

# Open3DSearch: Zero-Shot Precise Retrieval of 3D Shapes Using Text Descriptions

Xiong Li  
Zhejiang University of Technology  
Hangzhou, China  
lx.3958@gmail.com

Yikang Yan  
Zhenyu Wen\*  
Zhejiang University of Technology  
Hangzhou, China  
ligeimeiyuao@gmail.com  
wenluke427@gmail.com

Qin Yuan  
Beijing Institute of Technology  
Beijing, China  
yuanq1020@gmail.com

Fangda Guo  
Institute of Computing Technology,  
Chinese Academy of Sciences  
Beijing, China  
guofangda@ict.ac.cn

Zhen Hong  
Zhejiang University of Technology  
Hangzhou, China  
zhong1983@zjut.edu.cn

Ye Yuan  
Beijing Institute of Technology  
Beijing, China  
yuan-ye@bit.edu.cn

## ABSTRACT

With the rapid growth of 3D content, there is an increasing need for intelligent systems that can search for complex 3D shapes using simple natural language queries. However, existing approaches face significant limitations. They rely heavily on manually labeled datasets and use fixed similarity thresholds to determine matches, which restricts their ability to generalize and accurately retrieve novel or diverse 3D shapes. To bridge these gaps, this paper introduces *Open3DSearch*, the first attempt to address the challenge of open-domain text-to-shape precise retrieval. Our core idea is to transform 3D shapes into semantically representative 2D views, thereby enabling the task to be handled by mature large vision-language models (LVLMs) and allowing for explicit cross-modal matching judgments. To realize this concept, we design a view rendering strategy to mitigate potential information degradation during 3D-to-2D conversion while capturing the maximal amount of query-relevant information. To evaluate *Open3DSearch* and advance research in this field, we present the Uni3D-R benchmark dataset, designed to simulate precise associations between user queries and 3D shapes in open-domain contexts. Extensive quantitative and qualitative experiments demonstrate that *Open3DSearch* achieves state-of-the-art results.

## CCS CONCEPTS

• Information systems → Multimedia and multimodal retrieval.

\* Corresponding Author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '25, October 27–31, 2025, Dublin, Ireland.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 979-8-4007-2035-2/2025/10  
<https://doi.org/10.1145/3746027.3755533>

## KEYWORDS

Text-to-Shape Retrieval, Zero-Shot Precise Retrieval, Large Vision-Language Models

### ACM Reference Format:

Xiong Li, Yikang Yan, Zhenyu Wen, Qin Yuan, Fangda Guo, Zhen Hong, and Ye Yuan. 2025. Open3DSearch: Zero-Shot Precise Retrieval of 3D Shapes Using Text Descriptions. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, October 27–31, 2025, Dublin, Ireland. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3746027.3755533>

## 1 INTRODUCTION

With the rapid advancement of 3D content creation tools, 3D shapes have become essential assets in fields such as VR/AR, industrial design, and digital entertainment. Statistics show that major 3D shape platforms (e.g., Sketchfab, 3D Warehouse) now host millions of shapes, with numbers continuing to grow [13]. However, traditional tag-based retrieval systems struggle to help users find specific shapes efficiently. For example, a designer might search for "a modern chair suitable for Nordic-style living rooms, featuring wooden textures and a curved backrest." Such queries involve style, material, and structural characteristics, and employ highly flexible and open vocabulary that far exceeds the range of tags traditional systems can handle. Users often expend significant effort refining keywords and filtering through numerous similar results, rendering the process time-consuming and inefficient.

Figure 1 illustrates our research vision—open-domain text-to-shape precise retrieval—which enables users to directly acquire 3D shapes matching their needs through free-text descriptions. Previous approaches [7, 13, 20, 31, 32, 37] focus on learning joint representations for text and 3D data, mapping both of them into a shared embedding space and performing retrieval based on similarity. However, these methods exhibit two fundamental limitations: First, these methods depend on manually annotated fine-grained paired data for training, which inherently struggles with zero-shot scenarios. The high cost of such annotation limits the breadth of data coverage, making it infeasible to represent all object categories in open-domain contexts. Consequently, these methods often fail to generalize to unseen categories. Second, these methods rely on

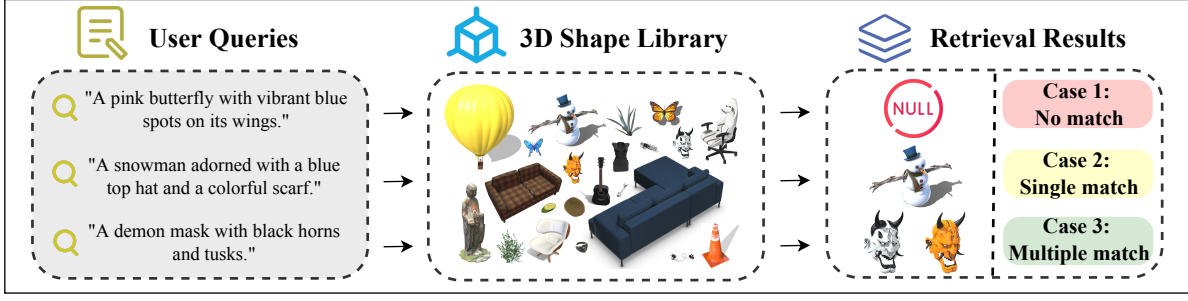


Figure 1: A demonstration of 3D Shape Retrieval Based on User-Provided Text Descriptions.

a fixed similarity threshold to determine whether a shape matches a query. However, in our task, the decision is based on semantic similarity, which is inherently ambiguous and lacks well-defined boundaries—particularly when dealing with 3D shapes. Additionally, similarity in the embedding space often lacks a consistent and transferable structure, making threshold-based decisions unreliable in zero-shot scenarios.

