Beyond Categorical Matching: Intra-Class Graded Relevance Estimation for Cross-Modal 3D Retrieval

Shunning Liu[1,2*], YingPeng Zhang[2*,†], Jianing Lin[3], Yifan Wang[1], Yang Zhang[1], and Chun Yuan[1†]

[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen, Guangdong, China
[2] Tencent IEG, Shenzhen, Guangdong, China
[3] School of Computer Science and Engineering, Beihang University, Beijing, China

Abstract. Mainstream 3D asset retrieval relies predominantly on category-level matching, overlooking intra-class variations in geometry, style, and texture critical for real-world applications. To address this, we redefine retrieval from binary categorical matching to intra-class graded relevance estimation, treating relevance as a continuous variable to capture fine-grained semantic nuances. We propose a tri-modal encoder incorporating a Semantic Conditioned Interaction module for cross-modal fusion and a Query-Guided Mixture-of-Experts module for dynamic modality weighting. This design explicitly decouples intra-class variations and prevents semantic collapse via a re-parameterization mechanism. Furthermore, we introduce an automated, scalable benchmark generation framework driven by a Multimodal Large Language Model to synthesize high-quality soft relevance labels, proposing a new benchmark: INGRE. Extensive experiments demonstrate our method achieves state-of-the-art results on INGRE while maintaining superior performance and robustness on standard classification tasks.

Keywords: 3D Retrieval · Multimodal Embedding · Intra-Class Graded Relevance Estimation

1 Introduction

3D asset retrieval plays a pivotal role in augmented reality, virtual reality, digital content creation, and 3D e-commerce [1, 38, 40, 57], where efficient asset discovery directly dictates user immersion and pipeline productivity. To address large-scale cross-modal retrieval, mainstream models predominantly adopt representation learning, encoding multi-modal assets into a shared feature space and employing distance metrics for retrieval [1–6, 35, 38, 40, 45, 47]. While this

* Equal contribution. † Corresponding authors. Contact the corresponding author through yuanc@sz.tsinghua.edu.cn. This work is supported by the SSTIC Grant (KJZD20230923115106012, KJZD20230923114916032, GJHZ20240218113604008). Get code at https://github.com/Listeningx/objaverse-benchmark-generation.

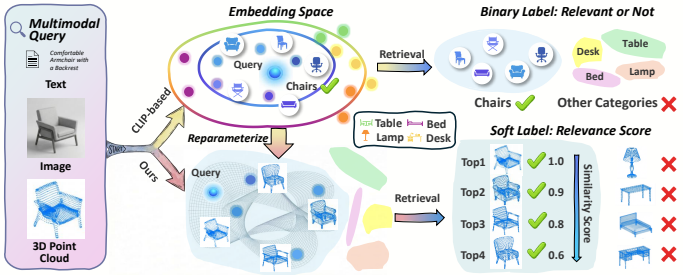


Fig. 1: Paradigm Shift from Binary Categorical Matching to Intra-Class Graded Relevance Estimation. Based on multimodal user queries, we encode them to embeddings for retrieval. While CLIP-based methods rely on binary categorical matching, our approach models continuous intra-class relevance to capture fine-grained variations, enabling precise ranking.

paradigm drives benchmark performance, it overlooks a critical characteristic of real-world scenarios: assets within the same semantic category exhibit significant intra-class variance in geometry, style, texture, and functionality. Users typically seek results ranked by fine-grained similarity rather than mere category-level matching. This misalignment between task formulation and practical needs imposes an intrinsic ceiling on performance.

To surmount this limitation, we transition the 3D asset retrieval paradigm from categorical matching to **intra-class graded relevance estimation**. Existing approaches typically define relevance as a binary attribute at the category level, where an asset is deemed relevant if it shares the semantic class with the query [5,7–9,37]. This discrete setting implicitly assumes intra-class homogeneity. As illustrated in Figure 1, under our proposed perspective, relevance is reformulated as a continuous variable designed to reflect structured relationships within semantic categories. The objective extends beyond inter-class discrimination to intra-class ranking, requiring the embedding space to distinguish categories while preserving subtle feature nuances.

Guided by this definition, we propose ReMU3D, a Reparameterized Multimodal Unified representation encoder for Intra-class Graded Relevance 3D Asset Retrieval. As attributes like style, material, and context are often ambiguous in raw 3D geometry, while image and text offer rich details, effective intra-class modeling requires structured cross-modal fusion. Furthermore, practical experiments[1] show that simply averaging embeddings from two modalities outperforms single-modality retrieval, motivating our multimodal joint-encoding adoption. Specifically, we introduce a Semantic Conditioned Interaction (SCI) module using image and text features to condition point cloud tokens via attention, facilitating deep integration. Subsequently, a Query-Guided Mixture-of-Experts (Q-MoE) module identifies input modality combinations via gating,

[1] Detailed experimental settings and results are provided in the supplementary material.

dynamically adjusting expert weights. Unlike prior methods relying on simple alignment [3–5, 10–13, 47], our framework supports mixed-modality training. By re-parameterizing the pre-aligned semantic space, we explicitly decouple intra-class variations and enhance discriminative power.

To rigorously evaluate graded intra-class relevance, we construct an automated, scalable, and transferable benchmark generation framework. Specifically, we engineer a refinement system driven by a Multimodal Large Language Model, which synthesizes multi-dimensional attributes of 3D assets to generate soft labels that align closely with human preferences. Furthermore, we adapt traditional retrieval metrics to accommodate intra-class graded relevance estimation. Using this pipeline, we build a new benchmark named INGRE. Extensive experiments demonstrate that our method not only excels on standard inter-class 3D classification benchmarks [1, 14, 15] but also achieves state-of-the-art results on our proposed INGRE. The framework exhibits superior precision and robustness compared to existing methods.

In summary, the main contributions of this work are as follows:

- We redefine 3D asset retrieval from binary classification to intra-class graded relevance estimation, emphasizing fine-grained semantic differences for high-quality retrieval.
- We propose a novel tri-modal encoder utilizing reparameterization to prevent intra-class semantic collapse while ensuring robustness against missing modalities.
- We introduce the first open-source framework for graded relevance benchmarks; employing automated annotation, we constructed INGRE to fill the gap in refined evaluation standards.

