3DAlign-DAER: Dynamic Attention Policy and Efficient Retrieval Strategy for Fine-grained 3D-Text Alignment at Scale

Yijia Fan[1*], **Jusheng Zhang**[1*], **Kaitong Cai**[1], **Jing Yang**[1], **Jian Wang**[2], **Keze Wang**[1,3†]

[1]Sun Yat-sen University
[2]Snap Inc.
[3]Guangdong Key Laboratory of Big Data Analysis and Processing

Abstract

Despite recent advancements in 3D-text cross-modal alignment, existing state-of-the-art methods still struggle to align fine-grained textual semantics with detailed geometric structures, and their alignment performance degrades significantly when scaling to large-scale 3D databases. To overcome this limitation, we introduce **3DAlign-DAER**, a unified framework designed to align text and 3D geometry via the proposed dynamic attention policy and the efficient retrieval strategy, capturing subtle correspondences for diverse cross-modal retrieval and classification tasks. *Specifically*, during the training, our proposed dynamic attention policy (DAP) employs the Hierarchical Attention Fusion (HAF) module to represent the alignment as learnable fine-grained token-to-point attentions. To optimize these attentions across different tasks and geometric hierarchies, our DAP further exploits the Monte Carlo tree search to dynamically calibrate HAF attention weights via a hybrid reward signal and further enhances the alignment between textual descriptions and local 3D geometry. During the inference, our 3DAlign-DAER introduces an Efficient Retrieval Strategy (ERS) to leverage efficient hierarchical searching in the large-scale embedding spaces, outperforming traditional methods (*e.g.*, KNN) in accuracy and efficiency. *Furthermore*, to facilitate text-3D alignment research and train our 3DAlign-DAER, we construct **Align3D-2M**, a large-scale dataset featuring 2M text-3D pairs, to provide sufficient fine-grained cross-modal annotations. Extensive and comprehensive experiments demonstrate the superior performance of our 3DAlign-DAER on diverse benchmarks.

Introduction

Text-3D alignment, *i.e.*, aligning natural language descriptions with 3D geometric representations, has widespread real-world applications, such as robotic manipulation, augmented reality, and large-scale retrieval. Although existing SOTA methods (Sarkar et al. 2025; Hegde, Valanarasu, and Patel 2023) demonstrate promising results in large-scale pretraining and global feature alignment (Zhou et al. 2024), they are still faced with two critical challenges: *i)*, existing retrieval methods often struggle at aligning fine-grained textual descriptions (Ren and Wang 2025; Cao et al. 2023) (*e.g.*, distinguishing *a ceramic mug with a handle* from *a simple drinking glass*) with corresponding local geometric structures (*e.g.*, the presence or absence of a handle). This is primarily because they operate on coarse-grained feature maps or attention mechanism (*e.g.*, global `[CLS]` token) and tend to overlook these critical geometric details that provide semantic discriminative information; *ii)*, these methods suffer from poor scalability, i.e., their performance drops sharply on larger 3D databases due to the increased difficulty of discriminating targets from challenging distractors. Thus, an effective retrieval method is required for robust sparse alignments to overcome the weaknesses of traditional KNN-based methods (Radford et al. 2021) at scale. Besides, a dedicated, fine-grained, and large-scale text-3D alignment benchmark is missing to set the foundation for rigorous training and standardized evaluation aimed at addressing the above two challenges. Although the existing dataset (*i.e.*, ObjaverseXL (Deitke, Liu, and et al. 2023)) contains scaled 3D assets, its annotated textual descriptions are mostly from noisy web sources and uncurated, which does not satisfy the need for fine-grained alignment (Zhang et al. 2025a,c).

To address these issues, we introduce our open-source dataset **Align3D-2M**. Constructed using our customized parallel rendering and annotation pipelines, it comprises 2 million curated text-3D pairs featuring fine-grained text-geometry correspondences. The significant scale and high quality of alignment within Align3D-2M enable the development and rigorous evaluation of models targeting complex real-world text-3D relationships for learning fine-grained alignment at scale. We further introduce our **3DAlign-DAER** framework to achieve fine-grained alignment and scalability, fully exploiting the potential of our Align3D-2M dataset. The central innovation of our 3DAlign-DAER lies in employing a Dynamic Attention Policy (DAP) to modulate the fine-grained token-to-point attentions learned by our Hierarchical Attention Fusion (HAF) module. Guided by a tailored hybrid reward from Monte Carlo Tree Search (MCTS) (Świechowski et al. 2022), our DAP iteratively calibrates attention weights, ensuring precise correspondence between textual descriptions and local geometric details. Extensive and comprehensive results demonstrate that the alignment quality of our 3DAlign-DAER is significantly enhanced through the dynamic attention refinement of DAP.

*These authors contributed equally.
†Corresponding author: kezewang@gmail.com

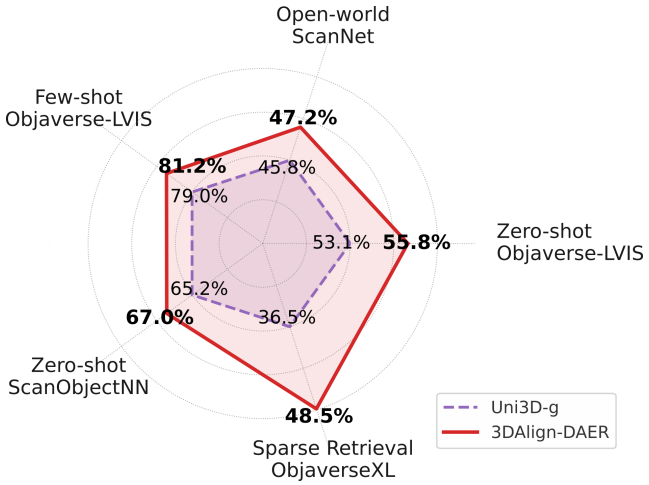

Figure 1: Our 3DAlign-DAER outperforms all task-specialized state-of-the-art methods on multiple 3D benchmarks and tasks (*i.e.*, few/zero-shot classification, large-scale retrieval, and open-world understanding).

Thus, 3DAlign-DAER advances in achieving robust retrieval performance at scale by making features more discriminative against hard distractors.

Specifically, during the training phase, our DAP employs an adaptive online search by using MCTS on attention maps from the HAF module. For each training sample, rather than relying on fixed attention weights, our DAP determines the optimal modulation strategy for the HAF weights. Our DAP is guided by our designed hybrid reward signal, which incorporates both dense feedback from the reduction in contrastive loss and sparse feedback based on retrieval performance on a validation set. By actively exploring how to enhance or suppress specific text token-to-3D patch associations based on this reward, our DAP forces the model to learn the precise and robust fine-grained correspondences necessary for challenging alignments.