**Our solution.** Recent advances in large vision–language models (LVLMs), such as GPT-4V[43] and Qwen-VL[4], have demonstrated robust zero-shot image–text reasoning by leveraging internet-scale training data, thereby enabling open-domain cross-modal matching. With well-designed prompts, these models can dynamically attend to query-specific information and conduct explicit matching analyses, mitigating decision rigidity and uncertainty associated with fixed similarity thresholds. Our key idea is to render 3D shapes into 2D views, allowing mature 2D-LVLMs to perform matching judgments.

The key challenge in realizing this idea is obtaining rendered views that accurately capture the query-described information despite the complexity of the pose search space and the lack of explicit objective guidance. Furthermore, retrieving a 3D shape from a large database can also introduce significant computational overhead. This arises from two factors: the sheer volume of stored 3D shapes and the need to render each shape multiple times into 2D representations from diverse viewing angles.

To address the aforementioned challenge, we introduce *Open3DSearch*. To effectively capture informative views for matching decisions, we first apply constraints on the camera’s position and orientation—based on the principles of perspective projection—to ensure the entire 3D shape remains visible. Based on this, we generate a sparse but representative set of candidate poses and render the corresponding views. We then evaluate the amount of query-relevant information retained in each view via semantic similarity measurements, which guide the selection of images for input into a 2D-LVLM. To reduce per-query computational overhead and improve retrieval efficiency, we introduce a two-stage strategy that limits the number of matching verifications. Upon receiving a query, repository 3D shapes are first prioritized and then undergo sequential verification with an early-stopping mechanism, terminating the process once a sufficient number of matches is found.

Our contributions are summarized as follows:

**1) Task & Dataset:** To the best of our knowledge, we are the first to introduce the open-domain text-to-shape precise retrieval task,

which allows users to accurately retrieve a wide range of arbitrary 3D shapes based solely on natural language descriptions, without being limited to predefined categories or domains. To facilitate this research, we construct the Uni3D-R benchmark dataset. This dataset includes 7,855 high-quality 3D shapes and 812 queries representing diverse matching scenarios.

**2) Methodology:** We propose *Open3DSearch*, which pioneers the integration of LVLMs into 3D shape retrieval, achieving zero-shot explicit cross-modal matching. By leveraging joint constraints of geometry and semantics, *Open3DSearch* can effectively capture key visual features in queries and generate discriminative rendered views. Furthermore, an efficient hierarchical retrieval strategy is implemented to dynamically optimize the matching-validation process, thereby significantly reducing computational overhead.

**3) Extensive Experiments:** To validate the robustness of our approach, we conduct extensive experimental evaluations. Results demonstrate that *Open3DSearch* achieves state-of-the-art retrieval performance in both target-present and target-absent scenarios, outperforming all customized baseline methods.

## 2 RELATED WORK

### 2.1 Text-3D Retrieval via Joint Embedding Learning

To support research in text-3D cross-modal retrieval, Chen et al.[7] constructed the first 3D shape retrieval dataset with natural language descriptions based on ShapeNet[6] tables and chairs. They were also the first to introduce joint embedding learning for this task. Subsequent studies [13, 31, 32, 37] improved retrieval performance within this paradigm but remained constrained in open-domain scenarios due to insufficient training data. With breakthroughs in vision language pre-trained models like CLIP[30] and ALIGN[19], some studies [14, 16, 25, 28, 39, 40, 45] attempt to align the 3D modality to CLIP’s embedding space to inherit its general representation ability acquired from billion-scale image-text pre-training data. These methods demonstrate robustness in zero-shot 3D shape classification but remain inadequate for precise retrieval due to: 1) They focus on global feature embeddings while neglecting the modeling of fine-grained attributes (e.g., part structures, material textures), making it challenging to differentiate candidate shapes with similar overall but local differences; 2) The retrieval process relies on distance metrics in an implicit embedding space, lacking an explicit interpretable mechanism to support accurate

matching judgments. In contrast, our work introduces an explicit matching paradigm that leverages LVLMs for zero-shot reasoning. By transforming 3D shapes into informative 2D views and enabling LVLMs to perform fine-grained semantic verification, we circumvent the dependency on large-scale training data and overcome the limitations of global-embedding-based retrieval.

## 2.2 Large Vision-Language Models

The breakthrough of Large Language Models[5, 11, 34, 41, 42] has revolutionized multimodal intelligence systems, particularly in developing LVLMs [3, 4, 8, 10, 21–23, 26, 43, 44] with open-domain visual reasoning capabilities. In 2D vision, early works such as BLIP-2[22] and Flamingo[3] bridged the gap from image understanding to visual reasoning by connecting visual encoders with LLMs. State-of-the-art 2D-LVLMs, including GPT-4V[43] and Qwen-VL[4], can now handle more sophisticated tasks involving spatial relationships and attribute reasoning through chain-of-thought prompting[35]. In contrast, 3D-LVLM development faces two major obstacles: 1) The complex topology[2, 17] of 3D data widens the semantic gap, making fine-grained correlations difficult; 2) The scarcity of large-scale instruction-following data severely restricts the alignment quality of the latent space and the model’s adherence to human intentions. Existing attempts [15, 29, 33, 38] still show notable limitations in 3D semantic understanding, especially in open-domain dialogue. For cross-modal retrieval, they often fail to decouple and verify the correspondence between attributes and candidate shapes, resulting in frequent mismatches. Instead of training a 3D-LVLM, we bridge the 3D-to-2D gap by generating representative rendered views, allowing mature 2D-LVLMs to perform zero-shot matching judgments.