2 Related Work

2.1 Single-Modal 3D Asset Retrieval

Single-modality 3D asset retrieval has progressed from handcrafted descriptors to deep geometric encoders. Early methods focused on invariant shape signatures for rigid matching: Shape Distributions [7] and spherical-harmonic representations [8] modeled global geometry statistics, while the Light Field Descriptor [37] leveraged multi-view projections. Although effective on clean CAD data, these descriptors are sensitive to deformation, topology noise, and intra-class appearance variation [52]. Deep learning shifted the paradigm to data-driven representation learning. View-based pipelines such as MVCNN [16], GVCNN [17], and RotationNet [18] improved discriminability by aggregating rendered views. Voxel and point-based models then became dominant for native 3D input, including volumetric learning [14], PointNet/PointNet++ [19, 20], and dynamic-graph modeling [21]. More recently, Transformer pretraining on point clouds (e.g., Point-BERT [56] and [50]) further improved generalization and robustness [51, 56].

2.2 Multimodal Representation Learning for Cross-Modal 3D Retrieval

Despite strong geometric retrieval performance, single-modality methods are inherently limited when user intent depends on textual semantics or visual style cues, motivating cross-modal retrieval frameworks [22, 38, 43]. CLIP [10] established large-scale image-text alignment and inspired 3D transfer. Early approaches such as PointCLIP and CLIP2Point [5, 12] projected point clouds into depth-like images and reused frozen vision-language encoders. Subsequent methods moved to joint multimodal optimization with explicit 3D supervision. ULIP/ULIP-2 [3, 13], OpenShape [47], and Uni3D [4] scale triplet-style alignment with large image-text-3D corpora, improving open-vocabulary retrieval and transfer across datasets. Recent work further enhances cross-modal geometry fusion [23, 24, 58]. In parallel, multimodal retrieval is being extended to compositional and scene-aware settings [25, 49]; rendering-centric retrieval alignment with Gaussian representations is also actively explored [55]. For open-set 3D object retrieval, recent CLIP-enhancement strategies include hypergraph-based multimodal representation learning [39], DAC [54] and CLIP-AdaM [42]. Overall, multimodal representation learning has shifted 3D retrieval toward semantically rich, query-flexible retrieval in open-world settings.

2.3 Evaluation for 3D Asset Retrieval

Benchmarks drive algorithmic iteration. The Princeton Shape Benchmark (PSB) [26] laid the foundation, while ModelNet [14] and ShapeNet [27] dominate the deep learning era. However, synthetic datasets lack real-world complexity. ScanObjectNN [15] addresses this with real-world scans containing noise and occlusions, challenging model robustness. Meeting large model data needs, Objaverse [1] offers over 800,000 high-quality models, becoming a standard for next-generation tasks. OmniObject3D [28] provides vast real-world scans, mitigating the synthetic-real domain gap. Xu et al. [36] introduced TexVerse, a large-scale dataset with detailed textures for texture-aware tasks. PartNet [29] and 3D-Future [40] are only manually annotated resources for part-level geometry and furniture attributes rather than query-conditioned graded relevance. Future research requires challenging fine-grained standards to reflect practical performance discrepancies [9, 30].

3 Method

3.1 Problem Definition

Let $\mathcal{X}$ be a 3D asset set and q be a query that may contain any subset of modalities (text, image, point cloud). Conventional retrieval defines binary relevance $y(q, x) \in \{0, 1\}$, which treats all in-class assets as equivalent. This binarization compresses the underlying continuous similarity manifold into a coarse supervision signal and overlooks the model's ability to capture fine-grained similarity

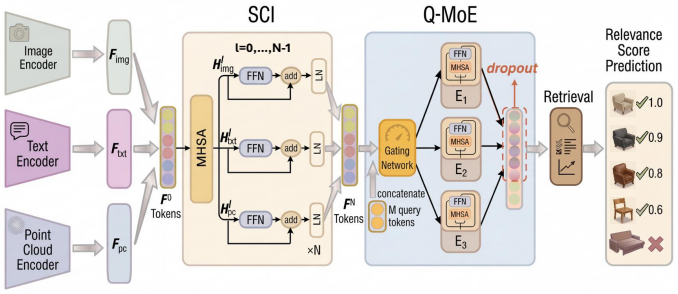


Fig. 2: Overall architecture of ReMU3D and the intra-class graded-relevance retrieval pipeline.

gradients; *e.g.*, two chairs can share the same label but differ substantially in style or intended usage.

We redefine retrieval as *intra-class graded relevance estimation* by predicting a continuous relevance score $s(q, x) \in [0, +\infty)$ for each candidate $x \in \mathcal{X}$. Given a query q, the goal is to learn an embedding function $f(\cdot)$ such that the induced ranking π_q over candidates maximizes agreement with the graded relevance ordering defined by $s(q, x)$. Formally, for any $x_i, x_j \in \mathcal{X}$, $s(q, x_i) > s(q, x_j)$ should imply $\langle f(q), f(x_i)\rangle > \langle f(q), f(x_j)\rangle$.

We retain a standard cross-modal contrastive learning objective because the contrastive mechanism itself encourages the model to capture differences among assets within the same class. Although each batch only provides binary positive and negative pairs, with sufficiently large training data and batch sizes the model naturally distinguishes "hard negatives" from "easy negatives." Coupled with our architecture, this is sufficient to achieve the desired graded relevance behavior. Our training objective is as follows:

$$\mathcal{L} = -\frac{1}{4N}\sum_{i=1}^{N}\left(\log\frac{\exp(\mathbf{e}_i^A\cdot\mathbf{e}_i^T/\tau)}{\sum_j \exp(\mathbf{e}_i^A\cdot\mathbf{e}_j^T/\tau)} + \log\frac{\exp(\mathbf{e}_i^T\cdot\mathbf{e}_i^A/\tau)}{\sum_j \exp(\mathbf{e}_i^T\cdot\mathbf{e}_j^A/\tau)} + \log\frac{\exp(\mathbf{e}_i^A\cdot\mathbf{e}_i^I/\tau)}{\sum_j \exp(\mathbf{e}_i^A\cdot\mathbf{e}_j^I/\tau)} + \log\frac{\exp(\mathbf{e}_i^I\cdot\mathbf{e}_i^A/\tau)}{\sum_j \exp(\mathbf{e}_i^I\cdot\mathbf{e}_j^A/\tau)}\right). \quad (1)$$

Here, A_i, I_i, and T_i denote the asset (3D), image, and text inputs in the i-th triplet; f_A, f_I, and f_T are the corresponding encoders; and $\mathbf{e}_i^A = \frac{f_A(A_i)}{\|f_A(A_i)\|}$, $\mathbf{e}_i^I = \frac{f_I(I_i)}{\|f_I(I_i)\|}$, $\mathbf{e}_i^T = \frac{f_T(T_i)}{\|f_T(T_i)\|}$ are their ℓ_2-normalized embeddings.

