Describe, Adapt and Combine: Empowering CLIP Encoders for Open-set 3D Object Retrieval

Supplementary Material

Sec. A provides more implementation details. Sec. B presents additional ablation studies on DAC. Sec. C discusses InternVL prompt choices and their effects. Sec. D demonstrates the applicability of DAC in scenarios where only point cloud data is available. Sec. E showcases more qualitative results. We also present additional experiments on both seen and unseen categories in Sec. F. Sec. G provides more comparisons with the state-of-the-arts by varying view numbers. Sec. H give some retrieval examples to gain some insights about the limitation of our framework. Finally, Sec. I investigates the impact of different multimodal large language models (MLLMs) on DAC performance.

A. More Dataset and Implementation Details

The four existing open-set 3DOR datasets, which are curated by Feng *et al.* [6] are described in detail below. *1) OS-ESB-core* is created based on ESB [9], which covers CAD objects of high genus (e.g., holes and tunnels) from the mechanical engineering domain. It includes only 98 training objects from 17 seen categories, 120 probe objects and 452 gallery objects from 24 unseen categories. *2) OS-NTU-core* has 378 training objects in 13 seen classes, 270 probe and 1,271 gallery objects in 54 unseen classes. Each object is coming from NTU [2]. *3)* OS-MN40-core is constructed from ModelNet40 [16]. It has 2,821 synthetic objects from 8 seen categories for training, 160 probe objects and 9,329 gallery objects from 32 unseen categories for testing. *4)* OS-ABO-core is a challenging, large-scale, real-word dataset with the 3D objects derived from real-word household items [4]. It contains 1,082 training objects in 4 seen categories, 85 probe objects and 5,455 gallery objects divided into 17 unseen categories.

For fair comparisons with off-the-shelf point cloud encoders, we further extend our framework by taking depth maps from point clouds. For this experiment, we have curated a zero-shot ZS-Objaverse-Core based on the large-scale Objaverse dataset. The original Objaverse [5] contains 46,832 shapes across 1,156 LVIS categories. We further split each category of Objaverse-LVIS into a query set and a target set with a 20%/80% ratio, resulting in a total of 8,798 query samples and 37,407 target samples. For the experiment, 10 depth maps are projected for each point cloud online following [21].

For a comprehensive comparison, we reimplement MV-CLIP [14] to evaluate its performance on open-set 3D datasets. For textual prompts, we use a pre-defined template: "a synthetic 3D model view of [cls] with different angles". It is important to note that we must provide *ground-truth category* to MV-CLIP for view selection, which is not suitable for open-set retrieval, where the category information for unseen categories is unknown. We select M_{selec} views based on entropy and perform mean pooling over them. Following MV-CLIP, M_{selec} is set to 4.

For ULIP-2 [18], we simply use the open-source PointBERT-CLIP ViT-G/14 pre-trained model as the backbone and directly employ the output features from the last layer of the model for retrieval. Similarly, For OpenShape [11], we use the PointBERT-CLIP ViT-B/32 and PointBERT-CLIP ViT-L/14 models, with the extracted features for retrieval directly. It is worth mentioning that we also experimented with fine-tuning both ULIP-2 and OpenShape for open-set setups. However, fine-tuning leads to worse performance, suggesting that the irregular point clouds are somewhat fragile representations that easily overfit to known categories in open-set setups.

B. More Ablations

B.1. Impact of Rank Number

We investigate the effect of decomposed matrix rank numbers in LoRA by setting it to 2, 4, 8, and 12. As shown in Table 1, increasing the rank from 2 to 8 leads to consistent improvements in mAP, NDCG, and ANMRR metrics. However, further increasing the rank to 12 results in a slight decline in performance. A rank of 8 strikes the best balance between performance and training complexity, and thus, all experiments are conducted with rank = 8.

LoRA Rank	mAP↑	NDCG↑	ANMRR↓
2	62.00	72.36	40.25
4	62.25	72.17	39.91
8	**62.40**	**72.63**	**39.82**
12	62.17	72.13	39.92

Table 1. Ablation of LoRA rank number with CLIP ViT-B/32 as the backbone on the OS-MN40-core dataset.

