Appendix of Multi-modal Relation Distillation for Unified 3D Representation Learning

1 More Ablation Studies

Hyper-parameter. We perform an ablation study to examine the impact of the tunable hyper-parameter λ in Eq. (13). The results are depicted in Tab. A.

From Tab. A, we find that the optimal performance across all three benchmarks is achieved when λ is set to 3. Consequently, we adopt $\lambda = 3$ for all the experiments within this study.

Table A: Ablation results (%) on the Impact of the hype-rparameter λ.

Setting	Objaverse-LVIS			ModelNet40			ScanObjectNN		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
Base	11.0	20.0	24.6	72.2	85.5	88.8	53.4	73.5	81.9
$\lambda = 1$	11.8	21.2	**26.0**	72.9	87.8	92.1	55.5	75.2	83.5
$\lambda = 2$	11.5	20.8	25.7	72.9	88.0	**92.8**	**55.8**	75.4	84.3
$\lambda = 3$	**11.8**	**21.2**	25.8	**74.2**	**88.3**	90.9	55.7	**75.7**	83.0
$\lambda = 4$	11.5	20.9	25.9	73.7	87.6	91.5	53.7	74.9	**84.4**
$\lambda = 5$	11.5	20.8	25.6	72.4	87.2	91.3	52.5	72.5	81.9

Downstream Finetuning. We follow the same fine-tuning protocols as ULIP on standard 3D classification and present the results in **Tab. B**. MRD largely surpasses the baseline and other counterparts in both the settings.

Data Source. Since ULIP2 generates image-text pairs for 3D point clouds using different protocols, we train MRD-T and MRD-S on ShapeNet with the image-text pairs generated by ULIP2 to evaluate the data source generalization ability of MRD. The hyper-parameters are set as stated in Sec. 4.1, consistent with the previous ablation studies in the main text. The results are shown in Fig. A. As depicted, MRD-T and MRD-S achieve better performance when trained with the image-text data from ULIP2, demonstrating the generalization ability.

2 Details in Cross-Modal Retrieval

For quantitative assessment in Sec. 4.3, we use the Objaverse test set, denoted as $\{(p_i, l_i)\}_{i=1}^{N}$ as our evaluation set. The encoded 3D point clouds are employed as cross-modal queries to retrieve the corresponding detailed text descriptions $\{t_i\}_{i=1}^{N}$ generated by Cap3D. A retrieval is deemed a success at the instance level

Table B: Results (%) standard 3D classification on ScanObjNN. $*$ indicates the results obtained by re-implementation.

	PointBERT-22M			PointBERT-38M		
Pretrained	-	ULIP	MRD	-	OpenShape	MRD
Ins. Acc.	83.1	86.4	89.6	82.5*	85.3*	89.2

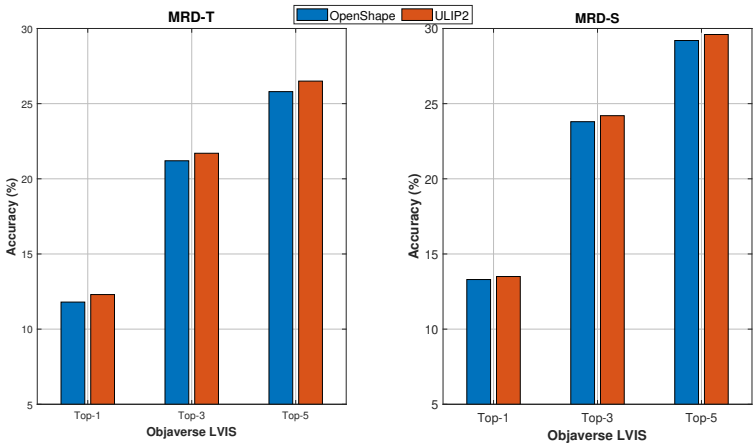

Fig. A: Comparison of zero-shot accuracy (%) on Objaverse When trained with different data sources.

if the retrieved item is the i-th item itself. Additionally, a more lenient criterion is applied to category-wise retrieval: if the retrieved item belongs to the same category l_i as the target, it is also considered a successful case.

For fair comparison, we utilize the released OpenShape-PointBert and Uni3D-B models as well as MRD-B, which are all on a comparable scale in terms of parameters. The retrieval accuracy is reported in Tab.2 in the main text. Additionally, more qualitative comparison is visualized in Fig. B, wherer we can find that MRD achieves the superior performance compared to the counterparts.

We further visualize additional image-query 3D shape results generated by MRD in Fig. C.

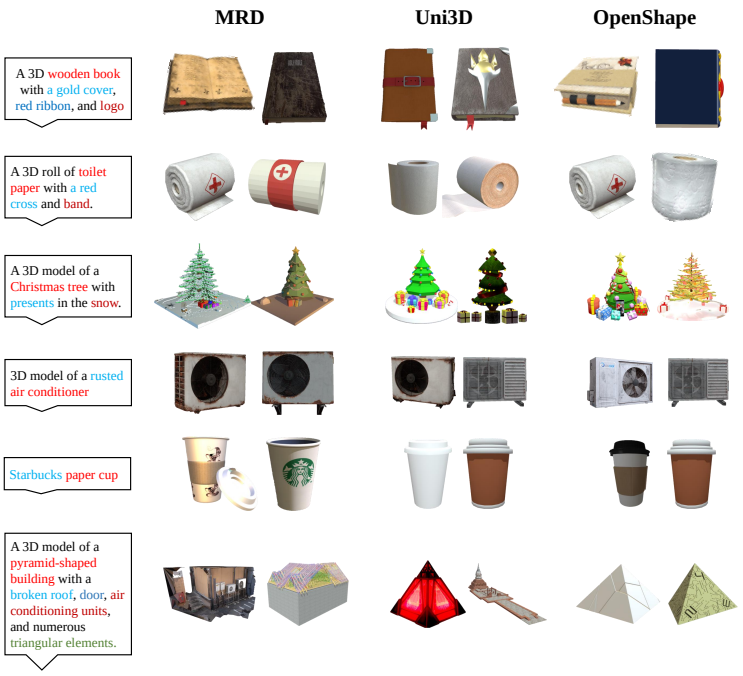

Fig. B: Comparison of text-query 3D shape retrieval results. For every query text, we retrieve two 3D shapes that match most closely. Text descriptions feature words in different colors to highlight the diverse attributes of the retrieval targets. The results for MRD, Uni3D, and OpenShape are presented across columns 2 to 4.

Fig. C: Visualization of more image-query 3D shape retrieval results. Input images are from unsplash.com.