3.2 Overview of ReMU3D

Figure 2 summarizes the proposed ReMU3D and its execution flow. The SCI module re-parameterizes the original CLIP space and deeply fuses the three modalities to produce detail-preserving embeddings, while Q-MoE supports arbitrary modality combinations and improves robustness via modality-adaptive aggregation. With these two modules, our ReMU3D can address the graded relevance estimation objective in Equation 1.

3.3 Semantic Conditioned Interaction (SCI)

To effectively model fine-grained intra-class variations, we design a **Semantic Conditioned Interaction (SCI)** module that enables structured cross-modal integration among image, text, and point cloud representations. The goal of SCI is not simple semantic alignment, but leveraging high-level appearance and linguistic cues to guide geometric representation learning.

Given an asset, we extract image features $\mathbf{F}_{img} \in \mathbb{R}^{L_i \times d}$, text features $\mathbf{F}_{txt} \in \mathbb{R}^{L_t \times d}$, and point cloud tokens $\mathbf{F}_{pc} \in \mathbb{R}^{L_p \times d}$, where L_i, L_t, and L_p denote the token lengths of each modality and d is the shared embedding dimension. In our work, we choose L_i and L_t as 1. We concatenate the three modalities along the token dimension to form a joint sequence:

$$\mathbf{F}^0 = [\mathbf{F}_{img}; \mathbf{F}_{txt}; \mathbf{F}_{pc}] \in \mathbb{R}^{(L_i+L_t+L_p)\times d}. \tag{2}$$

We assume simply aligning 3D representations with the pretrained CLIP space is insufficient to capture all attributes, as the 3D modality conveys far richer information than image and text, particularly in texture, structure, and functionality. Direct alignment can therefore cause modality information collapse [44, 46, 53]. Specifically, same-class assets with subtle differences remain indistinguishable in the pretrained CLIP space, mapping to the same point. To avoid this, we defer alignment to the unified multimodal semantic space after SCI encoding. Modality-specific encoders project outputs into a unified feature space, feeding the concatenated sequence into a Transformer with N stacked self-attention blocks. Each block applies global Multi-Head Self-Attention (MHSA) over concatenated tokens, then splits by modality for FFNs with residual connections and layer normalization:

$$\begin{aligned}
\mathbf{H}^l &= \mathrm{MHSA}(\mathbf{F}^l), \\
\mathbf{H}^l &\to \{\mathbf{H}^l_{img}, \mathbf{H}^l_{txt}, \mathbf{H}^l_{pc}\}, \\
\mathbf{F}^{l+1}_m &= \mathrm{LN}\big(\mathbf{H}^l_m + \mathrm{FFN}(\mathbf{H}^l_m)\big), \quad m \in \{img, txt, pc\}, \\
\mathbf{F}^{l+1} &= \mathrm{Concat}\big(\mathbf{F}^{l+1}_{img}, \mathbf{F}^{l+1}_{txt}, \mathbf{F}^{l+1}_{pc}\big), \quad l = 0, \ldots, N-1.
\end{aligned} \tag{3}$$

Unlike asymmetric cross-attention designs that explicitly separate query and key roles, SCI adopts a unified self-attention mechanism over the full multimodal sequence. This formulation allows every token to attend to all others, enabling global interaction across modalities. Within this shared space, image and text tokens primarily function as semantic conditions. Image features inject fine-grained appearance cues such as style, material, and color, while text features provide high-level semantic descriptors and contextual constraints. Through self-attention, these semantic tokens modulate the attention distribution over point cloud tokens, guiding geometric representations to emphasize semantically relevant attributes. Consequently, cross-modal interaction acts as a semantic refinement process rather than simple feature aggregation.

Meanwhile, point cloud tokens undergo deep intra-modal fusion across layers, progressively integrating structural information with injected semantic cues.

After N layers of interaction, the updated point cloud tokens encode a holistic representation that jointly captures geometry, appearance, and semantic context in a unified embedding space. This design preserves subtle intra-class differences, ultimately producing semantically enriched geometric embeddings suitable for intra-class graded retrieval.

Additionally, this structure naturally accommodates missing-modality settings: when any one or several modalities are absent, the SCI module can still perform fusion over the available tokens without changing its formulation.

3.4 Query-Guided Mixture-of-Experts (Q-MoE)

After the SCI module performs deep cross-modal interaction, inspired by [41], we further introduce a **Query-Guided Mixture-of-Experts (Q-MoE)** module to achieve modality-adaptive feature aggregation and robust embedding generation.

Specifically, given the fused token sequence $\mathbf{F}_{SCI}$ produced by SCI, we append a fixed number M of learnable query tokens $\mathbf{Q} \in \mathbb{R}^{M\times d}$ to the sequence:

$$\mathbf{F}_{MoE} = [\mathbf{F}_{SCI}; \mathbf{Q}], \tag{4}$$

where $\mathbf{Q}$ is randomly initialized and optimized jointly with the entire network. These query tokens act as task-oriented aggregators, progressively extracting information relevant to retrieval from the fused multimodal features.

The augmented sequence is then processed by a Mixture-of-Experts (MoE) layer. Concretely, we maintain a set of K expert networks $\{\mathcal{E}_k\}_{k=1}^{K}$, each implemented as a lightweight feed-forward transformation. A router function $\mathcal{R}(\cdot)$ analyzes the global feature context and infers the active modality set of the current input. Based on this inference, it produces routing weights $\mathbf{w} \in \mathbb{R}^K$ satisfying $\sum_{k=1}^{K} w_k = 1$, which are used to combine expert outputs:

$$\mathbf{H} = \sum_{k=1}^{K} w_k \, \mathcal{E}_k(\mathbf{F}_{MoE}). \tag{5}$$

Through this dynamic routing mechanism, the model adaptively adjusts expert contributions according to the modality composition and query characteristics, enabling flexible handling of full-modality and missing-modality scenarios within a unified framework.