B.2. Impact of Fusion Weight α

The hyper-parameter α governs the relative weights of text features to image ones. To study its effects, we adjust the fusion ratio α within the range of 0 to 1 and conduct experiments on OS-MN40-core. The results are summarized

in Figure 1. As shown, an appropriate choice of α is essential for good retrieval performance. For instance, with CLIP ViT-L/14, we observe the performances are gradually improved by increasing α from 0 to 0.25. When α is set to 0.25, we have the best mAP of 68.98%. However, further increasing α results in a decline in performance. Similar phenomena are also observed when using CLIP ViT-B/32, as well as other datasets. In Table 2, we further provide the optimal values for α across all the datasets. Interestingly, we find that on the OS-ESB-core dataset, a smaller α gives better results (*i.e.*, 0.1). In contrast, for other datasets, especially the OS-ABO-core dataset, a larger value is preferred. We assume that it is difficult to derive accurate text embeddings for OS-ESB-core, which consists of high-genus mechanical parts [9]. In contrast, for common semantic real-world categories, text embeddings from InternVL are more accurate, and thus more weights are needed. We set these values as our default configurations for our experiments.

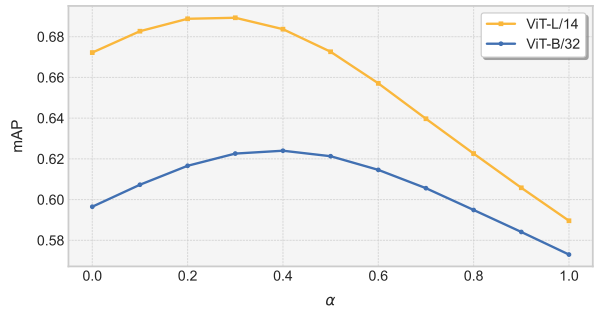

Figure 1. Impact of fusion weight α on OS-MN40-core.

Dataset	Backbone	α	mAP↑	NDCG↑	ANMRR↓
OS-ESB-core	ViT-B/32	$\alpha = 0.1$	58.70	24.27	45.67
	ViT-L/14	$\alpha = 0.1$	57.80	24.36	47.44
OS-NTU-core	ViT-B/32	$\alpha = 0.6$	59.21	27.06	44.58
	ViT-L/14	$\alpha = 0.3$	65.83	28.78	37.46
OS-MN40-core	ViT-B/32	$\alpha = 0.4$	62.40	72.63	39.82
	ViT-L/14	$\alpha = 0.25$	68.98	77.59	33.87
OS-ABO-core	ViT-B/32	$\alpha = 0.85$	66.10	59.01	36.12
	ViT-L/14	$\alpha = 0.7$	70.74	60.87	32.14

Table 2. Optimal α values across different datasets and backbones.

B.3. Impact of Normalization

After fusing image and text features with elementwise summation, it is beneficial to further utilize activation functions for normalization. We have evaluated three commonly-used activation functions: *ReLU*, *Tanh*, and *Sigmoid*, on the OS-ABO-core dataset. As shown in Table 3, with *ReLU*, DAC attains inferior results. However, when adopting *Sigmoid* and *Tanh*, we observe consistent improvements, with *Tanh* being more advantageous. Especially, for CLIP ViT-B/32, *Tanh* brings an improvement of over 1% mAP. It demonstrates that adopting *Tanh* for normalization is an effective option to further augment the discriminativeness of the derived 3D descriptors.

We also analyze why *Tanh* yields better results. The *Tanh* function maps input values to the range $[-1, 1]$, aligning well with the vector space used by CLIP for retrieval. This property helps the model produce evenly distributed outputs that are consistent with CLIP's embedding structure. In contrast, *ReLU* sets negative values to zero, retaining only positive values, which can lead to information loss and disrupt the symmetry required for the relative relationships in CLIP's vector space. On the other hand, *Sigmoid* restricts outputs to the $[0, 1]$ range, which can weaken vector directionality and produce many low-magnitude weights, making it less compatible with CLIP's feature distribution. Thus, *Tanh* emerges as the most suitable choice, preserving the necessary feature balance and enhancing the robustness of the 3D descriptors for retrieval.