## 3 METHOD

### 3.1 Problem Definition

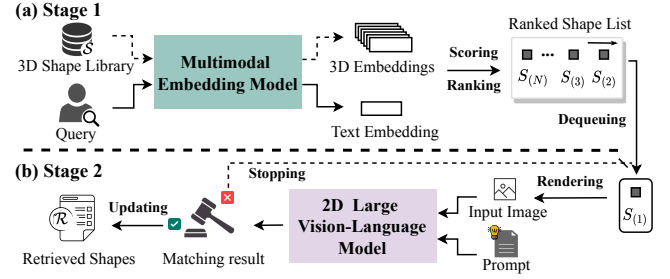
**Open-Domain Text-to-Shape Precise Retrieval (OD-T2SPR).** OD-T2SPR can be defined as a multi-modal retrieval problem. Given a library  $\mathcal{S} := \{S_1, S_2, \dots, S_N\}$  containing  $N$  3D shapes and an arbitrary text query  $q$ , the objective is to retrieve all shapes that match the description of query such that the retrieval result  $\mathcal{R}_q \subseteq \mathcal{S}$  aligns with ground truth  $\mathcal{G}_q \subseteq \mathcal{S}$ .  $|\mathcal{G}_q|$  can be zero (i.e., no match target), one, or greater than one (up to  $N$ ).

### 3.2 Overview

Figure 2 illustrates the overview of *Open3DSearch*, which consists of two stages. First, it prioritizes the 3D shapes from the entire library. For each shape  $S_i \in \mathcal{S}$ , its retrieval priority is determined by the semantic similarity to the query text  $q$ , with higher similarity corresponding to elevated ranking. The semantic embeddings  $E(q)$ ,  $\{E(S_i)\}_{i=1}^N$  used for this similarity computation are derived from OpenShape[25], a multimodal embedding model that supports text, image, and 3D point cloud modalities. The similarity is computed as follows:

$$\text{sim}(S_i, q) = \frac{\langle E(S_i), E(q) \rangle}{\|E(S_i)\| \|E(q)\|} \quad (1)$$

where  $\langle \cdot, \cdot \rangle$  denotes the inner product, and  $\|\cdot\|$  indicates the norm. Based on this, the ranked **shape** list is formalized as  $\mathcal{C} :=$



**Figure 2: The overall retrieval process of *Open3DSearch* is divided into two stages: a rough ranking based on semantic embeddings and a fine-grained, instance-by-instance verification based on image-text analysis.**

$\{S_{(1)}, S_{(2)}, \dots, S_{(N)}\}$ , where  $\text{sim}(S_{(1)}, q) \geq \text{sim}(S_{(2)}, q) \geq \dots \geq \text{sim}(S_{(N)}, q)$ .

Second, we conduct fine-grained matching verification for each selected 3D shape. Using our view rendering strategy (§3.3), each 3D shape  $S_{(i)} \in \mathcal{C}$  is converted into a set of representative 2D views. These views are then selectively analyzed by a 2D-LVLM, guided by our tailored prompt (§3.4). To improve efficiency, the retrieval process includes a dynamic stopping criterion: the system halts further verification if it encounters  $u$  consecutive negative matches. The stopping index  $t$  is formally defined as:

$$t = \min \left\{ t \in [u, N] \mid \bigwedge_{i=t-u+1}^t \text{Match}(S_{(i)}, q) = \text{No} \right\} \quad (2)$$

where  $\text{Match}(\cdot)$  denotes the binary judgment function.

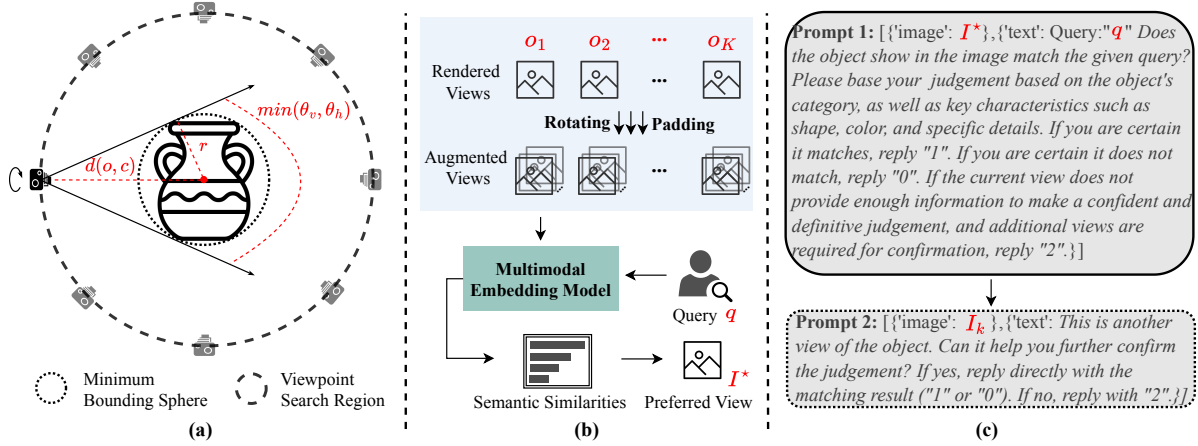
### 3.3 Semantic-Aware View Rendering

To meet the input requirements of 2D-LVLMs, we first render the 3D shape into 2D views. The rendering pose—defined by a rotation matrix  $R \in \text{SO}(3)$  and a position vector  $o \in \mathbb{R}^3$ —plays a crucial role in determining view quality. Poorly chosen poses may cause key features to be occluded or push the object out of frame, leading to missing critical information for accurate matching. To overcome this, our view rendering strategy first applies a physical visibility constraint (Figure 3(a)) to limit the pose search space. Then, it introduces a semantic consistency constraint (Figure 3(b)) to guide pose selection. This ensures that the final view not only fully captures the 3D shape but also aligns semantically with the query, improving the reliability of matching results.

**Physical Visibility Modeling.** We meet three key criteria to ensure the 3D shape is clearly and fully presented in the rendered image: 1) The object is centered to draw viewer attention; 2) The entire object is visible, avoiding any cropping or information loss; 3) The object fills most of the image, maximizing resolution and detail visibility.