To achieve efficient inference, especially when tackling large-scale sparse data retrieval , we design an **Efficient Retrieval Strategy (ERS)** for speed and accuracy. Our ERS leverages the highly discriminative and well-aligned representations produced by the training process of 3DAlign-DAER, which incorporates dynamic attention refinement in DAP. It constructs semantic and spatial hierarchies over the embedding space based on the learned representations, enabling an efficient tree-based search strategy to locate Top-K matches rapidly. This approach significantly boosts retrieval performance on large-scale datasets, overcoming the speed and accuracy limitations of traditional KNN, ANN, or Greedy Search techniques.

Comprehensive experiments show that our 3DAlign-DAER achieves SOTA on zero-shot classification, cross-modal retrieval (including large-scale), and few-shot learning, significantly outperforming strong baselines.

Our **main** contributions are summarized as follows: i) We propose **3DAlign-DAER**, a unified framework to dynamically refine fine-grained token-to-point attention and enhance sparse text-3D alignment. Our 3DAlign-DAER further incorporates an efficient retrieval strategy for scalable and large-scale retrieval. Our 3DAlign-DAER can handle multiple geometric hierarchies and tasks; ii) We construct **Align3D-2M**, a large-scale dataset comprising ∼2M fine-grained text–3D pairs. We develop pipelines for parallel point cloud rendering and annotation by curating multiple noisy datasets. Align3D-2M offers rich, clean, and consistent cross-modal annotations at scale, facilitating future research for large-scale 3D-text alignment; iii) Extensive experiments across diverse benchmarks demonstrate that our 3DAlign-DAER framework achieves new SOTA performance in 3D-text alignment tasks, including zero-shot / few-shot classification and large-scale cross-modal retrieval.

Related Works

3D-Text Cross-Modal Alignment. Learning joint representations for shapes and language often involves aligning global features using vision-language models like CLIP (Radford et al. 2021) and its variants (Wang, Mei, and Yuille 2023). Representative methods like ULIP (Xue et al. 2023), OpenShape (Liu et al. 2023b), and Uni3D (Zhou et al. 2024) leverage large datasets (Deitke, Liu, and et al. 2023) to train 3D encoders whose features align with text embeddings in a shared space, enabling zero-shot retrieval and classification. These methods achieve strong coarse alignment by scaling models and data, sometimes enhancing text descriptions. Other works (Luo et al. 2023; Luo, Johnson, and Lee 2024; Kabra et al. 2024) transfer 2D knowledge by projecting 3D data or incorporate auxiliary tasks like reconstruction. However, these methods rely on aligning global representations, limiting their ability to capture fine-grained correspondences between text phrases and specific shape parts.

Fine-Grained Alignment and Hierarchical Fusion. Addressing the limitation of global alignment requires fine-grained grounding. While the 2D vision-language tasks using cross-modal attention (Ye et al. 2019; Li et al. 2022; Zhang et al. 2025b), achieving this for general 3D shapes is challenging. Some efforts (Lu et al. 2024; Qi et al. 2019) explore multi-level features or apply attention in specific domains like 3D scene grounding. Yet, most existing 3D-text models (Le and et al. 2023) lack explicit token-to-point reasoning by interacting only from a global objective. There is a requirement for hierarchical fusion methods that can link local geometry to language. Our work advances in a dynamic mechanism to achieve finer alignment.

Search-Based Optimization for Alignment. Standard cross-modal learning (Li et al. 2023; Liu et al. 2023a) relies on static architectures and gradient-based optimization. Reinforcement learning (RL) (Schulman et al. 2017; Zhang et al. 2025d) or search strategies offer potential for dynamic refinement but remain largely unexplored for 3D-text alignment. While RL has been used elsewhere in vision-language (Shen et al. 2025), and search methods like Monte Carlo Tree Search (MCTS) (Świechowski et al. 2022; Silver et al. 2017) guide decisions in other domains (Zhang et al. 2024; OpenAI 2024b; Sun et al. 2023), their application to directly optimize cross-modal alignment attention is novel. Inspired