Backbone	Activation Function	mAP↑	NDCG↑	ANMRR↓
ViT-B/32	-	64.63	58.52	37.79
	ReLU	63.13	58.03	39.08
	Sigmoid	65.58	58.82	37.08
	Tanh	**66.10**	**59.01**	**36.12**
ViT-L/14	-	69.33	60.43	32.98
	ReLU	68.79	60.23	33.62
	Sigmoid	69.95	60.60	32.83
	Tanh	**70.74**	**60.87**	**32.14**

Table 3. Impact of Activation Functions on OS-ABO-core.

B.4. Training strategy

We further investigate the impact of different textual descriptions during LoRA training and their influence on text fusion during retrieval. We have tested two settings: the first uses a fixed, hand-crafted template ("a synthetic 3D model view of [cls] with different angles"), while the second employs descriptions generated by InternVL [3] for each class. The experimental results are summarized in Table 4.

As shown, both training strategies yield comparable performance when retrieval is conducted using only image features. However, descriptions generated by InternVL demonstrate superior results in scenarios where text fusion is applied during retrieval. This improvement can be attributed to the consistency between the descriptions used during training and those generated by InternVL for text fusion in the retrieval phase. Fine-tuning with InternVL-generated descriptions enables more seamless integration of visual and textual features, enhancing the retrieval performance.

We have also experimented with providing InternVL with view projections and label information during training, allowing it to generate a description for each training sam-

Description	Method	OS-ESB-core	OS-NTU-core	OS-MN40-core	OS-ABO-core
Hand-crafted	*Images only*	57.19 / 23.95 / 47.18	54.66 / 25.73 / 49.17	58.96 / 71.65 / 42.94	56.70 / 56.05 / 44.55
	Images with texts	**+1.15 / +0.23 / -0.76**	**+3.81 / +1.13 / -4.24**	**+2.88 / +0.82 / -2.65**	**+9.06 / +2.80 / -7.98**
Generated	*Images only*	57.45 / 23.96 / 47.13	54.57 / 25.80 / 49.08	59.35 / 71.89 / 42.72	56.45 / 55.91 / 45.33
	Images with texts	**+1.25 / +0.31 / -1.46**	**+4.64 / +1.26 / -4.50**	**+3.05 / +0.74 / -2.90**	**+9.65 / +3.10 / -9.21**

Table 4. Impact of different training text descriptions with ViT-B/32 as the backbone. Values are presented in mAP/NDCG/ANMRR format.

ple. While this approach may appear to better align with the testing process, we have observed that the unstable descriptions generated for individual samples lead the model to focus too much on the specific characteristics of each sample, neglecting the learning of more stable, category-level features. This phenomenon is more pronounced in the OS-ESB-core dataset with fewer samples. As shown in Table 5, using individual-level descriptions achieved only 48.78 mAP, which is significantly lower than the 58.70 mAP obtained with category-level descriptions. Therefore, we have opted to provide a single description per category rather than per sample, as this can yield more consistent and effective results.

Dataset	Description	mAP↑	NDCG↑	ANMRR↓
OS-ESB-core	*Individual-level*	48.78	21.90	54.37
	Category-level	**58.70**	**24.27**	**45.67**

Table 5. Impact of different description generation methods with ViT-B/32 as the backbone.

C. More Choices for InternVL Prompts

The selection of prompts is crucial for the responses generated by InternVL, yet finding a universal prompt that fits all scenarios is quite challenging. Using ViT-L/14 as the backbone, we have tested the impact of two different prompts:

- Prompt A: " There are images of an object from different angles. Describe this object in one sentence." (This is the default prompt used in our method.)

- Prompt B: " There are images of an object from different angles. Describe this object's shape information in one sentence."

Prompt A attempts to derive explicit category information, while prompt B attempts to derive descriptive shape information. The experimental results are detailed in Table 6. From the table, we can observe that in the OS-ESB-core, OS-NTU-core, and OS-MN40-core datasets, the results from both prompts are relatively similar. However, in the OS-ABO-core dataset, the effect of the prompt is more pronounced. For instance, the mAP for Prompt A is 68.40%, whereas Prompt B yields a lower mAP of 66.43%.

We randomly select a chair object from the OS-ABO-core dataset and input it into InternVL using different prompts, resulting in the following two outputs: "A classic brown leather armchair with a high back and rounded armrests, featuring subtle nailhead trim along its edges," and "This object has a rounded, high-backed shape with a broad seat, armrests, and a slight outward curve on the back and arms." We hypothesize that the statement generated by Prompt A benefits from more explicit label information, leading to superior results. Conversely, inaccurate label information could adversely affect the model's expressiveness. Therefore, identifying a universally applicable prompt is quite challenging, and further exploration into prompt design is essential.