For Condition 1, we constrain the camera optical axis (Z-axis) to point towards the center  $c$  of the shape’s minimum bounding sphere, as defined below:

$$\min_{c, r} \{r \mid \forall p \in S, \|p - c\| \leq r\} \quad (3)$$



**Figure 3: Core technical details of *Open3DSearch*. (a): Planar projection schematic of camera-to-3D shape spatial relationships. (b): Standardized workflow for preferred view acquisition. (c) Feedback-enabled hierarchical prompting framework.**

where  $p$  denotes a point in the shape and  $r$  is the radius of the minimum bounding sphere. Here, we employ the Welzl algorithm[36] to compute the minimum bounding sphere. Next, we adjust the camera distance  $d_{(o,c)}$  to satisfy Conditions 2 and 3. Calculating the optimal distance based on the 3D shape itself is complex and inefficient, as the irregularity of the shape would necessitate recalculations whenever the camera pose changes. An efficient alternative is to use the precomputed minimum bounding sphere as a reference. The optimal distance is derived via:

$$d_{(o,c)} = \max \left( \frac{r}{\sin(\theta_v/2)}, \frac{r}{\sin(\theta_h/2)} \right) \quad (4)$$

where  $\theta_v$  and  $\theta_h$  denote the camera's vertical and horizontal field of view, respectively. In this way, regardless of how the camera pose changes, the optimal distance  $d_{(o,c)}$  remains consistent, significantly reducing the computational burden. Ultimately, the search space for the camera position is confined to a spherical region centered at the center  $c$ , with the optimal distance  $d_{(o,c)}$  as the radius.

**Semantic Consistency Constraint.** After determining the region for the camera's position, the next step is to optimize the specific position and the rotation angle around the Z-axis to render an image that maximally supports the matching judgment between the 3D shape and the query text. In general, visual systems exhibit a certain degree of tolerance to changes in viewpoint, and the semantic differences between adjacent viewpoints are almost negligible. Therefore, we do not perform continuous searches on the sphere to determine the optimal position. Instead, we discretely sample multiple positions and select the best one. Specifically, we use the HEALPix grid[12] to sample the sphere uniformly, ensuring comprehensive coverage of the shape's various directions with fewer sampled positions. This process can be formalized as follows:

$$O = \text{HEALPix} \left( c, d_{(o,c)}, K \right) \quad (5)$$

where  $O := \{o_k \in \mathbb{R}^3, 1 \leq k \leq K\}$  denotes the set of coordinates for the sampled positions, and  $K := 12 \times w^2$  is the number of samples, determined by the set resolution  $w \in \mathbb{Z}^{\geq 1}$ . For each candidate

position  $o_k \in O$ , we randomly initialize a camera rotation and render the corresponding image  $I_k$  based on this. The rotation matrix  $R_k$  is obtained through Schmidt orthogonalization to ensure its rationality, with the calculation defined as follows:

$$R_k = [\alpha_k, \beta_k, \gamma_k]^T \quad (6)$$

$$\gamma_k = \frac{c - o_k}{\|c - o_k\|}, \quad \beta_k = \frac{\eta - \langle \eta, \gamma_k \rangle \gamma_k}{\|\eta - \langle \eta, \gamma_k \rangle \gamma_k\|}, \quad \alpha_k = \beta_k \wedge \gamma_k \quad (7)$$

where  $\alpha_k$ ,  $\beta_k$  and  $\gamma_k$  denote the unit vectors along the  $x$ ,  $y$ , and  $z$  axes of the camera coordinate system, respectively.  $\wedge$  denotes the cross product, and  $\eta$  is a random unit vector that is not collinear with the  $z$ -axis (i.e.,  $\langle \eta, \gamma \rangle \neq \pm \|\eta\| \|\gamma\|$ ). In theory, the image rendered from the optimal camera position should be semantically closest to the query text, and the embedding similarities  $\{sim(I_k, q)\}_{k=1}^K$  can be computed using Equation 1 to determine this position. However, the 3D shape may present unnatural orientations (e.g., inverted) in the rendered images at certain positions due to camera rotation angle issues, which could significantly impact the multimodal embedding model's understanding and encoding of semantics. To mitigate this impact, we capture images at multiple rotation angles for each camera position and compute the average semantic similarity, thereby robustly selecting the optimal position. Instead of repeatedly rendering images at different camera rotation angles, we perform rotation augmentation on the initially rendered image, thus reducing the number of renderings and improving computational efficiency. The calculation of the average semantic similarity for the  $k$ -th position can be expressed as follows:

$$sim_{avg}(I_k, q) = \frac{1}{G} \sum_{g=1}^G sim(I_k^{\theta_g}, q), \quad I_k^{\theta_g} = \text{Rotate}(I_k, \theta_g) \quad (8)$$

where  $G$  denotes the number of rotation angles,  $\theta_g$  denotes the  $g$ -th rotation angle, and  $\text{Rotate}(\cdot, \cdot)$  indicates the rotation operation applied to the image. The position with the highest average semantic similarity is determined to be the optimal one. Then, we select the image with the highest semantic similarity from all rotation angle images at that position and use it as the preferred view for input to



the 2D-LVLM. This process is formalized as follows:

$$I^* = \arg \max_{I_*^{\theta_g}} \text{sim} \left( I_*^{\theta_g}, q \right), \quad I_* = \arg \max_{I_k} \text{sim}_{\text{avg}} (I_k, q) \quad (9)$$

### 3.4 Feedback-Enabled Prompt

Given the selected image  $I^*$  and query text  $q$ , we guide the 2D-LVLM to perform matching judgments through customized prompts. To support a fully automated retrieval process and improve the reliability of judgment results, the design of prompts should address two critical requirements: 1) During the retrieval process, the 2D-LVLM is required to engage in high-frequency reasoning. The prompts should help suppress potential hallucinations[24], ensuring that it consistently provides effective responses; 2) A single view may provide insufficient semantic information, making accurate judgment difficult. The prompt should guide the 2D-LVLM in performing an enhanced analysis of inherent uncertainties in complex cases.