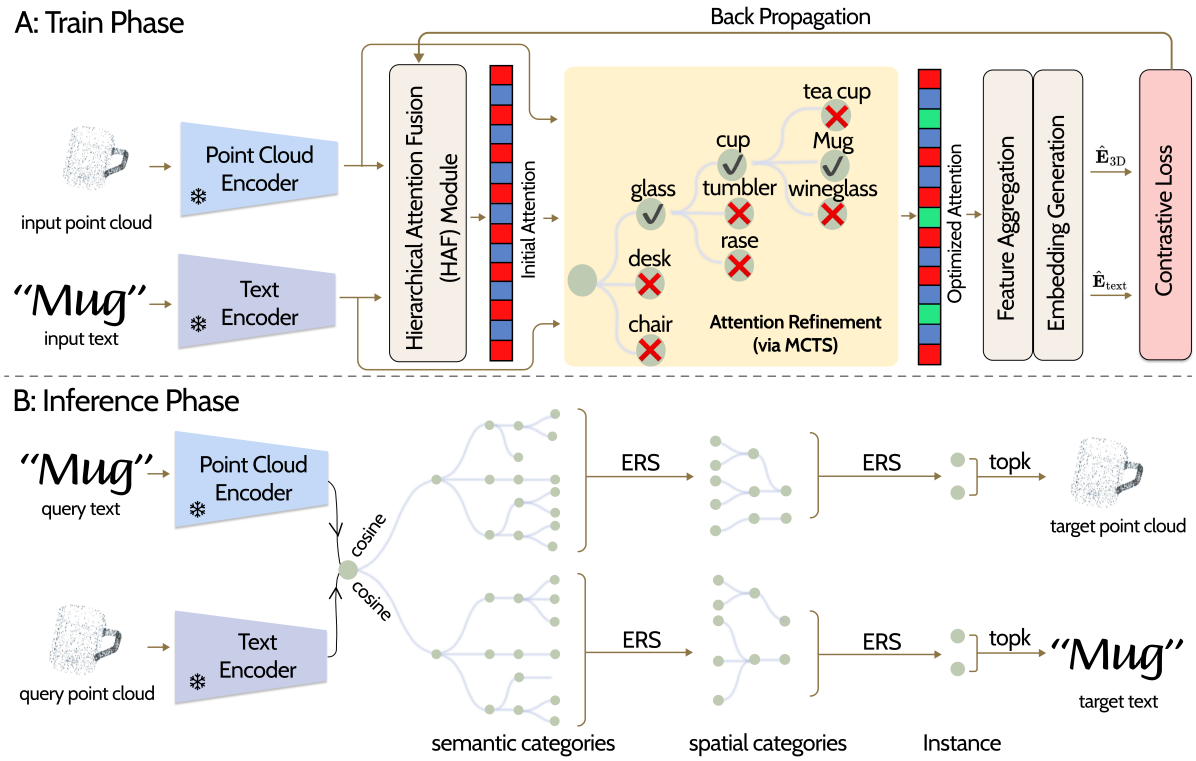

Figure 2: Overview of the 3DAlign-DAER framework. (A) Training Phase: Input modalities are initially processed by pre-trained encoders from Uni3D-g. Features are then refined through the Hierarchical Attention Fusion (HAF) module and an attention optimization module (based on MCTS) that implements the Dynamic Attention Policy (DAP). Our 3DAlign-DAER is trained using a contrastive loss on the final aggregated embeddings. (B) Inference Phase: For retrieval tasks, queries are encoded using the point cloud and text encoders belonging to the trained 3DAlign-DAER model. We propose an Efficient Retrieval Strategy (ERS) to navigate semantic and spatial categories to find the target.

by MCTS, our approach actively searches the space of attention configurations during inference to discover better fine-grained alignments, departing from conventional end-to-end training or predefined dynamic routing mechanisms.

3D Datasets for Pretraining. Progress has been tightly linked to dataset availability, moving from smaller annotated datasets like Text2Shape (Chen et al. 2018) and ShapeNet (Chang et al. 2015) to large web-scraped collections like Objaverse (Deitke et al. 2022). These large datasets enabled the scaling of models like OpenShape (Liu et al. 2023b) and Uni3D (Zhou et al. 2024), improving open-vocabulary capabilities. However, large datasets often contain weak textual annotations (e.g., tags, titles), limiting fine-grained understanding despite their scale.

Methodology

Align3D-2M Dataset

To enable 3D model retrieval based on open-text instructions, we constructed a large-scale multimodal 3D dataset named Align3D-2M. During the dataset construction process, we have integrated multiple publicly available 3D model repositories, such as Objaverse-XL, ShapeNet (Chang et al. 2015), Objaverse 1.0 (Deitke et al. 2022), ObjectNet3D (Xiang et al. 2016), Pix3D (Sun et al. 2018), and 3D_FUTURE (Fu et al. 2020), and developed a unique data cleaning and standardization pipeline. Public repositories like ObjaverseXL, while vast, often contain raw, uncleaned metadata and include non-descriptive strings (e.g., "a591c555d?hl=ru") unsuitable for detailed cross-modal training. Our primary goal is to create a large-scale dataset with consistent, semantically rich text annotations specifically tailored to effectively train advanced text-to-3D alignment models like our proposed 3DAlign-DAER.

To overcome the limitations of inconsistent or noisy source metadata and generate suitable training pairs, we first extract diverse 3D objects along with their available metadata (such as category labels and geometric attributes) from the integrated sources. For each object, we render a standard frontal view image. Subsequently, we utilize the multimodal capabilities of GPT-4o (OpenAI 2024a), providing it with both the rendered frontal image and the extracted metadata. Guided by prompt strategies incorporating this information and including randomization parameters to ensure diversity, GPT-4o is leveraged to batch-produce initial text descriptions designed for better cross-modal understanding. This process initially generated about 2 million descriptions.

To refine these descriptions, we employ a combination of

automated and human reviews. Initially, a language model (a BERT-based (Devlin et al. 2019; Liu et al. 2019) classifier) performs preliminary screening, filtering out descriptions with obvious issues like grammatical errors, excessive generalization, or those clearly not describing 3D objects (filtering approx. 2.7%). To verify the quality and suitability of the generated descriptions for training, we hired around 50 annotators who conducted a random sample review of about 10% of the remaining descriptions. They focus on evaluating the accuracy, specificity, and relevance of the description to the 3D object based on predefined criteria. This multi-stage process aims to ensure a baseline level of quality and consistency across the large generated dataset, resulting in about 2 million description-object pairs deemed suitable for training our 3DAlign-DAER. Specific review criteria and statistics are detailed in the Appendix "Details on Align3D-2M Text Annotation and Data Curation".

Model Architecture

Our 3DAlign-DAER comprises four key components: a text encoder, a 3D encoder, the Hierarchical Attention Fusion (HAF) module that incorporates MCTS-driven attention modulation according to our proposed Dynamic Attention Policy (DAP), and a contrastive learning objective.

Text Encoder We employ a Transformer-based architecture to process input natural language descriptions. Given an input text sequence, the encoder generates token-level features projected into a shared d-dimensional embedding space: $\mathbf{F}_{\text{text}} \in \mathbb{R}^{T\times d}$, where T is the number of text tokens and d is the feature dimension.

3D Encoder The 3D encoder processes point cloud data, extracting geometric and spatial information. We utilize a hierarchical architecture, potentially inspired by PointNet++, involving point set sampling, local feature aggregation, and global information fusion. This transforms the 3D geometry into features compatible with the text features: $\mathbf{F}_{\text{3D}} \in \mathbb{R}^{N}$, where N is the number of sampled points. Both $\mathbf{F}_{\text{text}}$ and $\mathbf{F}_{\text{3D}}$ reside in the same d-dimensional space.

Hierarchical Attention Fusion (HAF) Module The HAF module establishes fine-grained correspondences using cross-attention. Text and 3D features are linearly projected to Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$) matrices using learnable weights $\mathbf{W}_Q, \mathbf{W}_K^{\text{3D}}, \mathbf{W}_V^{\text{3D}}, \mathbf{W}_V^{\text{text}} \in \mathbb{R}^{d\times d}$. For text-to-3D attention:

$$\mathbf{Q}_{\text{text}} = \mathbf{F}_{\text{text}}\mathbf{W}_Q \qquad (\mathbf{Q}_{\text{text}} \in \mathbb{R}^{T\times d}) \tag{1}$$

$$\mathbf{K}_{\text{3D}} = \mathbf{F}_{\text{3D}}\mathbf{W}_K^{\text{3D}} \qquad (\mathbf{K}_{\text{3D}} \in \mathbb{R}^{N\times d}) \tag{2}$$

$$\mathbf{V}_{\text{3D}} = \mathbf{F}_{\text{3D}}\mathbf{W}_V^{\text{3D}} \qquad (\mathbf{V}_{\text{3D}} \in \mathbb{R}^{N\times d}) \tag{3}$$

For 3D-to-text attention (used in aggregation), text values are:

$$\mathbf{V}_{\text{text}} = \mathbf{F}_{\text{text}}\mathbf{W}_V^{\text{text}} \quad (\mathbf{V}_{\text{text}} \in \mathbb{R}^{T\times d}) \tag{4}$$

An initial cross-attention weight matrix is computed via scaled dot-product attention:

$$\mathbf{A}_{\text{initial}} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{text}}\mathbf{K}_{\text{3D}}^{\top}}{\sqrt{d}}\right) \in \mathbb{R}^{T\times N} \tag{5}$$

The softmax$(\cdot)$ is applied row-wise. This $\mathbf{A}_{\text{initial}}$ serves as the starting point for MCTS refinement.