Dataset	Prompt	mAP↑	NDCG↑	ANMRR↓
OS-ESB-core	A	57.80	24.36	47.44
	B	57.89	24.37	47.31
OS-NTU-core	A	65.83	28.78	37.46
	B	65.58	28.74	37.59
OS-MN40-core	A	68.98	77.59	33.87
	B	68.81	77.56	34.18
OS-ABO-core	A	**70.74**	**60.87**	**32.14**
	B	68.68	59.23	33.57

Table 6. Impact of different prompts.

Why not just let InternVL judge the categories? We also conduct a simple experiment to evaluate the performance of directly using InternVL for category classification. On the challenging OS-ESB-core dataset, even when provided with ground-truth category options, InternVL achieved only 11% accuracy. Incorrect category predictions in such cases can be catastrophic for our task, as they severely disrupt feature representation and retrieval. Therefore, instead of relying on InternVL for classification, we opted to use it for generating descriptive information, which provides more robust and generalized representations.

In conclusion, while prompt selection plays a critical role, the design space for prompts remains vast, and further exploration is required to optimize their effectiveness.

D. Extending to Point Cloud Retrieval

Setup. For this experiment, we curate ZS-Objaverse-Core based on Objaverse-LVIS, which is an annotated subset of

Objaverse [5], for zero-shot point cloud retrieval. For fair comparisons with off-the-shelf point cloud encoders OpenShape [11] and ULIP series [17, 18], our baseline only takes depth images from point clouds with the online projection scheme [21].

Results. As shown in Table 7, our method based on depth maps also achieves superior performance, surpassing ULIP-2 by +1.22% mAP. Note that all compared methods require huge resources to train on a large-scale 3D dataset of point cloud, text, and image triplets. By contrast, our baseline offers a much cheaper solution without 3D training.

Method	Backbone	mAP↑	NDCG↑	ANMRR↓
OpenShape (point cloud)	PointBERT-CLIP ViT-L/14	11.93	14.10	85.40
ULIP (point cloud)	PointMLP - SLIP	6.69	9.17	90.82
ULIP-2 (point cloud)	PointBERT - CLIP ViT-G/14	18.15	19.34	79.03
Ours (depth image)	**CLIP ViT-L/14**	**19.37**	**20.15**	**78.23**

Table 7. **Performance on ZS-Objaverse-Core**.

E. Qualitative Results

To further understand how our method improves the representations, we visualize correlation heatmaps. Specifically, we randomly select three categories from OS-ABO-core, with ten samples chosen from each category. We draw three heatmaps: the first represents features extracted from the original CLIP, the second shows features extracted using CLIP with LoRA added, and the third depicts features obtained after fusing with InternVL on top of the LoRA-enhanced features. The detailed visual results are illustrated in Figure 2. As shown, we observe that, compared to the original CLIP, the addition of LoRA significantly increases the similarity among samples within the same category while reducing similarity with samples from different categories. This effect becomes even more evident in the third figure, which incorporates text features extracted by InternVL. This indicates that DAC not only enhances intra-class similarity but also effectively diminishes inter-class similarity, thereby markedly improving the discriminative power of the retrieval features.

F. Retrieval on Seen and Unseen Categories

In real-world applications, the ability to retrieve 3D objects of both seen and unseen categories is crucial. In this section, we follow HGM^{2}R [6] and split the ModelNet40 dataset into two subsets: D_{S} (for seen categories) and D_{U} (for unseen categories). Each subset consists of 20 categories, and the 3D objects within each category are further divided into training sets $D_{\mathrm{S}}^{\mathrm{tr}}/D_{\mathrm{U}}^{\mathrm{tr}}$ and retrieval sets $D_{\mathrm{S}}^{\mathrm{re}}/D_{\mathrm{U}}^{\mathrm{re}}$, with 80% of the data used for training and 20% for retrieval. The models are trained on $D_{\mathrm{S}}^{\mathrm{tr}}$ and evaluated separately on the seen categories $D_{\mathrm{S}}^{\mathrm{re}}$ and unseen categories $D_{\mathrm{U}}^{\mathrm{re}}$.