To enhance reliability and reduce output variability in the open-domain text-to-shape retrieval task, we reformulate the open-ended response generation into a fixed-option classification problem with a strict output format. This not only curbs randomness in the 2D-LVLM's outputs but also introduces an essential "uncertain" option in addition to binary choices like "match" or "non-match." The uncertainty option is key to our feedback design: when the model lacks confidence and selects "uncertain," it triggers a feedback loop that incorporates additional views of the 3D shape for repeated analysis. This iterative refinement enables the model to improve its judgment and converge on a more confident and accurate result—especially valuable in open-domain settings where inputs can be highly varied and ambiguous. To maintain efficiency, the feedback mechanism draws from a queue of  $K - 1$  basic rendering views (excluding the primary view  $I_*$ ), randomly ordered to ensure diverse visual perspectives without adding significant computational cost.

## 4 UNI3D-R DATASET

Table 1 summarizes existing publicly available datasets for 3D shape retrieval. However, these datasets present two major limitations that make them insufficient for evaluating open-domain, fine-grained retrieval tasks: 1) The 3D shapes cover a narrow range of semantic categories, limiting diversity; 2) The relationships between query texts and 3D shapes are either oversimplified or poorly annotated, failing to capture complex matching scenarios. To address these gaps and enable robust evaluation, we introduce the Uni3D-R dataset, which comprises 7,855 3D shapes and 812 carefully curated query texts. Among these, 195 queries have no correct match, 262 have a single correct match, and 355 correspond to multiple valid targets. The dataset construction followed a structured pipeline combining public data sources with manual refinement and verification, carried out in three key steps.

**3D Shape Collection.** We collect 3D shapes from the Objaverse[9] dataset. Objaverse[9] provides over 800K 3D shapes primarily sourced from user uploads, which exhibit a high degree of randomness and diversity. To ensure data quality, the collection process was performed by human annotators who were instructed to select geometrically complete objects with vivid colors. Ultimately, the 3D library consists of 7,855 carefully selected 3D shapes.

**Query Text Collection.** After constructing the [shape](#) repository, we collect query texts using the Cap3D[27] dataset. Cap3D[27] employs GPT-4[1] to generate descriptive texts for shapes in Objaverse[9]. We first extract the subset corresponding to the pre-selected 3D shapes from Cap3D[27], and then human annotators review and select entries that conform to human descriptive styles. In the end, 812 texts were chosen to form the final query set.

**Matching Relation Annotation.** For the collected 3D shapes and query texts, we employ a two-stage annotation pipeline analogous to the proposed retrieval framework to establish their matching relationships. For each query, after obtaining the ranked 3D shape list  $C$ , human annotators sequentially verify matching results. The validation process is terminated when an obvious "stagnation" emerges (i.e., difficulty in identifying more matches), indicating that the matching targets for the current query had been essentially determined. Each query could match zero, one, or multiple targets. To ensure annotation objectivity and consistency, each query is independently evaluated by at least three annotators, with disputed cases undergoing final adjudication through majority voting.

## 5 EXPERIMENTS

### 5.1 Experimental Setup

Previous text-to-shape retrieval methods[7, 13, 31, 32, 37] are typically trained on closed categories (e.g., tables and chairs), lacking generalization capabilities and thus struggling to handle free descriptions in open-domain contexts. Moreover, their adopted evaluation metrics (Recall Rate at  $k$ [7] and Normalized Discounted Cumulative Gain[18]) inherently incorporate two critical assumptions incompatible with our task: 1) Each query contains at least one ground-truth match; 2) The retrieval system only returns a ranked list without explicitly determining matching relationships.

To enable a fair and meaningful comparison, we redesign the evaluation metrics and develop adapted baseline methods that better align with the nature of our task.

**5.1.1 Evaluation Metrics.** We adopt a divide-and-conquer evaluation framework that defines differentiated metrics based on the ground-truth matching status of the query.

**Match Absent.** When no shapes in the library semantically match the query, the system should refrain from returning irrelevant results. To evaluate this behavior, we introduce two metrics: False Positive Rate (FPR) and Average False Positives (AFP). FPR measures the likelihood of the system returning incorrect results, while AFP captures the average number of such irrelevant returns. The corresponding calculation formulas are as follows:

$$FPR = \frac{|\hat{Q}_0|}{|Q_0|}, \quad \hat{Q}_0 = \{q \mid \mathcal{R}_q \neq \emptyset, q \in Q_0\} \quad (10)$$

$$AFP = \frac{1}{|\hat{Q}_0|} \sum_{q \in \hat{Q}_0} |\mathcal{R}_q| \quad (11)$$

where  $Q_0$  and  $\mathcal{R}_q$  denote the set of queries with zero matches and the retrieval results of query  $q$ , respectively.

**Match Present.** When at least one matching shape exists in the library, the system should return all relevant results and avoid misjudgments and omissions. We quantify the retrieval accuracy

**Table 1: Comparison with Existing 3D shape Retrieval Datasets.**

Dataset	3D Shapes			Query Texts	
	Number	Scope	Source	Number	Match-Distribution [0 match/1 match/>1 matches]
Text2Shape[7]	15058	Chair,Table	Synthetic	75344	[0/75344/0]
TextANIMAR[20]	186	Animals	Real	150	[0/150/0]
Uni3D-R	7855	Not Specified	Hybrid	812	[195/262/355]

through modified precision, recall, and F1-score metrics, as shown below:

$$Precision = \frac{1}{|Q_{\exists}|} \sum_{q \in Q_{\exists}} \frac{|\mathcal{R}_q \cap \mathcal{G}_q|}{\max(\epsilon, |\mathcal{R}_q|)} \quad (12)$$

$$Recall = \frac{1}{|Q_{\exists}|} \sum_{q \in Q_{\exists}} \frac{|\mathcal{R}_q \cap \mathcal{G}_q|}{|\mathcal{G}_q|} \quad (13)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (14)$$

where  $Q_{\exists}$  and  $\mathcal{G}_q$  denote the set of queries with matches and the ground-truth matching results of query  $q$ , respectively.  $\epsilon$  is a small positive constant ensuring non-zero denominators.