After expert aggregation, we extract the updated representations corresponding to the M query tokens and apply a pooling operation (e.g., mean pooling) to obtain the final embedding. Then we apply ℓ_2 normalization to produce a unit-length representation, which is used for cosine-similarity-based retrieval:

$$f(q) = \mathrm{Pool}(\mathbf{H}_{query}), \hat{f}(q) = \frac{f(q)}{\|f(q)\|_2}, \tag{6}$$

where $\mathbf{H}_{query} \in \mathbb{R}^{M\times d}$ denotes the output features of the query tokens.

By combining query-guided aggregation with modality-aware expert routing, Q-MoE enables adaptive feature integration while maintaining a compact and stable embedding space. This design enhances robustness in fine-grained intra-class graded relevance ranking tasks while missing modalities.

4 Benchmark Construction

To make evaluation consistent with the proposed paradigm and model design, we construct a scalable pipeline for generating intra-class graded relevance labels. The framework is driven by a Multimodal Large Language Model (MLLM) and is designed to produce preference-aligned soft labels without manual annotation. In addition, the pipeline is modular and transferable across datasets. More implementation details of benchmark construction are provided in the supplementary material.

4.1 Soft Label Generation Strategies

Considering the diversity of 3D datasets, we use two complementary label generation strategies: one for geometry-dominant assets (often without RGB) and one for large-scale complex assets.

Strategy I: Structured Case-Based Grading for Geometry-Dominant Assets. For clean geometry with limited appearance cues, we build category-wise query cases and score candidates with a fixed hierarchy of dimensions (e.g., color, function, and texture). The MLLM outputs structured graded relevance scores, and we add inter-class distractors to preserve retrieval difficulty.

Strategy II: Adaptive Multi-Step Grading for Large-Scale Complex Assets. For large-scale and heterogeneous datasets such as Objaverse [1], we introduce a multi-step adaptive scoring framework. Due to the scale and varying data quality, we first curate a set of 10,742 high-quality gold-standard assets to ensure reliable evaluation. First, given a query object, the MLLM analyzes its intrinsic characteristics and infers the most appropriate evaluation focus, including relevant comparison dimensions and potential application scenarios. From a predefined dimension bank, it selects the five most critical dimensions and assigns a weight to each, forming a query-specific rubric. Next, the MLLM generates detailed multi-dimensional descriptions for all candidate assets according to the selected dimensions. Together with the previously determined dimension weights, the model computes dimension-wise scores and aggregates them into a final composite relevance score. A consistency verification step is then performed to ensure that the generated ranking is logically coherent and semantically reasonable.

4.2 Overview of Benchmark: INGRE

Through the above two strategies, we construct a comprehensive benchmark dedicated to evaluating intra-class graded relevance in 3D asset retrieval, INGRE. As summarized in Fig. 3, INGRE covers diverse datasets with unified

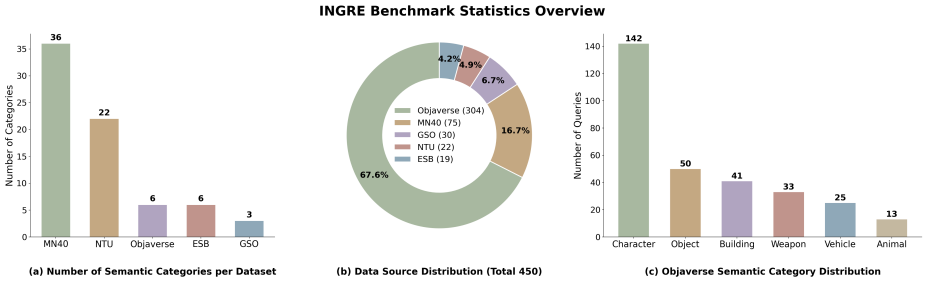


Fig. 3: Overview of INGRE, including (a) category distribution over all source datasets, (b) source datasets and their proportion, and (c) category distribution in Objaverse subset.

graded-relevance annotation. It integrates data from ESB [9], NTU [15], ModelNet40 [14], GSO [31], and Objaverse [1], covering both geometry-only and RGB-enhanced assets. In total, INGRE spans over 70 semantic categories, including characters, weapons, animals, clothing and accessories, industrial parts, and other common asset types encountered in practical retrieval applications. All annotations are generated through our automated MLLM-driven pipeline, ensuring scalability and transferability. We fully open-source the benchmark and its construction framework, facilitating future research.

5 Experiments

5.1 Experimental Setup

Evaluation Metrics Standard retrieval metrics include Recall, mAP, NDCG, and ANMRR, with the latter three considering ranking positions. However, their typical binary relevance (0 or 1) ignores similarity variations within semantic categories. To address this, we replace binary labels with continuous relevance from our annotation pipeline and calculate NDCG. Specifically, for a ranked list of length K, we use:

$$\mathrm{DCG@}K = \sum_{i=1}^{K} \frac{2^{\mathrm{rel}_i} - 1}{\log_2(i+1)}, \quad \mathrm{NDCG@}K = \frac{\mathrm{DCG@}K}{\mathrm{IDCG@}K}, \tag{7}$$

where rel_i is the graded relevance of the item at rank i, and $\mathrm{IDCG@}K$ is the maximum possible DCG obtained by an ideal ranking. We primarily report NDCG@5 and NDCG@10, since real-world retrieval applications focus more on a small number of top-ranked 3D assets. In addition, we introduce Kendall's Tau [32] and Spearman's Rho [33] to further evaluate the quality of relative ordering among objects within the same semantic class. We augment ground-truth lists with zero-relevance objects from different categories. Consequently, our protocol simultaneously measures intra-class ranking quality and inter-class separability.

Table 1: Main results on INGRE. Higher is better for all metrics. The best results are **bolded**, and the second-best results are underlined. To clearly highlight the advantages of our model, we report the best-performing version of each baseline method.

Method	NDCG@5	NDCG@10	NDCG@20	NDCG@50	Kendall's τ	Spearman's ρ
ULIP [3]	0.69	0.71	0.72	0.88	0.22	0.30
OpenShape [47]	0.76	0.76	0.79	0.88	0.30	0.41
Uni3D [4]	0.71	0.73	0.75	0.88	0.32	0.43
ReMU3D	**0.94**	**0.93**	**0.92**	**0.91**	**0.39**	**0.52**

Model and Implementation Details Our ReMU3D builds upon Uni3D [4] to support arbitrary combinations of image, point cloud, and text modalities. The SCI module has 4 Transformer layers (16 heads, RoPE, MLP ratio 4.0), and Q-MoE uses 3 experts (8 blocks each, 16 heads) plus 30 learnable query tokens per modality combination. Fused features are mean-pooled and projected to the CLIP embedding space with a two-layer MLP.