MCTS-driven Dynamic Attention Policy

To refine the initial alignment $\mathbf{A}_{\text{initial}}$, our DAP introduces MCTS to dynamically modulate the attention weights, searching for an optimized distribution $\mathbf{A}_{\text{optimized}}$ that maximizes a reward signal.

State (s): Represents the current attention matrix $\mathbf{A} \in \mathbb{R}^{T\times N}$. The root state corresponds to $\mathbf{A}_{\text{initial}}$.

Action (a): A modification operation $a : \mathbb{R}^{T\times N} \to \mathbb{R}^{T\times N}$, yielding $\mathbf{A}' = \text{Normalize}(\text{op}(\mathbf{A}, \Delta, \mathbf{M}))$. Here, op is the operation (e.g., addition), Δ is the magnitude, $\mathbf{M}$ is a binary mask, and Normalize re-applies row-wise softmax. Actions include enhancing or suppressing attention in $\mathbf{M}$.

Reward (R): Evaluates the quality of reaching a state, incorporating a dense internal loss-oriented reward and a sparse external retrieval-oriented reward:

$$R_{\text{total}} = \alpha \cdot R_{\text{internal}} + (1-\alpha)\cdot R_{\text{external}} \tag{6}$$

where $\alpha \in [0,1]$ is a balance factor. R_{internal} denotes the decrements in the contrastive loss (see §) after each MCTS action. R_{external} denotes the performance score of the final downstream retrieval task based on metrics (*i.e.*, weighted Recall@K and mAP). These scores can be efficiently approximated during MCTS rollouts. MCTS iteratively builds a **search tree** via four steps:

Selection: Traverse the tree from the root by selecting child nodes maximizing the UCT score :

$$\text{UCT}(s,a) = \overline{Q}(s,a) + c\cdot\sqrt{\frac{2\ln N(s)}{N(s,a)+\epsilon}} \tag{7}$$

where $\overline{Q}(s,a)$ is the average reward after action a from state s, $N(s)$ and $N(s,a)$ are visit counts, c is the exploration constant, and ϵ prevents division by zero. Selection stops at an expandable leaf node.

Expansion: Add one or more child nodes corresponding to valid actions $a \in C(s)$ applicable to the leaf node's state s.

Simulation (Rollout): Estimate the reward from a new node s' by simulating subsequent actions using a fast policy π_{rollout} for a fixed depth d: $\hat{R} = \text{Simulate}(s', \pi_{\text{rollout}}, d)$. The reward is calculated via Eq. 6.

Backpropagation: Propagate the estimated reward $\hat{R}$ up the path from the expanded node to the root, updating visit counts $N(s), N(s,a)$ and average rewards $\overline{Q}(s,a)$ using incremental mean updates:

$$\begin{aligned} N(s) &\leftarrow N(s)+1 \\ N(s,a) &\leftarrow N(s,a)+1 \\ \overline{Q}(s,a) &\leftarrow \overline{Q}(s,a) + \frac{\hat{R}-\overline{Q}(s,a)}{N(s,a)} \end{aligned} \tag{8}$$

After a fixed budget, the action corresponding to the most promising path from the root yields $\mathbf{A}_{\text{optimized}}$.

Attention-Guided Feature Aggregation

The optimized attention $\mathbf{A}_{\text{optimized}}$ guides cross-modal feature aggregation:

$$\mathbf{Z}_{\text{text}} = \mathbf{A}_{\text{optimized}} \cdot \mathbf{V}_{\text{3D}} \qquad (\mathbf{Z}_{\text{text}} \in \mathbb{R}^{T \times d}) \tag{9}$$

$$\mathbf{Z}_{\text{3D}} = \mathbf{A}_{\text{optimized}}^{\top} \cdot \mathbf{V}_{\text{text}} \qquad (\mathbf{Z}_{\text{3D}} \in \mathbb{R}^{N \times d}) \tag{10}$$

These features represent each modality informed by the other. Global embeddings $\mathbf{E}_{\text{text}}, \mathbf{E}_{\text{3D}} \in \mathbb{R}^{d'}$ are obtained by pooling (Pool) and non-linear transformations ($g_{\text{text}}, g_{\text{3D}}$):

$$\mathbf{E}_{\text{text}} = g_{\text{text}}(\text{Pool}(\mathbf{Z}_{\text{text}})) \quad \mathbf{E}_{\text{3D}} = g_{\text{3D}}(\text{Pool}(\mathbf{Z}_{\text{3D}})) \tag{11}$$

$$\hat{\mathbf{E}}_{\text{text}} = \mathbf{E}_{\text{text}} / \|\mathbf{E}_{\text{text}}\|_2 \qquad \hat{\mathbf{E}}_{\text{3D}} = \mathbf{E}_{\text{3D}} / \|\mathbf{E}_{\text{3D}}\|_2 \tag{12}$$

Contrastive Learning Objective

Using the InfoNCE loss (van den Oord, Li, and Vinyals 2019), the text-to-3D loss ($\mathcal{L}_{\text{t2v}}$) is for a batch of B pairs:

$$\mathcal{L}_{\text{t2v}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(\hat{\mathbf{E}}_{\text{text},i} \cdot \hat{\mathbf{E}}_{\text{3D},i} / \tau\right)}{\sum_{j=1}^{B} \exp\left(\hat{\mathbf{E}}_{\text{text},i} \cdot \hat{\mathbf{E}}_{\text{3D},j} / \tau\right)} \tag{13}$$

where τ is a temperature parameter. A symmetric loss $\mathcal{L}_{\text{v2t}}$ is for 3D-to-text alignment. The final bidirectional loss is:

$$\mathcal{L}_{\text{bidirectional}} = \mathcal{L}_{\text{t2v}} + \mathcal{L}_{\text{v2t}} \tag{14}$$

Minimizing this loss aligns positive pairs and separates negative pairs in the embedding space.