As shown in Table 8, our method has competitive performance on seen categories compared to other approaches by solely relying on multi-view images. More importantly, we achieve superior results on unseen categories, reaching an mAP of 86.27%, surpassing previous state-of-the-art HGM^{2}R [6] by a large margin. Notably, our model demonstrates a reduced performance gap between seen and unseen categories, with only a 4.85% performance difference—significantly lower than the 11.87% gap observed with HGM^{2}R and the 19.73% gap with InfoNCE. This indicates that our approach is particularly effective for retrieving unknown categories, allowing for an enhancement in unseen performance while maintaining competitive results on seen categories. Thus, our method is especially advantageous in complex environments, where retrieval performance for unseen categories is crucial.

Method	On Seen Categories		On Unseen Categories	
	mAP↑	Recall@100↑	mAP↑	Recall@100↑
TCL [7]	93.50	82.14	73.92	71.76
MMJM [12]	91.99	80.78	73.07	71.38
SDML [8]	88.50	78.50	74.69	72.39
CMCL [10]	90.99	79.60	75.21	72.49
MMSAE [15]	88.72	78.61	76.03	72.94
MCWSA [19]	85.70	76.83	72.89	70.56
PROSER [20]	87.71	77.78	74.93	72.56
InfoNCE [13]	93.65	82.19	73.92	71.64
HGM^{2}R [6]	**94.10**	**82.47**	82.23	78.21
Ours (ViT-B/32)	87.96	77.89	**83.00**	**79.10**
Ours (ViT-L/14)	91.12	80.46	**86.27**	**81.76**

Table 8. Separate retrieval results on both seen and unseen categories.

G. More Experiments on View Numbers

We further study the impact of the number of view images on the OS-MN40-core dataset. The adjustment of view numbers affects two components: the number of views input into CLIP and the number of views input into InternVL. In our experimental setup, the view counts for both components are kept the same. As shown in Figure 4, the mAP values increase with the number of views across different backbones. It suggests that additional views provide more detailed and accurate information about 3D objects, giving better retrieval performance.

We also compare with other competing methods under the same number of views, as summarized in Table 9. The results indicate that our approach, particularly with the ViT-L/14 backbone, outperforms other methods in both the 4-view and 12-view settings. Specifically, we achieve 64.12% mAP and 68.08% mAP for the ViT-L/14 backbone, respectively, surpassing the previous best method HGM^{2}R greatly.

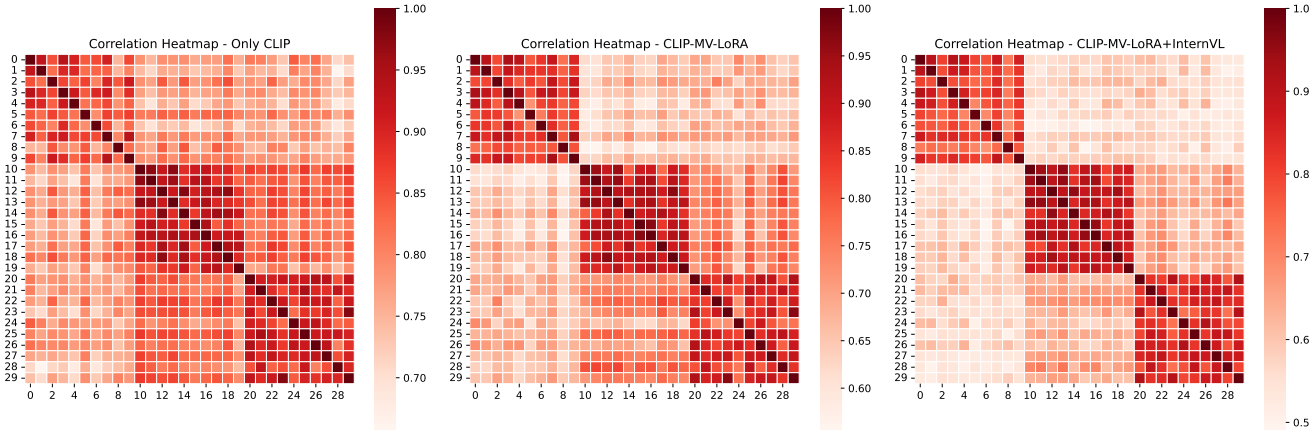

Figure 2. Correlation heatmaps of features from 3 randomly selected classes, each with 10 samples on OS-ABO-core. Darker red indicates higher similarity.

