Table 1: Backbone details of baseline methods.

Baseline	Image & Text encoders	3D Backbone
PointCLIP	CLIP with ResNet50	N/A
PointCLIP v2	CLIP with ViT-B/16	N/A
ReCon	CLIP with ViT-B/32	PointBERT
CG3D	SLIP with ViT-B/16	PointTransformer
CLIP2Point	CLIP with ViT-B/32	ViT-B/32 as 2D depth encodei
ULIP	SLIP with ViT-L/16	PointBERT

Table 2: Comparison between various image and text encoders.

Image & Text Encoder	3D Backbone	Training Data	ModelNet40	Objaverse-LVIS
OpenCLIP ViT-bigG-14	SparseConv	ShapeNet	69.5	8.7
CLIP ViT-L/14	SparseConv	ShapeNet	68.1	8.5
OpenCLIP ViT-bigG-14	SparseConv	Objaverse	79.2	41.6
CLIP ViT-L/14	PointBERT	Objaverse	77.0	42.3

Figure 1: **Point cloud input 3D shape retrieval.** In each example, the left image shows the input shape, the first row on the right shows the retrieved shapes using ULIP (retrained), and the second row on the right shows the retrieved shapes using OpenShape.

Table 3: Image-shape alignment.

(a) Image-shape matching on size-196 batches of image-shape pairs. S2I and I2S are matching accuracy from shape to image and vice versa.

Model	S2I	I2S	Avg.
ULIP (retrained)	71.94	77.04	74.49
OpenShape	**92.86**	**94.90**	**93.88**

(b) Real-world image input retrieval.

Model	R@10	R@100
ULIP (retrained)	7.0	22.5
OpenShape	**43.9**	**69.3**