Overall Loss Function

Our overall training objective in Phase 2 is a bidirectional cross-modality contrastive loss computed with MCTS-refined attentions, with additional regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{bidirectional}} + \gamma \cdot \mathcal{L}_{\text{reg}} \tag{15}$$

where $\mathcal{L}_{\text{bidirectional}}$ uses embeddings derived from $\mathbf{A}_{\text{optimized}}$, $\mathcal{L}_{\text{reg}}$ includes regularization terms (e.g., weight decay) weighted by γ. The MCTS reward (Eq. 6) guides the *search* for the optimal attention matrix $\mathbf{A}_{\text{optimized}}$, which subsequently influences $\mathcal{L}_{\text{bidirectional}}$, rather than directly contributing to the gradient via backpropagation in this formulation.

Training Algorithm

We introduce our two-stage training algorithm of 3DAlign-DAER (details in Appendix), which contains two phases:

1^{st}-Phase: *Contrastive Pre-training*. We pretrain only with $\mathcal{L}_{\text{bidirectional}}$ to warmup and obtain the initial attention $\mathbf{A}_{\text{initial}}$ (Eq. 5). This establishes a baseline alignment, potentially fine-tuning encoders and training projection heads.

2^{nd}-Phase: *MCTS-Guided Optimization*. MCTS is introduced to refine attention. The training loop involves: (1) Forward pass for $\mathbf{F}_{\text{text}}, \mathbf{F}_{\text{3D}}$. (2) Compute $\mathbf{A}_{\text{initial}}$. (3) Run MCTS search starting from $\mathbf{A}_{\text{initial}}$ to find $\mathbf{A}_{\text{optimized}}$, using the current model for reward evaluation. (4) Generate final embeddings $\hat{\mathbf{E}}_{\text{text}}, \hat{\mathbf{E}}_{\text{3D}}$ using $\mathbf{A}_{\text{optimized}}$. (5) Compute $\mathcal{L}_{\text{bidirectional}}$ (Eq. 14). (6) Backpropagate the loss to update model parameters. (7) Update MCTS statistics. MCTS search is performed periodically (e.g., every $k = 10$ steps).

ERS for Efficient Inference

During inference, the full MCTS is computationally expensive. We propose ERS, a lightweight hierarchical search strategy for efficient retrieval. It navigates a pre-computed or dynamically built hierarchical index over the 3D database. When retrieving for a query text embedding $\mathbf{q} = \hat{\mathbf{E}}_{\text{text}}$, ERS selects child nodes a (representing sub-category s_a) using a modified UCT-like score:

$$\begin{aligned} \text{UCT}_{\text{Lite}}(s, a) &= \lambda_1 \cdot \text{sim}(\mathbf{q}, \boldsymbol{\mu}_{s_a}) \\ &+ \lambda_2 \cdot \frac{N_{\text{success}}(s, a)}{N(s, a) + \epsilon} + \lambda_3 \cdot \sqrt{\frac{\ln N(s)}{N(s, a) + \epsilon}} \end{aligned} \tag{16}$$

where $\text{sim}(\mathbf{q}, \boldsymbol{\mu}_{s_a})$ is the cosine similarity to the child's representative embedding $\boldsymbol{\mu}_{s_a}$, N_{success} counts past retrieval successes through a, N are visit counts, $\lambda_1, \lambda_2, \lambda_3$ are weights balancing similarity, historical success, and exploration, and ϵ prevents division by zero. This efficiently prunes the search space.

Experiments

We conduct a comprehensive evaluation of our 3DAlign-DAER across several key tasks. We further investigate the quality and generalization ability of the learned 3D representations. Furthermore, detailed ablation studies analyzing the contribution of different components and strategies (including MCTS optimization and reward functions), evaluations on open-world understanding (ScanNet), hyperparameter sensitivity analysis, and computational overhead assessments for both training and inference are provided in the Appendix. All experiments are performed on one A100 GPU.

Zero-shot Shape Classification

Experimental Setup We select three commonly used benchmarks: Objaverse-LVIS (Deitke et al. 2022), ModelNet40 (Wu et al. 2015), and ScanObjectNN (Uy et al. 2019). For comparison models, we choose ULIP-2 (Xue et al. 2024), ULIP, Uni3D-B, Uni3D-L, Uni3D-g, PointCLIP (Zhang et al. 2021), ReCon++-L (Qi et al. 2024), and OpenShape-PointBERT (Liu et al. 2023b) to compare against our model, with each model using its officially provided inference hyperparameter configuration. We follow the evaluation setup of OpenShape, sampling 10,000 points for point clouds from Objaverse-LVIS and ModelNet40. For ScanObjectNN, we use 2,048 colorless sampled points from the OBJ ONLY version, maintaining alignment with the configuration used for testing the baseline Uni3d model. Our model is pre-trained on Align-2M and uses the previously defined ERS search during inference. Specific hyperparameters can be found in Appendix "Implementation Details and Hyperparameters".

Experimental Results As shown in Table 1, our 3DAlign-DAER achieves a Top-1 accuracy of 55.8% on the Objaverse-LVIS dataset, significantly outperforming the second-best model, ReCon++-L, at 53.1%. On the ModelNet40 dataset, 3DAlign-DAER obtains a Top-1 accuracy of 88.5%, also leading all comparison models. On the ScanObjectNN dataset, 3DAlign-DAER reaches a Top-1 accuracy of

Method	Venue	Objaverse-LVIS (%)			ModelNet40 (%)			ScanObjectNN (%)		
		Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
PointCLIP	CVPR 2022	1.8	4.0	5.9	19.4	28.5	34.9	10.4	20.9	30.5
ULIP	CVPR 2023	6.1	13.5	17.8	60.5	79.1	84.5	51.4	71.0	80.3
OpenShape-PointBERT	NeurIPS 2023	46.7	69.0	77.1	84.5	96.4	98.1	52.1	79.6	88.8
ULIP-2	CVPR 2024	26.7	44.9	52.5	75.0	88.0	93.1	51.7	72.6	82.4
Uni3D-B	ICLR 2024	51.6	74.0	80.9	86.2	96.6	97.8	63.7	81.7	90.3
Uni3D-L	ICLR 2024	53.0	47.2	81.4	86.4	96.7	98.2	64.1	81.9	90.5
Uni3D-g	ICLR 2024	54.2	76.1	81.9	86.7	96.9	99.1	65.2	82.9	91.3
ReCon++-L (shapeLLM)	ECCV 2024	53.1	75.2	81.6	86.4	94.8	95.9	63.5	80.1	90.5
3DAlign-DAER (Ours)		**55.8**	**77.0**	**83.1**	**88.5**	**98.6**	**99.5**	**67.0**	**86.2**	**93.1**

Table 1: Zero-Shot 3D Classification performance comparison on Objaverse-LVIS, ModelNet40, and ScanObjectNN benchmarks.