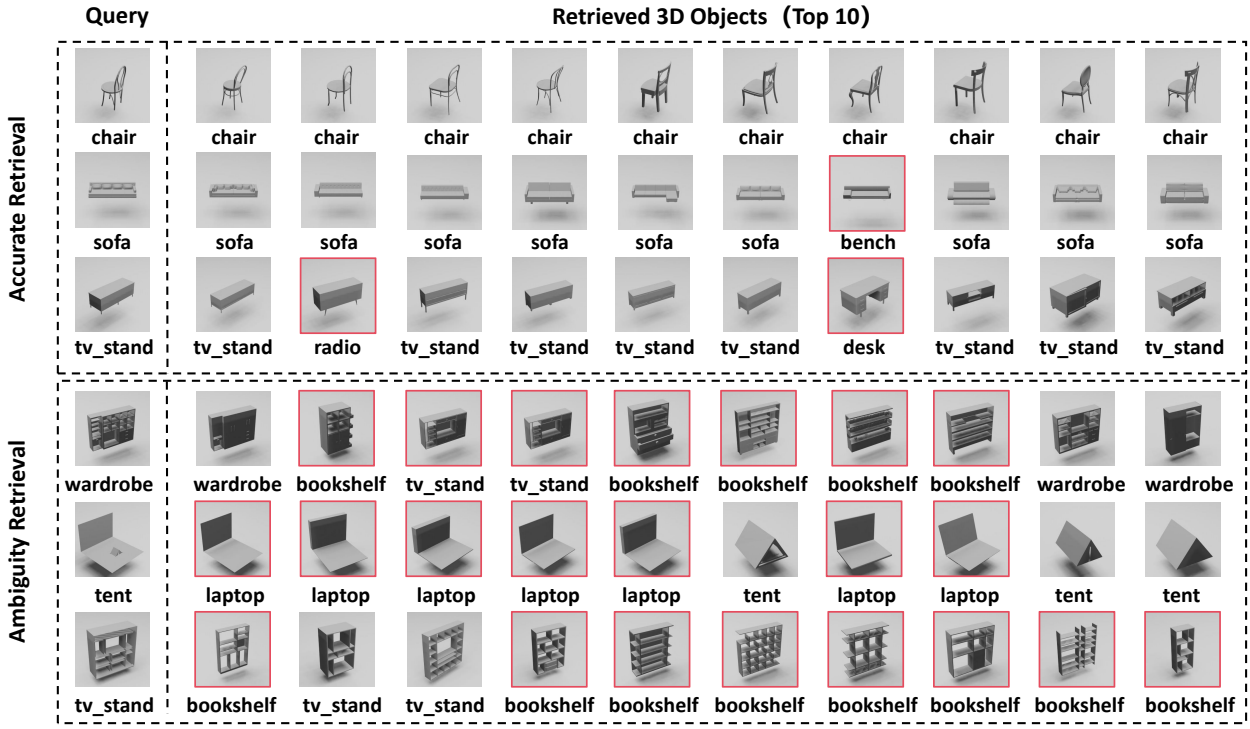

Figure 3. Retrieval examples on OS-MN40-core. Incorrect matches are in red boxes.

H. More Visualization

To gain more insights into our framework, we provide some retrieval examples of our method, especially including some failure cases, on OS-MN40-core. As shown in Figure 3, for objects of easy categories (*e.g.*, chair), our method produces discriminative 3D representations for accurate retrieval. However, it fails when two objects have similar global appearances but from distinct categories. For instance, a tent instance (row 5) globally looks like a laptop object. Yet, notice that a laptop has ver distinct local features on the integrated keyboard. The keyboard serves as a strong discriminative cue for identifying a laptop. In the future, we plan to emphasizing these local features during the representation learning process, which could potentially avoid these failure cases.

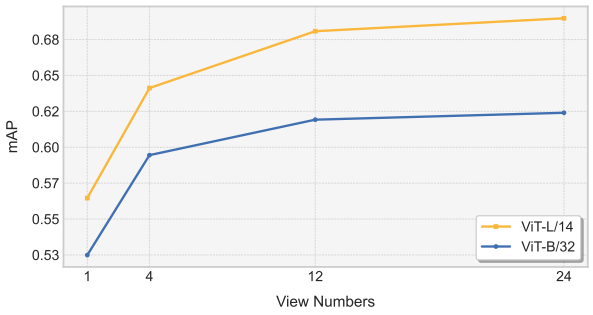

Figure 4. Impact of View Numbers with different backbones.