**5.1.2 Baselines.** To demonstrate the advantages of our proposed pipeline, we construct the following two baseline methods as competitors.

**Similarity Threshold-based Retrieval (ST-R).** This method utilizes a multi-modal embedding model (i.e., OpenShape[25]) to compute the semantic similarity between the query text and each shape in the 3D shape library, determining matching objects based on a set similarity threshold. Those with similarity scores exceeding this threshold are considered matches, while the rest are non-matches. Our observations indicate that the similarity scores for matched pairs predominantly fall between 0.12 to 0.22. Consequently, we uniformly sample five equally spaced thresholds within this interval for testing. Furthermore, we employ a small number of queries to identify an optimal threshold. Specifically, we randomly select 50 queries with ground-truth matches and then evaluate F1 scores by setting thresholds from 0.12 to 0.22 in increments of 0.001. The optimal threshold ( $\xi = 0.164$ ) was identified as the value that yielded the highest F1 score.

**3D LVLM-based Retrieval (3D-LVLM-R).** This method employs a representative 3D-LVLM (e.g., PointLLM[38], ShapeLLM[29], MiniGPT-3D[33]) to verify the matching relationship between the 3D shapes and the queries. To ensure a fair comparison, this method follows the same two-stage retrieval process as *Open3DSearch*, where matching decisions are made sequentially after an initial ranking, and the retrieval stopping condition remains identical. These models can directly process complete 3D shapes (provided in point cloud format). Hence, we use a simple prompt: `[[{'point cloud':  $S_i$ }, {'text': 'Query: "q" Does this 3D object match the given query? Please reply "Yes" or "No".'}]]`

**5.1.3 Implementation Details.** All experiments are conducted on an Ubuntu server equipped with an NVIDIA L20 GPU. In the experiments, we use the Trimesh library provided by Python to load 3D shapes and perform rendering based on predefined camera poses. The resolution of the rendered images is set to  $800 \times 600$ . The

2D-LVLM employs the Qwen-VL[4] Max version, performing on-line inference via API. OpenShape[25] and MiniGPT-3D[33] are deployed and executed locally using open-source code. The parameter  $u$  for terminating retrieval and the resolution  $w$  for HEALPix grid sampling are set to 5 and 1, respectively.

## 5.2 Experimental Results

**5.2.1 Quantitative Results.** Table 2 summarizes the quantitative evaluation results comparing *Open3DSearch* with baseline methods. Overall, *Open3DSearch* demonstrates the most comprehensive performance across all metrics. When ground-truth matches are absent, our method achieves only an FPR of 0.262 and an AFP of 1.706, indicating its exceptional capability in rejecting incorrect matches. In contrast, while ST-R gains some advantages by adopting high thresholds (e.g.,  $\xi = 0.220$ ), this comes at the expense of significant losses in precision and recall. These results reveal the inherent limitation of ST-R in balancing diverse query scenarios through a fixed threshold. **For 3D-LVLM-R, all three variants yield an FPR of 1 and AFP values in the thousands, suggesting severe hallucinations in shape-text matching, where numerous incompatible 3D shapes are mistakenly identified as matches.** When ground-truth matches exist, *Open3DSearch* outperforms ST-R ( $\xi = 0.164$ ) with 26 and 18.5 percentage point improvements in precision and recall, respectively, while achieving a better balance between these metrics. This further validates its adaptive capability in addressing heterogeneous queries. Although 3D-LVLM-R attains a near-perfect recall, its near-zero precision and F1-score severely compromise overall performance.

**5.2.2 Qualitative Results.** To more clearly illustrate the strengths and limitations of each method, we provide several query examples in Figure 4. For 3D-LVLM-R, the system exhibits near-total failure in rejecting non-matching shapes, with a particularly pronounced prevalence of false matches among top-ranked targets. This observation aligns with the hallucination issues revealed in quantitative evaluations. For ST-R, qualitative results clearly expose its inadequacy in handling distractors. In most cases (e.g., cases 2–6), matching and non-matching targets interleave in the ranked list, making them inherently indistinguishable through a single threshold. Furthermore, cases 6 highlight ST-R’s risk of missing correct matches. In contrast, by integrating the powerful visual discrimination capabilities of the 2D-LVLM, *Open3DSearch* enables the precise identification of all matching targets from similar candidates, showcasing higher robustness.

## 5.3 Parameter Sensitivity Evaluation

To provide theoretical support and guidance for parameter tuning in the practical deployment of *Open3DSearch*, we assess whether its







Query Text	Ranked 3D Shape List $\mathcal{C}$ (Top 8)	Predicted Matches			GT Matches
		3D-LVLM-R	ST-R	Open3DSearch	
white baseball bat.		[1-8]	[1-2]	[None]	[None]
a green and blue pot with a face design, resembling a vase.		[1-8]	[1-4]	[5]	[5]
wooden shield with a red dragon emblem.		[1-8]	[1-8]	[5]	[5]
A vintage black and white camera model on a tripod.		[1-8]	[1-5]	[1, 5]	[1, 5]
a man wearing a white suit.		[1-8]	[1-3]	[3]	[3]
a small glass container with blue liquid on a wooden base.		[1-8]	[None]	[1, 6]	[1, 6]

Figure 4: Qualitative comparison of retrieval performance with baseline methods. 3D-LVLM-R employs MiniGPT-3D[33], and the ST-R threshold is set to 0.164.

Table 2: Quantitative comparison of retrieval performance with baseline methods. \* denotes the pre-defined optimal setting.