During training, we train with precomputed CLIP embeddings for efficiency and optimize Eq. 1. The temperature is initialized as $\ln(1/0.07) \approx 2.66$. Benefiting from our SCI and Q-MoE modules, we dynamically adjust the proportion of different combinations based on the model's performance on the validation set. Meanwhile, combinations that include point clouds are assigned higher sampling ratio. AdamW, cosine decay, and 500-step warmup are used. The base learning rate is 1×10^{-4}, with weight decay 0.1 and layer-wise decay 0.95 for the point encoder. We use batch size 36 per GPU, 8-step accumulation on 8 NVIDIA RTX H20 GPUs (effective batch size 2,304), and DeepSpeed ZeRO-2 [34] with checkpointing. ReMU3D contains 589M trainable parameters. By precomputing frozen features, the online encoding cost is reduced to 167–210 GFLOPs (0.03 s per query), while substantially improving retrieval quality (e.g., NDCG@10 from 0.76 to 0.93). Please see more details in supplementary materials.

5.2 Intra-Class Graded Relevance Estimation

As discussed previously, we reformulate the 3D asset retrieval task as an intra-class graded relevance estimation problem and construct a rigorous benchmark, INGRE. We conduct comprehensive experiments on INGRE and evaluate performance using our enhanced ranking-oriented metrics.

Quantitative Comparisons As shown in Table 1, ULIP [3], OpenShape [47] and Uni3D [4] reflect the current frontier of 3D embedding methods, providing a strong basis for comparison. On INGRE, ReMU3D consistently outperforms SOTA. Notably, it achieves 0.94 on NDCG@5 and 0.93 on NDCG@10, substantially surpassing the runner-up OpenShape (0.76). These margins highlight its capability to distinguish fine-grained semantics. In addition, ReMU3D also excels in global ranking, achieving Kendall's τ of 0.39 and Spearman's ρ of 0.52, surpassing Uni3D by 0.07 and 0.09. Moreover, it reaches 0.91 on NDCG@50 while

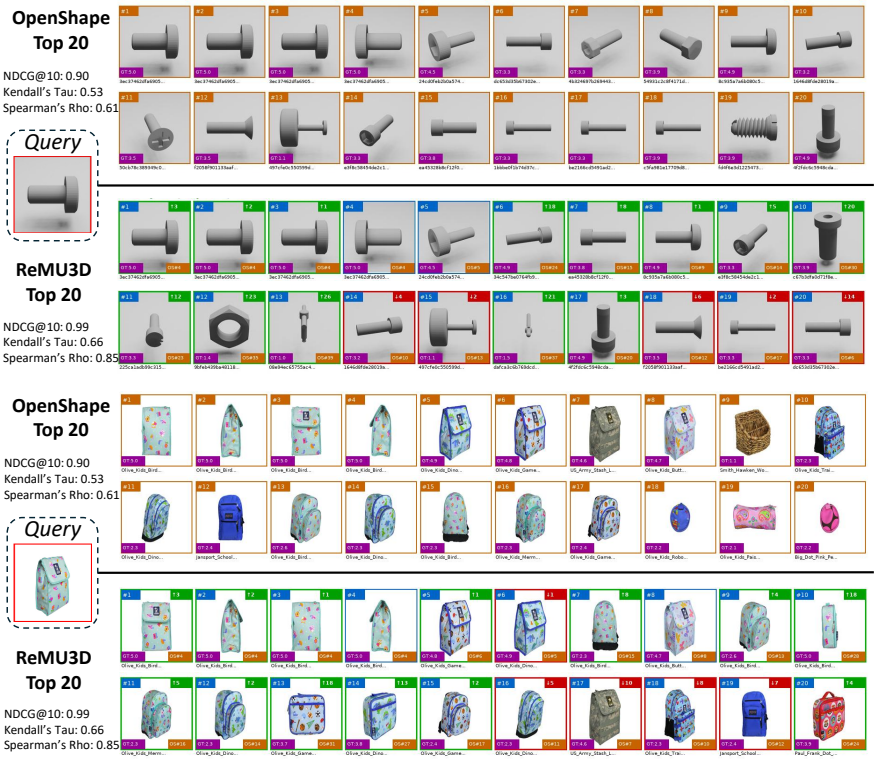


Fig. 4: Qualitative comparison cases on ESB and GSO datasets.

ULIP and OpenShape plateau at 0.88, validating its effectiveness in broader search scopes.

Qualitative Comparisons Beyond quantitative results, we choose OpenShape [47] as the primary baseline for qualitative comparisons. To provide a concrete evaluation, we select two representative cases for comparative analysis. The first case involves an industrial component lacking color information, where retrieval relies strictly on geometry. Compared to OpenShape, ReMU3D demonstrates superior structural discrimination by successfully suppressing the rankings of candidates at positions 14, 15, and 18, which exhibit significant morphological discrepancies from the query. In the second scenario focusing on daily necessities, ReMU3D effectively balances semantic and visual attributes; it assigns lower rankings to bags with distinct color mismatches while successfully recalling a correct bag instance at rank 20. Conversely, OpenShape exhibits semantic drift in the lower rankings, retrieving pencil cases instead of the queried large bags at ranks 18, 19,

Table 2: Zero-shot classification on Objaverse-LVIS [1], ModelNet40 [14] and ScanObjectNN [15]. Best results are **bold**, second-best are underlined. *ReMU3D supports arbitrary combinations of modalities, so we report the best results among all combinations.