Method	S2T			T2S		
	RR@1	RR@5	NDCG@5	RR@1	RR@5	NDCG@5
Text2Shape	0.83	3.37	0.73	0.40	2.37	1.35
Y2Seq2Seq	6.77	19.30	5.30	2.93	9.23	6.05
TriCoLo	16.33	45.52	12.73	10.25	29.07	19.85
Parts2Words	19.38	47.17	15.30	12.72	32.98	23.13
COM3D	20.03	48.32	15.62	13.12	33.48	23.89
Uni3D-B	22.51	51.39	17.05	14.57	36.28	25.04
Uni3D-L	25.02	53.55	18.71	15.21	37.57	27.13
Uni3D-g	26.23	54.71	19.42	16.51	38.32	28.06
SCA3D	27.22	55.56	19.04	16.67	38.90	28.17
3DAlign-DAER	**28.11**	**56.58**	**19.93**	**17.53**	**39.88**	**29.06**

Table 2: Comparison of cross-modal retrieval performance.

67.0%, showing a more pronounced advantage compared to Uni3D-g's 65.2% and ReCon++-L's 63.5%. Not only does 3DAlign-DAER lead in Top-1 accuracy, but it also comprehensively surpasses all comparison models in Top-3 and Top-5 accuracy metrics.

Cross-Modal Retrieval

Experimental Setup We conduct standard text-shape bidirectional retrieval experiments on the Text2Shape (Chen et al. 2018) dataset, including Shape-to-Text (S2T) and Text-to-Shape (T2S) retrieval. We adopt standard retrieval evaluation metrics: RR@1, RR@5, and NDCG@5. We select representative methods including Text2Shape, Y2Seq2Seq (Han et al. 2018), TriCoLo (Ruan et al. 2023), Parts2Words (Tang et al. 2023), COM3D (Wu et al. 2024), SCA3D (Ren et al. 2025), as well as Uni3D-B, Uni3D-L and Uni3D-g. Our 3DAlign-DAER utilizes the unified embedding space learned from pre-training on Align-2M and employs the ERS to efficiently match queries (text or shape) with items in the target gallery through hierarchical search to perform the cross-modal retrieval task. Specific hyperparameter settings can be found in the Appendix.

Experimental Results As shown in Table 2, our 3DAlign-DAER achieves new state-of-the-art performance on the Text2Shape cross-modal retrieval task. It obtains the highest RR@1 scores for both Shape-to-Text (S2T) at 28.11% and Text-to-Shape (T2S) at 17.53%. This performance surpasses prior works, strong Uni3D baselines, and notably improves upon the previous leading method, SCA3D, by approximately 0.9% absolute RR@1 gain in both directions, demonstrating the learning of highly discriminative joint embeddings. The outstanding performance of 3DAlign-DAER on the cross-modal retrieval task stems from the fine-grained cross-modal alignment achieved through its DAP for dynamic attention modulation. Compared to baseline models like Uni3D-g/L, our DAP in 3DAlign-DAER can delve deeper into mining and optimizing the subtle correspondences between text and local features of 3D shapes, thereby learning joint embeddings that are richer in semantic information and more sensitive to detailed features.

Attention Heatmap Visualization Comparison

To visually verify the effectiveness of our 3DAlign-DAER in focusing attention, we perform an attention map visualization analysis, which projects the attention weights generated by the model when processing 3D point cloud data onto a 2D plane, generating intuitive heatmaps. Comparing the heatmaps generated by our 3DAlign-DAER with those from baseline models like Uni3D (see Figure 3), one can observe that 3DAlign-DAER's attention activations are significantly more concentrated and precise, accurately focusing on the core semantic regions of the objects. For example, for a cup, attention is focused on the cup body and handle. For a chair, attention is focused on its contours and structure. This justifies that our DAP and MCTS optimization can achieve more refined and robust fine-grained alignment.

Large-Scale Text-to-3D Retrieval on ObjaverseXL

Experimental Setup To evaluate the performance of 3DAlign-DAER with its efficient retrieval strategy via ERS against several established baselines on a realistic, large-scale text-to-3D retrieval task, we utilize a subset of 1 million diverse 3D models from ObjaverseXL. Our 3DAlign-DAER is compared against Uni3D (Zhou et al. 2024) and

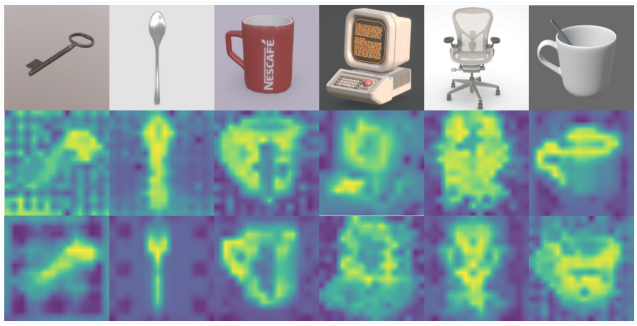

Figure 3: Attention Heatmap Visualization Comparison. Top: Original 3D objects. Middle: Attention heatmaps from Uni3D. Bottom: Attention heatmaps from 3DAlign-DAER.

Method	R@1 (%)	R@5 (%)	R@10 (%)
Uni3D-g + KNN	4.1	9.8	14.5
OpenDlign + KNN	3.8	9.2	13.9
Uni3D-g + FAISS-IVF	33.2	55.4	65.8
Uni3D-g + FAISS-HNSW	35.8	58.1	68.2
Uni3D-g + DiskANN	36.5	59.0	69.1
3DAlign-DAER + ERS	**48.5**	**69.2**	**78.6**

Table 3: Text-to-3D retrieval performance on the ObjaverseXL 1M subset. 3DAlign-DAER with ERS surpasses traditional methods and strong ANN baselines.

OpenDlign (Mao, Jing, and Mikolajczyk 2024) via standard k-Nearest Neighbors (KNN) retrieval. Furthermore, to benchmark against SOTA approximate search techniques, we include comparisons with leading ANN libraries, i.e., FAISS-IVF (Douze et al. 2025), FAISS-HNSW (Douze et al. 2025), and DiskANN (Jayaram Subramanya et al. 2019), which apply to the embeddings generated by the Uni3D baseline. We use the standard retrieval metrics, i.e., Recall@1 (R@1), Recall@5 (R@5), and Recall@10 (R@10).

