3D, respectively. Furthermore, methods like MixCon3D and CLIP-GS focus on information quantity stacking. Although they incorporate multi-view and texture information to optimize 3D representation, they neglect the detailed mapping between such information and 3D geometry. SpecBridge introduces CLIP-prior-based observation inference to guide the model in learning viewpoint-invariant

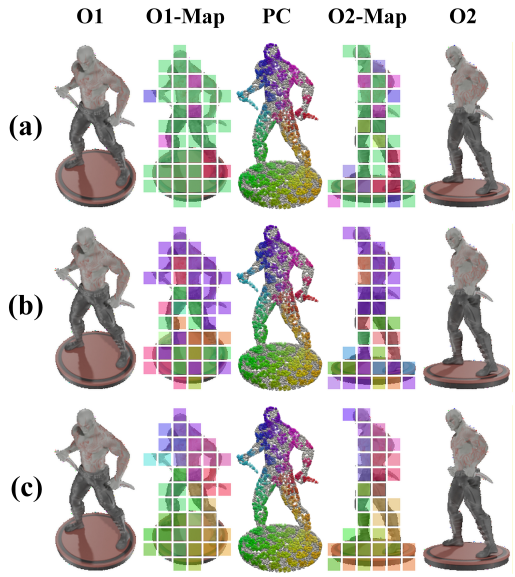

Figure 4: 2D-3D Structural Correspondence Visualization. Rows (a)-(c) represent OpenShape, MixCon3D, and SpecBridge. PC: color-coded 3D point clusters. O1/O2: distinct observations. O1/O2-Map colors signify image patches closest to 3D point clusters of the same color in the shared feature space.

3D representations. Ultimately, SpecBridge attains state-of-the-art performance with 90.6% and 56.6% Top-1 accuracy on ModelNet40 and Objaverse-LVIS, respectively, surpassing the sub-optimal PointCache method that relies on test-time optimization, validating the superiority of intrinsic spectral alignment.

Cross-modal Retrieval. Table 3 presents cross-modal retrieval results on Objaverse-LVIS. Because cross-modal retrieval tasks require model to learn high-quality cross-modal mappings, they are far more challenging than classification tasks. Methods such as OpenShape and Uni3D, which rely on data expansion, are clearly unable to capture the fine-grained alignment between complex 3D geometries and standard vision-language priors. While MixCon3D and CLIP-GS mitigate the modal discrepancy in visual learning by incorporating multi-view and texture information, they are still constrained by the semantic chasm in 3D-Text cross-retrieval tasks. SpecBridge leverages SEMA and TSSA to effectively overcome the modality discrepancies in 3D-2D and 3D-Text learning, respectively, thus achieving state-of-the-art performance. Notably, SpecBridge achieves significant performance improvements of 12.9% and 14.8% on the Text→3D and 3D→Text tasks, respectively.

2D-3D Structural Correspondence Visualization. The visualization in Figure 4 shows patch-to-cluster matching, where point clouds are split into 128 colored clusters and each ViT patch is assigned to its nearest 3D token. OpenShape produces coarse maps because its global alignment lacks part-level supervision. MixCon3D uses many-to-one 2D–3D learning with multi-view images, and roughly separates the statue's upper body (purple) and lower body (green). In contrast, SpecBridge integrates CLIP-guided observation

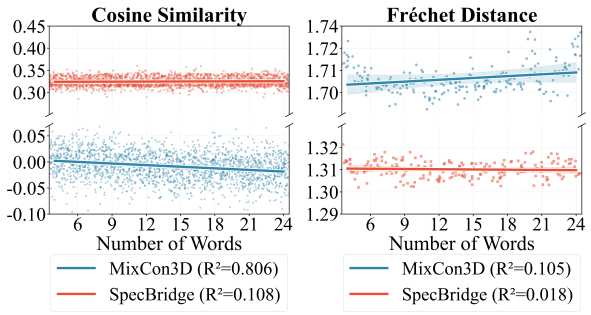

Figure 5: Linguistic Robustness of 3D-Text Alignment. Left: cross-modal cosine similarity for 10,000 random 3D-text pairs per word count of description. Right: Frechet Distance (FD) values computed over 10 random sets of 1,000 samples per word count of description. Charts illustrate distributions for MixCon3D (blue) and SpecBridge (red) with linear regression lines and R^2 values.

inference into SEMA, learning viewpoint-invariant 3D representations and yielding finer part segmentation, such as the left-hand dagger (red) and pedestal (yellow).