Method	Input Mod.	Objaverse-LVIS [1]			ModelNet40 [14]			ScanObjectNN [15]		
		Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
PointCLIP v2 [5]	P	4.7	9.5	12.9	63.6	77.9	85.0	7.3	12.0	14.8
CLIP2Point [12]	P	2.7	5.8	7.9	49.5	71.3	81.2	7.5	14.2	18.2
ReCon [23]	P	1.1	2.7	3.7	61.2	73.9	78.1	10.0	16.8	19.6
ULIP [3]	P	6.2	13.6	17.9	60.4	79.0	84.4	7.6	15.1	19.7
OpenShape [47]	P	39.1	60.8	68.9	85.3	96.2	97.4	47.2	72.4	84.7
ULIP-2 [13]	P	46.3	–	75.0	84.0	–	97.2	–	–	
Uni3D [4]	P	47.2	68.8	76.1	86.8	97.3	98.4	66.5	**83.5**	90.1
PointCLIP v2 [5]	I	24.9	40.9	47.6	50.7	69.1	74.6	–	–	–
OpenView [58]	P + I	52.7	74.6	81.1	87.5	97.0	98.5	–	–	–
ReMU3D*	Agnostic	**53.2**	**76.2**	**83.5**	**89.3**	**98.6**	**98.9**	**68.1**	81.7	**90.5**

and 20. These differences ultimately lead to improvements across all evaluation metrics for ReMU3D[2].

5.3 Zero-Shot 3D Asset Classification

To demonstrate the generalization ability of our method, and to show that improving intra-class fine-grained ranking does not sacrifice general recognition performance, we evaluate ReMU3D on the zero-shot shape classification task. Experiments are conducted on three benchmark datasets: Objaverse-LVIS [1], ModelNet40 [14], and ScanObjectNN [15]. ModelNet40 and ScanObjectNN are widely used datasets containing 40 and 15 common categories, respectively, while Objaverse-LVIS is a labeled and curated subset of Objaverse with 46,832 shapes across 1,156 LVIS categories [1]. We compare ReMU3D against prior state-of-the-art methods in the zero-shot shape classification task, as shown in Table 2. According to the corrected results in Table 2, ReMU3D remains competitive across all datasets. These results indicate that our modality-agnostic design improves recognition robustness overall, especially in Top-1 accuracy.

5.4 Ablation Study

To validate the design choices in ReMU3D, we conduct controlled ablations on module composition, training strategy, expert count, and pooling operator. Quantitative results are summarized in Table 3, and visual trends for pooling and expert count are shown in Figure 5.

Modules of ReMU3D Table 3 shows that removing either core component causes a substantial performance drop. The full model reaches 0.94/0.93 on

[2] We further compare our method with other SOTA models, with detailed qualitative results provided in the supplementary materials.

NDCG@5/10, while removing SCI reduces them to 0.57/0.61; removing Q-MoE yields 0.66/0.70. The same trend appears for rank-correlation metrics (τ, ρ), confirming that SCI and Q-MoE are both essential for preserving intra-class ordering.

Training Strategies We compare our joint training scheme against two alternatives: two-stage training and fixed-ratio optimization. Two-stage means the training process is divided into two distinct phases: initially training the SCI with a frozen Q-MoE, and subsequently optimizing the Q-MoE while keeping the SCI parameters fixed. In the fixed-ratio setting, we utilize predefined, constant ratios for modality combination, rather than optimizing them dynamically during training. As reported in Table 3, two-stage training performs worst, indicating that training them independently disrupts the continuity of the embedding space and consequently degrades the consistency of the learned representations. Fixed-ratio training is stronger but still clearly inferior to the full strategy, showing that adaptive sampling strategy is important for robust and diverse alignment.

Number of Experts and Pooling Strategy Figure a studies the impact of expert count in Q-MoE. Performance improves from a small number of experts to a moderate scale, then saturates when too many experts are used. This trade-off indicates that a moderate expert count offers the best balance between model capacity and optimization stability. Figure b compares different pooling operators under the same model and training setup. The selected pooling strategy consistently provides stronger ranking metrics.

Effect of Soft-label Supervision. Although INGRE provides graded relevance scores, ReMU3D is intentionally trained without directly regressing these labels in order to evaluate its zero-shot graded-relevance retrieval capability. To investigate whether explicitly fitting the soft labels is beneficial, we additionally incorporate two common soft-label objectives, namely MSE loss and KL divergence, on top of the original InfoNCE objective. As shown in Table 5, neither variant yields further improvement over the original training objective, with both even exhibiting slight performance degradation. These results suggest that the performance gains of ReMU3D do not arise from directly fitting the MLLM-generated relevance scores, but rather from learning discriminative representations through contrastive supervision.

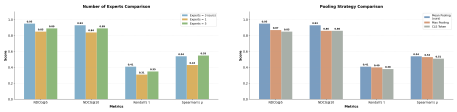
5.5 Visualization Analysis

Cross-modal Retrieval Visualization Experimental results on INGRE-Objaverse, evaluated via NDCG@5 and NDCG@10 in Figure 6, highlight modality alignment dominance, where diagonal elements yield the highest scores. Notably, V-to-V[3] retrieval peaks at 0.93, surpassing I-to-I (0.89) and T-to-T (0.85) baselines. While single-modal retrieval faces significant semantic gaps, exemplified

[3] For notational simplicity, we denote the point cloud by v.

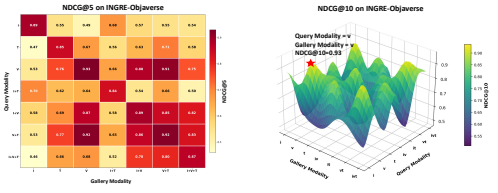
Method	NDCG @5	NDCG @10	Kendall's τ	Spearman's ρ
ReMU3D	0.94	0.93	0.41	0.54
w/o SCI	0.57	0.61	0.15	0.22
w/o Q-MoE	0.66	0.70	0.24	0.35
two-stage	0.41	0.47	0.04	0.07
fixed ratio	0.68	0.70	0.18	0.26

Table 3: Ablation results and training-strategy comparison.



(a) Number of experts. **(b)** Pooling strategy.

Fig. 5: Ablation visualizations.



(a) 2D heatmap of NDCG@5. **(b)** 3D surface plot of NDCG@10.

Fig. 6: Visualization of cross-modal retrieval.

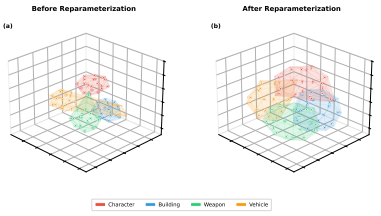


Fig. 7: t-SNE visualization of the embedding spaces before and after reparameterization.

by the low T-to-I score of 0.47, multi-modal fusion effectively bridges these disparities. Specifically, multi-modal queries outperform single-modal counterparts; for instance, the fused I+T query scores 0.64 when retrieving V, substantially improving over the single Image query (0.49). This trend culminates with the I+V+T alignment (0.87), confirming that integrating complementary modalities mitigates single-modal limitations and enhances robustness.