Experimental Results Table 3 reports text-to-3D retrieval results on the 1M-scale ObjaverseXL. Traditional KNN—used in Uni3D and OpenDlign—performs poorly (R@1 ¡ 5%) due to its reliance on local proximity. Strong ANN methods (FAISS-IVF/HNSW, DiskANN) improve R@1 to 33.2–36.5%, but remain limited by the same locality assumption. In contrast, 3DAlign-DAER with ERS achieves significant gains, reaching **48.5%** R@1, **69.2%** R@5, and **78.6%** R@10. This large margin demonstrates that ERS exploits hierarchical, semantically guided exploration of the embedding space—beyond purely local neighbor search—enabling accurate retrieval of globally relevant candidates in the diverse 1M-scale ObjaverseXL dataset.

Few-shot Linear Probing

Experimental Setup To assess few-shot performance, we follow the standard linear-probing protocol: the pre-trained 3DAlign-DAER encoder is frozen, and only a linear classifier is trained on limited labeled samples. Experiments are conducted on Objaverse-LVIS with 1, 2, 4, 8, and 16 shots per class. We compare against representative baselines—including PointCLIP V2, ULIP, OpenShape (SparseConv and PointBERT), Uni3D, and the recent ReCon++-L (shapeLLM). All results are averaged over 10 random seeds for stability.

Method	**1-s**	**2-s**	**4-s**	**8-s**	**16-s**
PointCLIP V2	14.0	18.0	22.0	25.0	28.0
ULIP	6.0	11.0	20.0	26.0	32.0
ULIP Retrained	20.0	27.0	35.0	41.0	45.0
OpenShape-SparseConv	22.0	28.0	36.0	43.0	48.0
OpenShape-PointBERT	26.0	33.0	41.0	46.0	50.0
Uni3D-g	42.0	54.0	64.0	74.0	79.0
ReCon++-L (shapeLLM)	43.0	55.0	65.0	75.0	80.0
3DAlign-DAER (Ours)	**47.0**	**60.0**	**69.0**	**78.0**	**82.0**

Table 4: Few-shot Linear Probing on Objaverse-LVIS (Top-1 Accuracy %)

Experimental Results Table 4 compares 3DAlign-DAER with several advanced baselines, including ReCon++-L (shapeLLM), on few-shot linear probing (Top-1 accuracy). Across all settings (1, 2, 4, 8, 16 shots), 3DAlign-DAER achieves the best performance, consistently surpassing all competitors. The advantage is most pronounced in extremely data-scarce cases (1–2 shots), highlighting the high intrinsic quality, discriminability, and strong low-data generalization of the learned 3D representations—even when only a simple linear classifier is trained.

Conclusion

In this paper, we presented the 3DAlign-DAER framework for fine-grained alignment and large-scale sparse data retrieval in 3D-text cross-modal understanding. Besides, we construct and release the large-scale dataset Align3D-2M, containing 2 million high-quality sample pairs. Extensive and comprehensive experiments have validated the superiority of our 3DAlign-DAER and particularly demonstrated the significant advantages of our ERS in sparse retrieval scenarios. Our future work will focus on enhancing search efficiency and exploring broader applications.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62276283, in part by the China Meteorological Administration's Science and Technology Project under Grant CMA-JBGS202517, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515012985, in part by Guangdong-Hong Kong-Macao Greater Bay Area Meteorological Technology Collaborative Research Project under Grant GHMA2024Z04, in part by Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant 23hytd006, and in part by Guangdong Provincial High-Level Young Talent Program under Grant RL2024-151-2-11.

References

Cao, Y.; Zeng, Y.; Xu, H.; and Xu, D. 2023. CoDA: Collaborative Novel Box Discovery and Cross-modal Alignment for Open-vocabulary 3D Object Detection. In *NeurIPS*.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; Xiao, J.; Yi, L.; and Yu, F. 2015. ShapeNet: An Information-Rich 3D Model Repository. arXiv:1512.03012.

Chen, K.; Choy, C. B.; Savva, M.; Chang, A. X.; Funkhouser, T.; and Savarese, S. 2018. Text2Shape: Generating Shapes from Natural Language by Learning Joint Embeddings. arXiv:1803.08495.

Deitke, M.; Liu, R.; and et al., M. W. 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. *arXiv preprint arXiv:2307.05663*.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2022. Objaverse: A Universe of Annotated 3D Objects. *arXiv preprint arXiv:2212.08051*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *ACL*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; and Jégou, H. 2025. The Faiss library. arXiv:2401.08281.

Fu, H.; Jia, R.; Gao, L.; Gong, M.; Zhao, B.; Maybank, S.; and Tao, D. 2020. 3D-FUTURE: 3D Furniture shape with TextURE. arXiv:2009.09633.

Han, Z.; Shang, M.; Wang, X.; Liu, Y.-S.; and Zwicker, M. 2018. Y2Seq2Seq: Cross-Modal Representation Learning for 3D Shape and Text by Joint Reconstruction and Prediction of View and Word Sequences. arXiv:1811.02745.

Hegde, D.; Valanarasu, J. M. J.; and Patel, V. M. 2023. CLIP goes 3D: Leveraging Prompt Tuning for Language Grounded 3D Recognition. *arXiv preprint arXiv:2303.11313*.

Jayaram Subramanya, S.; Devvrit, F.; Simhadri, H. V.; Krishnawamy, R.; and Kadekodi, R. 2019. DiskANN: Fast Accurate Billion-point Nearest Neighbor Search on a Single Node. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *NeurIPS*, volume 32. Curran Associates, Inc.

Kabra, R.; Matthey, L.; Lerchner, A.; and Mitra, N. J. 2024. Leveraging VLM-Based Pipelines to Annotate 3D Objects. In *ICML*, volume 235 of *Proceedings of Machine Learning Research*. PMLR.

Le, T.-N.; and et al., T. V. N. 2023. TextANIMAR: Text-based 3D animal fine-grained retrieval. *Computers and Graphics*, 116: 162–172.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning. arXiv:2304.08485.

Liu, M.; Shi, R.; Kuang, K.; Zhu, Y.; Li, X.; Han, S.; Cai, H.; Porikli, F.; and Su, H. 2023b. OpenShape: Scaling Up 3D Shape Representation Towards Open-World Understanding. arXiv:2305.10764.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Lu, W.; Zhao, D.; Premebida, C.; Zhang, L.; Zhao, W.; and Tian, D. 2024. Multi-scale Feature Fusion with Point Pyramid for 3D Object Detection. arXiv:2409.04601.

Luo, T.; Johnson, J.; and Lee, H. 2024. View Selection for 3D Captioning via Diffusion Ranking. *arXiv preprint arXiv:2404.07984*.

Luo, T.; Rockwell, C.; Lee, H.; and Johnson, J. 2023. Scalable 3D Captioning with Pretrained Models. *arXiv preprint arXiv:2306.07279*.