Linguistic Robustness of 3D-Text Alignment. Figure 5 evaluates semantic alignment robustness under increasing textual complexity. While MixCon3D exhibits noticeable performance decay, evidenced by decreasing cosine similarity and rising FD as word counts increase, SpecBridge maintains a remarkably stable and superior alignment profile. Specifically, SpecBridge yields higher similarity and lower FD across the entire linguistic spectrum. The near-zero slopes in the linear fits for SpecBridge substantiate its resilience to information density. Such robustness is facilitated by the TSSA module, which utilizes the 2D modality to propagate dense CLIP priors across the foundational bridge. Consequently, SpecBridge effectively surmounts the semantic chasm, ensuring consistent 3D text grounding for complex descriptive scenarios.

5 Conclusion

This paper presents SpecBridge, a framework synergizing spectral graph theory with transitive semantic learning for 3D-2D-Text pre-training, which consists of two core modules: SEMA and TSSA. SEMA explores CLIP priors to guide 3D-2D modalities in establishing viewpoint-invariant associations within the eigen-spectral domain. TSSA constructs a 3D→2D→Text semantic propagation chain, which serves to mitigate the semantic chasm between 3D geometries and abstract textual descriptions. Benchmark evaluations substantiate the superiority of this approach in surmounting modality discrepancies. Furthermore, given the demand of open vocabulary 3D understanding for complex scenes, as well as the superior geometric representation capability of the recently emerging 3D Gaussian Splatting over point clouds, our future research will focus on two key directions: extending object-level understanding tasks to complex indoor and outdoor scenarios, and developing a unified 3D representation learning paradigm compatible with both point clouds and 3D Gaussian Splatting.

References

[Chai *et al.*, 2024] Xiongli Chai, Feng Shao, Baoyang Mu, Hangwei Chen, Qiuping Jiang, and Yo-Sung Ho. Plainpcqa: No-reference point cloud quality assessment by analysis of plain visual and geometrical components. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6207–6223, 2024.

[Chang *et al.*, 2015] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[Cheng *et al.*, 2024] Wencan Cheng, Hao Tang, Luc Van Gool, and Jong Hwan Ko. Handdiff: 3d hand pose estimation with diffusion on image-point cloud. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2024.

[Cherti *et al.*, 2023] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2023.

[Collins *et al.*, 2022] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2022.

[Deitke *et al.*, 2023] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2023.

[Fang *et al.*, 2025] Shuangkang Fang, I. Shen, Yufeng Wang, Yi-Hsuan Tsai, Yi Yang, Shuchang Zhou, Wenrui Ding, Takeo Igarashi, Ming-Hsuan Yang, et al. Meshllm: Empowering large language models to progressively understand and generate 3d mesh. In *International Conference on Computer Vision, ICCV*, 2025.

[Fu *et al.*, 2021] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129(12):3313–3337, 2021.

[Gao *et al.*, 2024] Yipeng Gao, Zeyu Wang, Wei-Shi Zheng, Cihang Xie, and Yuyin Zhou. Sculpting holistic 3d representation in contrastive language-image-3d pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2024.

[Hess *et al.*, 2024] Georg Hess, Adam Tonderski, Christoffer Petersson, Kalle Åström, and Lennart Svensson. Lidarclip or: How i learned to talk to point clouds. In *Winter Conference on Applications of Computer Vision, WACV*, 2024.

[Huang *et al.*, 2024] Rui Huang, Xuran Pan, Henry Zheng, Haojun Jiang, Zhifeng Xie, Cheng Wu, Shiji Song, and Gao Huang. Joint representation learning for text and 3d point cloud. *Pattern Recognition*, 147:110086, 2024.

[Huang *et al.*, 2025] Jiangyong Huang, Baoxiong Jia, Yan Wang, Ziyu Zhu, Xiongkun Linghu, Qing Li, Song-Chun Zhu, and Siyuan Huang. Unveiling the mist over 3d vision-language understanding: Object-centric evaluation with chain-of-analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2025.

[Jain *et al.*, 2025] Ayush Jain, Alexander Swerdlow, Yuzhou Wang, Sergio Arnaud, Ada Martin, Alexander Sax, Franziska Meier, and Katerina Fragkiadaki. Unifying 2d and 3d vision-language understanding. In *International Conference on Machine Learning, ICML*, 2025.

[Jiang *et al.*, 2024] Ying Jiang, Chang Yu, Tianyi Xie, Xuan Li, Yutao Feng, Huamin Wang, Minchen Li, Henry Lau, Feng Gao, Yin Yang, et al. Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. In *Special Interest Group on GRAPHics and Interactive Techniques Conference, SIGGRAPH*, 2024.

[Jiao *et al.*, 2025] Siyu Jiao, Haoye Dong, Yuyang Yin, Zequn Jie, Yinlong Qian, Yao Zhao, Humphrey Shi, and Yunchao Wei. Clip-gs: Unifying vision-language representation with 3d gaussian splatting. In *International Conference on Computer Vision, ICCV*, 2025.

[Lee *et al.*, 2025] Junha Lee, Chunghyun Park, Jaesung Choe, Yu-Chiang Frank Wang, Jan Kautz, Minsu Cho, and Chris Choy. Mosaic3d: Foundation dataset and model for open-vocabulary 3d segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2025.