Method	Match Absent		Match Present		
	FPR ↓	AFP ↓	Precision ↑	Recall ↑	F1 ↑
<i>Similarity Threshold-based Retrieval (ST-R)</i>					
$\xi = 0.120$	1.000	32.133	0.110	0.932	0.197
$\xi = 0.145$	0.933	7.890	0.303	0.818	0.442
* $\xi = 0.164$	0.687	3.433	0.433	0.609	0.506
$\xi = 0.170$	0.579	2.761	0.443	0.534	0.485
$\xi = 0.195$	0.205	1.475	0.300	0.257	0.277
$\xi = 0.220$	0.041	1.250	0.116	0.091	0.102
<i>3D LVLM-based Retrieval (3D-LVLM-R)</i>					
PointLLM[38]	1.000	7568.554	0.001	1.000	0.001
ShapeLLM[29]	1.000	7820.610	0.000	1.000	0.001
MiniGPT-3D[33]	1.000	4475.667	0.005	1.000	0.011
Ours	0.262	1.706	0.693	0.794	0.740

retrieval performance remains consistent under different parameter configurations. We focus on two key parameters: the HEALPix grid sampling resolution (parameter  $w$ ) and the retrieval termination parameter (parameter  $u$ ).

**5.3.1 The Impact of HEALPix Grid Sampling Resolution.** We set the HEALPix resolution  $w$  to  $\{1, 2, 3\}$ , corresponding to 12, 48, and 108 uniformly distributed camera viewpoints, respectively, and

Table 3: Performance under different parameter settings. \* denotes the default setting.

Param.	Match Absent		Match Present		
	FPR ↓	AFP ↓	Precision ↑	Recall ↑	F1 ↑
<i>HEALPix Grid Sampling Resolution (<math>w</math>)</i>					
* $w = 1$	0.262	1.706	0.693	0.794	0.740
$w = 2$	0.287	1.714	0.666	0.763	0.711
$w = 3$	0.251	1.878	0.703	0.744	0.723
<i>Retrieval Stopping Parameter (<math>u</math>)</i>					
$u = 3$	0.190	1.162	0.730	0.658	0.692
* $u = 5$	0.262	1.706	0.693	0.794	0.740
$u = 7$	0.297	1.793	0.671	0.798	0.729
$u = 9$	0.405	2.316	0.601	0.807	0.689

collected the retrieval performance under each setting (see Table 3). The experimental results show that as  $w$  increases from 1 to 3, the variations in all metrics are within 15%, without any significant trend of performance improvement. This indicates that even at lower sampling densities, the proposed HEALPix-based view rendering strategy consistently captures text-relevant critical information, exhibiting strong robustness. Therefore, considering rendering costs and efficiency requirements in practical applications, setting  $w$  to 1 suffices for meeting retrieval accuracy demands and represents a more pragmatic choice.

**Table 4: Ablation studies on our *Open3DSearch*. VRS and PFM refer to the view rendering strategy and the prompt feedback mechanism, respectively.**

Abl.	Match Absent		Match Present		
	FPR ↓	AFP ↓	Precision ↑	Recall ↑	F1 ↑
w/o VRS	0.169	1.515	0.590	0.540	0.564
w/o PFM	0.133	1.269	0.687	0.649	0.667
w/o VRS & PFM	0.056	1.182	0.504	0.354	0.416
Full	0.262	1.706	0.693	0.794	0.740

**5.3.2 The Impact of Retrieval Stopping Parameter.** To analyze the impact of the retrieval termination parameter  $u$  on system performance, we conduct tests under  $u \in \{3, 5, 7, 9\}$ , with results summarized in Table 3. In scenarios where ground-truth matches are absent, we found that a smaller  $u$  helps reduce the FPR and MFP, exhibiting a linear trend. For scenarios with ground-truth matches, the system achieved the highest F1 score at  $u = 5$ , indicating a good balance between precision and recall. An overly small  $u$  (e.g.,  $u = 3$ ) tends to cause premature termination of the retrieval process, resulting in a significant decrease in recall by missing correct matches. Conversely, setting  $u$  too high (e.g.,  $u = 9$ ) may improve recall but also introduces more false positives, thereby reducing precision. In summary, we believe that the parameter  $u$  should be adjusted flexibly to meet different user interaction requirements.

## 5.4 Ablation Study

We investigate the contributions of key components to the overall pipeline through ablation experiments, including the view rendering strategy and the prompt feedback mechanism. Following the principle of controlled variables, we remove or simplify these components and measure performance changes to analyze their importance.

**View Rendering Strategy.** To verify the effectiveness of the view rendering strategy, we design a simplified version for comparison. In this version, the preferred view  $I^*$  is rendered from a camera pose not constrained by semantics. First, the camera position is randomly sampled from the spherical region determined by the 3D shape center  $c$  and distance  $d_{(o,c)}$ . Second, for camera orientation, only the optical axis (i.e., Z-axis) is constrained to point toward the 3D shape center, while the other two axes are randomly computed using Equation 7. When the preferred view  $I^*$  is insufficient for the 2D-LVLM to make a definitive judgment (i.e., returning '1' or '0'), the above steps are repeated to obtain additional views for auxiliary decision-making.

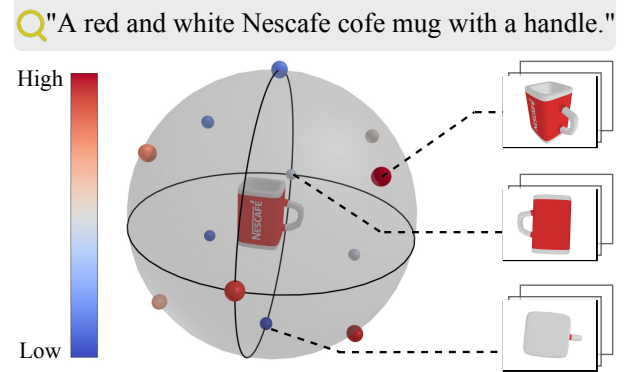
**Prompt Feedback Mechanism.** We assess the contribution of the prompt feedback mechanism to accurate retrieval by removing it. This is achieved by modifying the prompt. Specifically, the ending of the primary prompt (see Figure 3(c)) is changed to: "If it matches, reply "1"; otherwise, reply "0". The pipeline makes a definitive judgment based solely on the preferred view  $I^*$  without performing multi-round interaction.