Method	Number of Views	
	4-v	12-v
TCL [7]	46.46	47.36
MMJM [12]	46.81	48.11
SDML [8]	49.60	50.75
CMCL [10]	50.06	51.38
MMSAE [15]	50.85	52.09
MCWSA [19]	47.28	48.78
PROSER [20]	48.45	49.00
InfoNCE [13]	46.46	47.37
HGM^{2}R [6]	63.36	64.20
Ours (ViT-B/32)	59.45	61.92
Ours (ViT-L/14)	**64.12**	**68.08**

Table 9. Performance comparison on view numbers.

I. More Choices of MLLM

Our framework is compatible with any off-the-self pre-trained MLLM, enabling seamless integration of the latest advancements in multimodal learning. To study it, we experiment with different choices for multi-modal large language models (MLLMs). Table 10 shows that DAC's performance improves progressively as InternVL [3] scales from 1B to 8B parameters. It suggests that MLLMs with stronger reasoning capabilities lead to better results. Furthermore, the use of Qwen2.5-VL [1] further enhances DAC performance, highlighting the potential of DAC.

MLLM	mAP↑	NDCG↑	ANMRR↓
InternVL-1B [3]	61.48	72.64	40.60
InternVL-4B [3]	62.40	72.63	39.82
InternVL-8B [3]	63.08	72.93	39.13
Qwen2.5-VL-3B [1]	63.24	73.16	38.88
Qwen2.5-VL-7B [1]	66.72	75.99	35.86

Table 10. More Choices of MLLM.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[2] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3d model retrieval. In *Computer graphics forum*, pages 223–232, 2003. 1

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 2, 6

[4] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *CVPR*, pages 21126–21136, 2022. 1

[5] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, pages 13142–13153, 2023. 1, 4

[6] Yifan Feng, Shuyi Ji, Yu-Shen Liu, Shaoyi Du, Qionghai Dai, and Yue Gao. Hypergraph-based multi-modal representation for open-set 3d object retrieval. *IEEE TPAMI*, 2023. 1, 4, 6

[7] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3d object retrieval. In *CVPR*, pages 1945–1954, 2018. 4, 6

[8] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 635–644, 2019. 4, 6

[9] Subramaniam Jayanti, Yagnanarayanan Kalyanaraman, Natraj Iyer, and Karthik Ramani. Developing an engineering shape benchmark for cad models. *Computer-Aided Design*, 38(9):939–953, 2006. 1, 2

[10] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. Cross-modal center loss for 3d cross-modal retrieval. In *CVPR*, pages 3142–3151, 2021. 4, 6

[11] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *NeurIPS*, 36, 2023. 1, 4

[12] Weizhi Nie, Qi Liang, An-An Liu, Zhendong Mao, and Yangyang Li. Mmjn: Multi-modal joint networks for 3d shape recognition. In *ACM MM*, pages 908–916, 2019. 4, 6

[13] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4, 6

[14] Dan Song, Xinwei Fu, Weizhi Nie, Wenhui Li, and Anan Liu. Mv-clip: Multi-view clip for zero-shot 3d shape recognition. *arXiv preprint arXiv:2311.18402*, 2023. 1

[15] Yiling Wu, Shuhui Wang, and Qingming Huang. Multi-modal semantic autoencoder for cross-modal retrieval. *Neurocomputing*, 331:165–175, 2019. 4, 6

[16] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920, 2015. 1

[17] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *CVPR*, pages 1179–1189, 2023. 4

[18] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *CVPR*, pages 27091–27101, 2024. 1, 4

[19] Jiahao Zheng, Sen Zhang, Zilu Wang, Xiaoping Wang, and Zhigang Zeng. Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition. *IEEE TMM*, 2022. 4, 6

[20] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *CVPR*, pages 4401–4410, 2021. 4, 6

[21] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Point-clip v2: Prompting clip and gpt for powerful 3d open-world learning. In *ICCV*, pages 2639–2650, 2023. 1, 4