Extending evaluation to NDCG@10 provides a topological perspective corroborating NDCG@5 heatmap findings. The 3D surface plot exhibits a distinct diagonal ridge, confirming optimal alignment when query and gallery modalities match. Notably, V-to-V retrieval maintains a peak score of 0.93, identical to NDCG@5 performance. This stability across retrieval depths (k=5 vs. k=10) indicates the model concentrates highly relevant results at the top, maintaining consistent trends for both top-5 and top-10 candidates. Furthermore, surface topology illustrates multi-modal fusion enhancements; while pure cross-modal retrieval (e.g., Text-to-Image) occupies lower regions, fusion combinations (e.g., VT, IVT) create transitional slopes significantly elevating performance above single-modal baselines, reinforcing fusion robustness.

Visualization of Embedding Space Evolution To intuitively illustrate the impact of reparameterizing the pretrained CLIP semantic space, we employ t-SNE [48] to visualize the embedding distributions of 20 randomly selected samples from four distinct categories: Character, Building, Weapon, and Vehicle. As shown in Figure 7 (a), the original representation exhibits highly condensed clusters, where samples within the same category are densely grouped and oc-

cupy a limited volume of the feature space. In contrast, Figure 7 (b) reveals a significant shift following reparameterization: the spatial distribution of each category expands substantially, resulting in a more dispersed arrangement of individual samples. This increased intra-class variance suggests that the proposed method effectively mitigates the issue of semantic collapse, enabling the capture of fine-grained distinctions among intra-class instances while preserving clear inter-class separability.

Table 4: Label reliability.

Comparison	τ
Human–Human	0.76
MLLM–Human	0.68
Cross-MLLM	0.58

Table 5: Soft-label loss ablation.

Objective	N@5	N@10
InfoNCE	0.94	0.93
+ MSE	0.93	0.93
+ KL	0.92	0.91

5.6 Reliable analysis of INGRE

To evaluate the reliability of the generated labels, we further conduct a consistency study. Specifically, two independent 3D experts manually annotated a stratified random subset of 100 query-centered candidate lists. We additionally regenerated labels using a different MLLM (Claude Sonnet 4.6) with the identical prompts. As reported in Table 4, the agreement between the MLLM and human experts ($\tau = 0.68$) is close to the inter-human agreement ($\tau = 0.76$), indicating that the INGRE annotations are well aligned with expert judgments. Furthermore, the positive agreement between different MLLMs ($\tau = 0.58$) demonstrates that the generated labels remain reasonably consistent when the labeling model changes. Finally, as shown in Sec 5.3, ReMU3D also achieves strong performance on mainstream zero-shot embedding benchmarks, providing additional evidence that the proposed benchmark enables effective model learning.

6 Conclusion

We redefine cross-modal 3D retrieval as intra-class graded relevance estimation and present ReMU3D, a reparameterized multimodal unified 3D encoder employing SCI and Q-MoE for structured fusion and dynamic aggregation. Our method achieves strong discriminability for fine-grained ranking while preserving robust generalization. Leveraging an automated MLLM-driven pipeline, we introduce INGRE, showing state-of-the-art graded retrieval performance and competitive results on conventional benchmarks.

References

1. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: CVPR. pp. 13142–13152 (2023)

2. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: ICLR (2023)
3. Xue, L., Gao, M., Xing, C., Xu, J., Wu, H., Xiong, C., Xu, R., Socher, R., Niebles, J.C.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: CVPR. pp. 1179–1189 (2023)
4. Zhou, J., Wang, J., Ma, B., Liu, Y. S., Huang, T., Wang, X.: Uni3D: Exploring Unified 3D Representation at Scale. In: ICLR (2024)
5. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: CVPR. pp. 8552–8562 (2022)
6. Sanghi, A., Chu, H., Lambourne, J.G., Wang, Y., Cheng, C.Y., Niemeyer, M., Tulsiani, S.: Clip-forge: Towards zero-shot text-to-shape generation. In: CVPR. pp. 18603–18613 (2022)
7. Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D.: Shape distributions. ACM TOG **21**(4), 807–832 (2002)
8. Funkhouser, T., Min, P., Kazhdan, M., Chen, J., Halderman, A., Dobkin, D., Jacobs, D.: A search engine for 3d models. ACM TOG **22**(1), 83–105 (2003)
9. Uy, M.A., Huang, Q.H., Sung, M., Birdal, T., Guibas, L.J.: Deformation-aware 3d model embedding and retrieval. In: ECCV. pp. 397–413 (2020)
10. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
11. Han, Z., Shang, M., Liu, Z., Vong, C.M., Liu, Y.S., Zwicker, M., Han, J., Chen, C.P.: Y2seq2seq: Cross-modal representation learning for 3d shape and text by joint reconstruction and prediction. In: ICCV. pp. 9097–9106 (2019)
12. Huang, T., Dong, X., Ren, Y., Zhang, Y., Liu, W., Ran, R., Lu, J.: Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In: ICCV. pp. 22157–22167 (2023)
13. Xue, L., Yacoob, N., Niebles, J.C.: Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In: CVPR (2024)
14. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: CVPR. pp. 1912–1920 (2015)
15. Uy, M.A., Pham, Q.H., Hua, B.S., Nguyen, T., Yeung, S.K.: Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In: ICCV. pp. 1588–1597 (2019)
16. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3d shape recognition. In: ICCV. pp. 945–953 (2015)
17. Feng, Y., Zhang, Z., Zhao, X., Ji, R., Gao, Y.: Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In: CVPR. pp. 264–272 (2018)
18. Kanezaki, A., Matsushita, Y., Nishida, Y.: Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In: CVPR. pp. 5010–5019 (2018)
19. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR. pp. 652–660 (2017)
20. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: NeurIPS (2017)
21. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM TOG **38**(5), 1–12 (2019)
22. Li, Y., Su, H., Qi, C.R., Fish, N., Cohen-Or, D., Guibas, L.J.: Joint embeddings of shapes and images via cnn image purification. ACM TOG **34**(6), 1–12 (2015)