The experimental results are shown in Table 4. It can be observed that these two components significantly improve performance in terms of precision, recall, and F1-score, especially the view selection

strategy (which caused the F1-score to increase by 31.2%). It is noteworthy that although the FPR experienced adverse effects—rising by 9.3 and 12.9 percentage points, respectively—in the absence of ground-truth matches, the maximum MFP of 1.706 indicates that the risk of mismatches remains within an acceptable range.

## 5.5 Visualization Analysis

The effectiveness of *Open3DSearch* hinges on the determination of rendered views, which directly impacts the accuracy of matching judgments performed by the 2D-LVLM. To better understand this process, we present an intuitive case study. As shown in Figure 5, we visualize the candidate camera positions  $\{o\}_{k=1}^K$  on the constrained spherical region, along with some rendered views. The color of each position depends on its average semantic similarity  $\{sim_{avg}(I_k, q)\}_{k=1}^K$ . It can be observed that the positional constraints significantly enhance the 3D shape details in the rendered views (e.g., the clearly visible "Nescafe" text). Additionally, rendered views from the highest similarity position tend to convey the semantic information described in the query more completely and clearly. This demonstrates that selecting the primary view from this position in *Open3DSearch* effectively reduces matching ambiguity, thereby improving the judgment accuracy of the 2D-LVLM.



**Figure 5: A visualization case of the proposed view rendering strategy. The points denote candidate camera positions, and their colors indicate the corresponding average semantic similarity (computed by Equation 8).**

## 6 CONCLUSION

This paper proposes the *Open3DSearch* to address the demand for open-domain text-to-shape precise retrieval, pioneering the integration of LVLMs into 3D cross-modal retrieval tasks. By transforming implicit semantic matching on 3D shapes into explicit visual-language reasoning on 2D images, our approach overcomes the inherent limitations of traditional embedding-based methods in zero-shot generalization and decision ambiguity. Supported by the constructed Uni3D-R benchmark dataset, *Open3DSearch* demonstrates superior performance across diverse matching scenarios through comprehensive experiments.



## REFERENCES

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [2] Eman Ahmed, Alexandre Saint, Abd El Rahman Shabayek, Kseniya Cherenkova, Rig Das, Gleb Gusev, Djamilia Aouada, and Bjorn Ottersten. 2018. A survey on deep learning advances on different 3D data representations. *arXiv preprint arXiv:1808.01462* (2018).
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.
- [4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966* 1, 2 (2023), 3.
- [5] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qishui Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954* (2024).
- [6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).
- [7] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. 2019. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 100–116.
- [8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500* (2023).
- [9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13142–13153.
- [10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [11] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793* (2024).
- [12] Krzysztof M Gorski, Eric Hivon, Anthony J Banday, Benjamin D Wandelt, Frode K Hansen, Mstvos Reinecke, and Matthias Bartelmann. 2005. HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal* 622, 2 (2005), 759.
- [13] Zhizhong Han, Mingyang Shang, Xiyang Wang, Yu-Shen Liu, and Matthias Zwicker. 2019. Y2Seq2Seq: Cross-modal representation learning for 3D shape and text by joint reconstruction and prediction of view and word sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 126–133.
- [14] Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal Patel. 2023. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2028–2038.
- [15] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems* 36 (2023), 20482–20494.
- [16] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. 2023. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22157–22167.
- [17] Anastasia Ioannidou, Elisavet Chatzilaria, Spiros Nikolopoulos, and Ioannis Kompatsiaris. 2017. Deep learning advances in computer vision with 3d data: A survey. *ACM computing surveys (CSUR)* 50, 2 (2017), 1–38.
- [18] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.
- [19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.
- [20] Trung-Nghia Le, Tam V Nguyen, Minh-Quan Le, Trong-Thuan Nguyen, Viet-Tham Huynh, Trong-Le Do, Khanh-Duy Le, Mai-Khiem Tran, Nhat Hoang-Xuan, Thang-Long Nguyen-Ho, et al. 2023. TextANIMAR: text-based 3D animal fine-grained retrieval. *Computers & Graphics* 116 (2023), 162–172.
- [21] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326* (2024).
- [22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
- [23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26296–26306.
- [24] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253* (2024).
- [25] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. 2024. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems* 36 (2024).
- [26] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525* (2024).
- [27] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. 2023. Scalable 3d captioning with pretrained models. *Advances in Neural Information Processing Systems* 36 (2023), 75307–75337.
- [28] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2023. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning*. PMLR, 28223–28243.
- [29] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. 2024. Shapellm: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision*. Springer, 214–238.
- [30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [31] Yue Ruan, Han-Hung Lee, Yiming Zhang, Ke Zhang, and Angel X Chang. 2024. TriCoLo: Trimodal contrastive loss for text to shape retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5815–5825.
- [32] Chuan Tang, Xi Yang, Bojian Wu, Zhizhong Han, and Yi Chang. 2023. Parts2words: Learning joint embedding of point clouds and texts by bidirectional matching between parts and words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6884–6893.
- [33] Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Yixue Hao, Long Hu, and Min Chen. 2024. Minigpt-3d: Efficiently aligning 3d point clouds with large language models using 2d priors. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 6617–6626.
- [34] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- [36] Emo Welzl. 2005. Smallest enclosing disks (balls and ellipsoids). In *New Results and New Trends in Computer Science: Graz, Austria, June 20–21, 1991 Proceedings*. Springer, 359–370.
- [37] Hao Wu, Ruochong Li, Hao Wang, and Hui Xiong. 2024. COM3D: Leveraging Cross-View Correspondence and Cross-Modal Mining for 3D Retrieval. *arXiv preprint arXiv:2405.04103* (2024).
- [38] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2024. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*. Springer, 131–147.
- [39] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. 2023. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1179–1189.
- [40] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. 2024. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27091–27101.
- [41] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305* (2023).
- [42] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115* (2024).



- [43] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421* 9, 1 (2023), 1.
- [44] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800* (2024).
- [45] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. 2022. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8552–8562.