[Liu *et al.*, 2023] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. In *Advances in Neural Information Processing Systems, NeurIPS*, 2023.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations, ICLR*, 2019.

[Qi *et al.*, 2024a] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision, ECCV*, 2024.

[Qi *et al.*, 2024b] Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. Gpt4point: A unified framework for point-language understanding and generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack

Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning, ICML*, 2021.

[Sun *et al.*, 2025] Hongyu Sun, Qiuhong Ke, Ming Cheng, Yongcai Wang, Deying Li, Chenhui Gou, and Jianfei Cai. Point-cache: Test-time dynamic and hierarchical cache for robust and generalizable point cloud analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2025.

[Tang *et al.*, 2025] Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Jinfeng Xu, Yixue Hao, Long Hu, and Min Chen. More text, less point: Towards 3d data-efficient point-language understanding. In *AAAI Conference on Artificial Intelligence, AAAI*, 2025.

[Wang *et al.*, 2025] Guangming Wang, Yu Zheng, Yuxuan Wu, Yanfeng Guo, Zhe Liu, Yixiang Zhu, Wolfram Burgard, and Hesheng Wang. End-to-end 2d-3d registration between image and lidar point cloud for vehicle localization. *IEEE Transactions on Robotics*, 41:4643–4662, 2025.

[Wu *et al.*, 2015] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2015.

[Xu *et al.*, 2025] Runsen Xu, Shuai Yang, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm-v2: Empowering large language models to better understand point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15, 2025.

[Xue *et al.*, 2023] Le Xue, Mingfei Gao, Chen Xing, Roberto Martin-Martin, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2023.

[Xue *et al.*, 2024] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martin-Martin, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2024.

[Yang *et al.*, 2024] Timing Yang, Yuanliang Ju, and Li Yi. Imov3d: Learning open vocabulary point clouds 3d object detection from only 2d images. In *Advances in Neural Information Processing Systems, NeurIPS*, 2024.

[Yang *et al.*, 2025a] Xuemeng Yang, Licheng Wen, Tiantian Wei, Yukai Ma, Jianbiao Mei, Xin Li, Wenjie Lei, Daocheng Fu, Pinlong Cai, Min Dou, et al. Drivearena: A closed-loop generative simulation platform for autonomous driving. In *International Conference on Computer Vision, ICCV*, 2025.

[Yang *et al.*, 2025b] Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3d-mem: 3d scene memory for embodied exploration and reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2025.

[Yi *et al.*, 2017] Li Yi, Hao Su, Xingwen Guo, and Leonidas J. Guibas. Syncspeccnn: Synchronized spectral cnn for 3d shape segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017.

[Yu and Song, 2024] Hai-Tao Yu and Mofei Song. Mm-point: Multi-view information-enhanced multi-modal self-supervised 3d point cloud understanding. In *AAAI Conference on Artificial Intelligence, AAAI*, 2024.

[Yu *et al.*, 2022] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2022.

[Zelnik-Manor and Perona, 2004] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems, NeurIPS*, 2004.

[Zha *et al.*, 2025] Yixin Zha, Chuxin Wang, Wenfei Yang, and Tianzhu Zhang. Exploring semantic masked autoencoder for self-supervised point cloud understanding. In *International Joint Conference on Artificial Intelligence, IJCAI*, 2025.

[Zhai *et al.*, 2025] Hongjia Zhai, Hai Li, Zhenzhe Li, Xiaokun Pan, Yijia He, and Guofeng Zhang. Panogs: Gaussian-based panoptic segmentation for 3d open vocabulary scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2025.

[Zhang *et al.*, 2023] Zicheng Zhang, Wei Sun, Xiongkuo Min, Qiyuan Wang, Jun He, Quan Zhou, and Guangtao Zhai. Mm-pcqa: Multi-modal learning for no-reference point cloud quality assessment. In *International Joint Conference on Artificial Intelligence, IJCAI*, 2023.

[Zhang *et al.*, 2025a] Huang Zhang, Long Yu, Guoqi Wang, Shengwei Tian, Zaiyang Yu, Weijun Li, and Xin Ning. Cross-modal knowledge transfer for 3d point clouds via graph offset prediction. *Pattern Recognition*, 162:111351, 2025.

[Zhang *et al.*, 2025b] Yifan Zhang, Junhui Hou, Siyu Ren, Jinjian Wu, Yixuan Yuan, and Guangming Shi. Self-supervised learning of lidar 3d point clouds via 2d-3d neural calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):9201–9216, 2025.

[Zhou *et al.*, 2024] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations, ICLR*, 2024.

[Zhu *et al.*, 2025] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In *International Conference on Computer Vision, ICCV*, 2025.