23. Qi, Z., Dong, R., Fan, G., Ge, Z., Zhang, X., Ma, K., Yi, L.: Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In: ICML. pp. 28223–28243 (2023)
24. Zhang, R., Wang, L., Qiao, Y., Gao, P., Li, H.: Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In: CVPR. pp. 21769–21780 (2023)
25. Koo, J., Huang, I., Achlioptas, P., Guibas, L.J., Sung, M.: Partglot: Learning shape part segmentation from language reference games. In: CVPR. pp. 16505–16514 (2022)
26. Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.: The princeton shape benchmark. In: Shape Modeling Applications. pp. 167–178 (2004)
27. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Qixing, H., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
28. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., et al.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: CVPR. pp. 803–814 (2023)
29. Mo, K., Zhu, S., Chang, A. X., Yi, L., Tripathi, S., Guibas, L. J., Su, H.: PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In: CVPR. pp. 909–918 (2019)
30. Mo, K., Guerrero, P., Yi, L., Su, H., Wonka, P., Mitra, N., Guibas, L.J.: Structurenet: Hierarchical graph networks for 3d shape generation. ACM TOG **38**(6), 1–19 (2019)
31. Downs, A., Francis, J., Ko, J., Kinman, B., Hickman, D., Reymann, K., McDonald, R., Vlimant, N., Stump, E.: Google Scanned Objects: A high-quality dataset of 3D scanned household items. In: ICRA. pp. 2553–2560 (2022)
32. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1/2), 81–93 (1938)
33. Spearman, C.: The proof and measurement of association between two things. The American Journal of Psychology **15**(1), 72–101 (1904)
34. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In: KDD. pp. 3505–3506 (2020)
35. Afham, M., Dissanayake, I., Dissanayake, D., et al.: Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In: CVPR. pp. 9902–9912 (2022)
36. Ahmed, M., Li, X., Prajapati, A., Elhoseiny, M.: 3dcompat200: Language-grounded compositional understanding of parts and materials of 3d shapes. arXiv preprint arXiv:2501.06785 (2025)
37. Chen, D.Y., Tian, X.P., Shen, Y.T., Ouhyoung, M.: On visual similarity based 3d model retrieval. Comput. Graph. Forum **22**(3), 223–232 (2003)
38. Chen, K., Choy, C.B., Savva, M., Chang, A.X., Funkhouser, T., Savarese, S.: Text2shape: Generating shapes from natural language by learning joint embeddings. In: Asian conference on computer vision. pp. 100–116. Springer (2018)
39. Feng, Y., Ji, S., Liu, Y.S., Du, S., Dai, Q., Gao, Y.: Hypergraph-based multi-modal representation for open-set 3d object retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(4), 2206–2223 (2023)
40. Fu, H., Jia, R., Gao, L., Gong, M., Zhao, B., Maybank, S., Tao, D.: 3d-future: 3d furniture shape with texture. International Journal of Computer Vision **129**(12), 3313–3337 (2021)

41. Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., Yue, X.: Onellm: One framework to align all modalities with language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26584–26595 (2024)
42. He, X., Ma, L., Cheng, Y., Wang, Z., Wang, Y., Zhou, Y., Bai, X.: Clip-adam: Adapting multi-view clip for open-set 3d object retrieval. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1022–1032 (2025)
43. Hou, W., Diao, Z., Peng, J.: Sketch-based 3d shape retrieval via cross-modal contrastive learning and difficulty-aware uncertainty regularization. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 521–534. Springer (2024)
44. Jiang, H., Liu, L., Wang, X., He, Y., Sui, W., Su, Z., Liu, W., Wang, X.: Spa3r: Predictive spatial field modeling for 3d visual reasoning. arXiv preprint arXiv:2602.21186 (2026)
45. Jing, L., Vahdani, E., Tan, J., Tian, Y.: Cross-modal center loss for 3d cross-modal retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3142–3151 (2021)
46. Jing, Z., Sun, Y., Li, Y.Y., Janarthanan, S., Deng, A., Hu, P.: Structure-aware fusion with progressive injection for multimodal molecular representation learning. arXiv preprint arXiv:2510.23640 (2025)
47. Liu, M., Xu, C., Jin, H., Chen, L., Xu, Z., Su, H.: Openshape: Scaling up 3d shape representation towards open-world understanding. In NeurIPS (2023)
48. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
49. Pan, Z., Lu, Y., Liu, H.: Metafind: Scene-aware 3d asset retrieval for coherent metaverse scene generation. arXiv preprint arXiv:2510.04057 (2025)
50. Pang, Y., Tay, E.H.F., Yuan, L., Chen, Z.: Masked autoencoders for 3d point cloud self-supervised learning. World Scientific Annual Review of Artificial Intelligence **1**, 2440001 (2023)
51. Pang, Y., Tay, E.H.F., Yuan, L., Chen, Z.: Masked autoencoders for 3d point cloud self-supervised learning. World Scientific Annual Review of Artificial Intelligence **1**, 2440001 (2023)
52. Tangelder, J.W., Veltkamp, R.C.: A survey of content based 3d shape retrieval methods. Proceedings Shape Modeling Applications, 2004. pp. 145–156 (2004)
53. Wang, Z., Li, Y., Liu, T., Zhao, H., Wang, S.: Ov-uni3detr: Towards unified open-vocabulary 3d object detection via cycle-modality propagation. In: European Conference on Computer Vision. pp. 73–89. Springer (2024)
54. Wang, Z., Zhou, Y., Liu, Z., Yu, R., Bai, S., Wang, Y., He, X., Bai, X.: Describe, adapt and combine: Empowering clip encoders for open-set 3d object retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21026–21036 (2025)
55. Willis, K.D., Pu, Y., Luo, J., Chu, H., Du, T., Lambourne, J.G., Solar-Lezama, A., Matusik, W.: Fusion 360 gallery: A dataset and environment for programmatic cad construction from human design sequences. ACM Transactions on Graphics (TOG) **40**(4), 1–24 (2021)
56. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J.: Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19313–19322 (2022)

57. Zhang, Y., Zhang, L., Ma, R., Cao, N.: Texverse: A universe of 3d objects with high-resolution textures. arXiv preprint arXiv:2508.10868 (2025)
58. Zhou, Z., Wang, P., Liang, Z., Bai, H., Zhang, R.: Cross-modal 3d representation with multi-view images and point clouds. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3728–3739 (2025)