Mao, Y.; Jing, J.; and Mikolajczyk, K. 2024. OpenDlign: Enhancing Open-World 3D Learning with Depth-Aligned Images. *arXiv preprint arXiv:2404.16538*.

OpenAI. 2024a. GPT-4o System Card. arXiv:2410.21276.

OpenAI. 2024b. OpenAI o1 System Card. arXiv:2412.16720.

Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. Deep Hough Voting for 3D Object Detection in Point Clouds. arXiv:1904.09664.

Qi, Z.; Dong, R.; Zhang, S.; Geng, H.; Han, C.; Ge, Z.; Yi, L.; and Ma, K. 2024. ShapeLLM: Universal 3D Object Understanding for Embodied Interaction. *arXiv preprint arXiv:2402.17766*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Ren, J.; and Wang, H. 2025. Enhanced Cross-modal 3D Retrieval via Tri-modal Reconstruction. arXiv:2504.01476.

Ren, J.; Wu, H.; Xiong, H.; and Wang, H. 2025. SCA3D: Enhancing Cross-modal 3D Retrieval via 3D Shape and Caption Paired Data Augmentation. *arXiv preprint arXiv:2502.19128*.

Ruan, Y.; Lee, H.-H.; Zhang, Y.; Zhang, K.; and Chang, A. X. 2023. TriCoLo: Trimodal Contrastive Loss for Text to Shape Retrieval. arXiv:2201.07366.

Sarkar, S. D.; Miksik, O.; Pollefeys, M.; Barath, D.; and Armeni, I. 2025. CrossOver: 3D Scene Cross-Modal Alignment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347.

Shen, H.; Zhang, Z.; Zhao, K.; Zhang, Q.; Xu, R.; and Zhao, T. 2025. VLM-R1: A stable and generalizable R1-style Large Vision-Language Model. https://github.com/om-ai-lab/VLM-R1. Accessed: 2025-02-15.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; and Hassabis, D. 2017. Mastering the game of Go without human knowledge. *Nature*, 550(7676): 354–359.

Sun, F.; Liu, Y.; Wang, J.-X.; and Sun, H. 2023. Symbolic Physics Learner: Discovering governing equations via Monte Carlo tree search. arXiv:2205.13134.

Sun, X.; Wu, J.; Zhang, X.; Zhang, Z.; Zhang, C.; Xue, T.; Tenenbaum, J. B.; and Freeman, W. T. 2018. Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling. In *CVPR*.

Tang, C.; Yang, X.; Wu, B.; Han, Z.; and Chang, Y. 2023. Parts2Words: Learning Joint Embedding of Point Clouds and Texts by Bidirectional Matching between Parts and Words. arXiv:2107.01872.

Uy, M. A.; Pham, Q.-H.; Hua, B.-S.; Nguyen, D. T.; and Yeung, S.-K. 2019. Revisiting Point Cloud Classification: A New Benchmark Dataset and Classification Model on Real-World Data. In *ICCV*.

van den Oord, A.; Li, Y.; and Vinyals, O. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Wang, F.; Mei, J.; and Yuille, A. 2023. SCLIP: Rethinking Self-Attention for Dense Vision-Language Inference. *arXiv preprint arXiv:2312.01597*.

Wu, H.; Li, R.; Wang, H.; and Xiong, H. 2024. COM3D: Leveraging Cross-View Correspondence and Cross-Modal Mining for 3D Retrieval. In *ICME*, 1–6.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3D ShapeNets: A deep representation for volumetric shapes . In *CVPR*, 1912–1920. Los Alamitos, CA, USA: IEEE Computer Society.

Xiang, Y.; Kim, W.; Chen, W.; Ji, J.; Choy, C.; Su, H.; Mottaghi, R.; Guibas, L.; and Savarese, S. 2016. ObjectNet3D: A Large Scale Database for 3D Object Recognition. In *ECCV*.

Xue, L.; Gao, M.; Xing, C.; Martín-Martín, R.; Wu, J.; Xiong, C.; Xu, R.; Niebles, J. C.; and Savarese, S. 2023. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *CVPR*, 1179–1189.

Xue, L.; Yu, N.; Zhang, S.; Panagopoulou, A.; Li, J.; Martín-Martín, R.; Wu, J.; Xiong, C.; Xu, R.; Niebles, J. C.; and Savarese, S. 2024. ULIP-2: Towards Scalable Multimodal Pre-training for 3D Understanding. arXiv:2305.08275.

Ye, L.; Rochan, M.; Liu, Z.; and Wang, Y. 2019. Cross-Modal Self-Attention Network for Referring Image Segmentation. arXiv:1904.04745.

Zhang, D.; Zhoubian, S.; Hu, Z.; Yue, Y.; Dong, Y.; and Tang, J. 2024. ReST-MCTS*: LLM Self-Training via Process Reward Guided Tree Search. arXiv:2406.03816.

Zhang, J.; Cai, K.; Fan, Y.; Liu, N.; and Wang, K. 2025a. MAT-Agent: Adaptive Multi-Agent Training Optimization. In *NeurIPS*.

Zhang, J.; Cai, K.; Fan, Y.; Wang, J.; and Wang, K. 2025b. CF-VLM: CounterFactual Vision-Language Fine-tuning. In *NeurIPS*.

Zhang, J.; Fan, Y.; Cai, K.; Huang, Z.; Sun, X.; Wang, J.; Tang, C.; and Wang, K. 2025c. DrDiff: Dynamic Routing Diffusion with Hierarchical Attention for Breaking the Efficiency-Quality Trade-off. arXiv:2509.02785.

Zhang, J.; Huang, Z.; Fan, Y.; Liu, N.; Li, M.; Yang, Z.; Yao, J.; Wang, J.; and Wang, K. 2025d. KABB: Knowledge-Aware Bayesian Bandits for Dynamic Expert Coordination in Multi-Agent Systems. arXiv:2502.07350.

Zhang, R.; Guo, Z.; Zhang, W.; Li, K.; Miao, X.; Cui, B.; Qiao, Y.; Gao, P.; and Li, H. 2021. PointCLIP: Point Cloud Understanding by CLIP. *arXiv preprint arXiv:2112.02413*.

Zhou, J.; Wang, J.; Ma, B.; Liu, Y.-S.; Huang, T.; and Wang, X. 2024. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations (ICLR)*.

Świechowski, M.; Godlewski, K.; Sawicki, B.; and Mańdziuk, J. 2022. Monte Carlo Tree Search: a review of recent modifications and applications. *Artificial Intelligence Review*, 56(3): 2497–2562.