ULIP-2: Towards Scalable Multimodal Pre-training for 3D Understanding

Le Xue[1], Ning Yu[1], Shu Zhang[1], Artemis Panagopoulou[1,3], Junnan Li[1], Roberto Martín-Martín[4], Jiajun Wu[2], Caiming Xiong[1], Ran Xu[1], Juan Carlos Niebles[1,2], Silvio Savarese[1,2]

[1]Salesforce AI Research [2]Stanford University [3]University of Pennsylvania [4]University of Texas at Austin

CVPR SEATTLE, WA JUNE 17-21, 2024

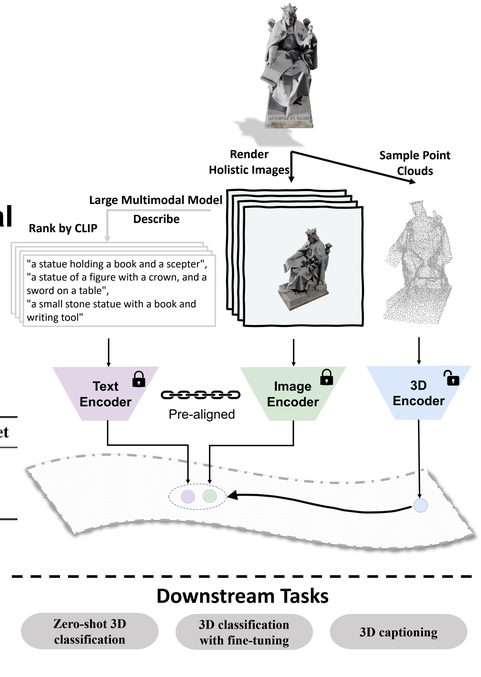
Introduction:

- Recent works like ULIP have shown promising 3D representation learning by aligning features from 3D, 2D counterparts, and language.
- However, existing methods often lack scalability and fail to deliver comprehensive language descriptions.
- We introduce ULIP-2, which utilizes LMMs to automatically produce holistic textual descriptions from only 3D data, eliminating manual annotations.
- Besides better scalability, ULIP-2 also sets new SOTA on various downstream tasks.

Two Large-scale Tri-modal Datasets are Released:

- ULIP-Objaverse Triplets
- ULIP-ShapeNet Triplets

Modality	ULIP-Objaverse	ULIP-ShapeNet
Point Clouds	~ 800k	~ 52.5k
Images	~ 10 million	~ 3 million
Language	~ 100 million	~ 30 million



Proposed Approach:

- In ULIP-2, **only 3D shape data is required.**
- We extract 3D point clouds from the surface.
- Then render images from various viewing angles.
- We then leverage BLIP-2 to generate holistic texts for each rendering.
- For each image, we generate 10 sentences, rank using CLIP, and aggregate the top-1 descriptions to form a holistic language modality.
- We scale both the tri-modal datasets and the encoders (3D and CLIP) for pre-training.

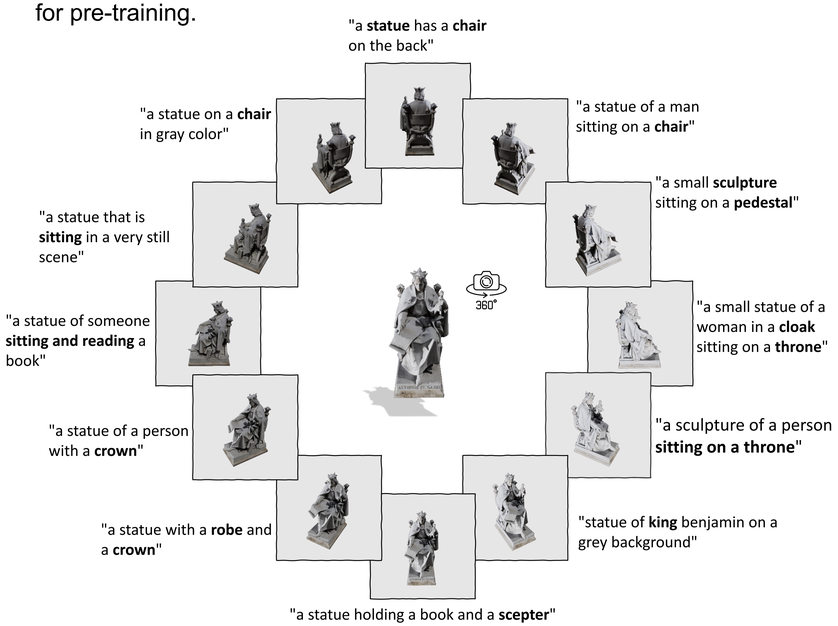


Illustration of our scalable tri-modal dataset creation framework.

Experiments:

- **Zero-Shot 3D Classification**

Model	Pre-train dataset	Pre-train method	Manual captions?	Objaverse-LVIS top-1	Objaverse-LVIS top-5	ModelNet40 top-1	ModelNet40 top-5
PointCLIP [58]	–	–	–	1.9	5.8	19.3	34.8
PointCLIPv2 [62]	–	–	–	4.7	12.9	63.6	85.0
ReCon [34]	ShapeNet	ReCon [34]	✓	1.1	3.7	61.2	78.1
CLIP2Point [11]	ShapeNet	CLIP2Point [11]	✗	2.7	7.9	49.5	81.2
Point-BERT [55]	ShapeNet	OpenShape [22]	✓	10.8	25.0	70.3	91.3
Point-BERT [55]	Objaverse(no LVIS) + ShapeNet	OpenShape [22]	✓	38.8	68.8	83.9	97.6
Point-BERT [55]	Objaverse + ShapeNet	OpenShape [22]	✓	46.5	76.3	82.6	96.9
Point-BERT [55]	Objaverse + ShapeNet + (2 extra)	OpenShape [22]	✓	46.8	77.0	84.4	**98.0**
Point-BERT [55]	ShapeNet	ULIP [52]	✓	2.6	8.1	60.4	84.0
		ULIP-2	✗	16.4	34.3	75.2	95.0
	Objaverse(no LVIS) + ShapeNet	ULIP [52]	✓	21.4	41.9	68.6	86.4
		ULIP-2	✗	46.3	75.0	84.0	97.2
	Objaverse + ShapeNet	ULIP [52]	✓	34.9	61.0	69.6	85.9
		ULIP-2	✗	**50.6**	**79.1**	**84.7**	97.1

- **Finetune for 3D Classification on ScanObjectNN**

Model	#Params(M)	Overall Acc	Class-mean Acc
PointNeXt (from scratch)	1.4	87.5	85.9
PointNeXt (w/ ULIP-2)	1.4	**91.5** (↑ 4.0)	**90.9** (↑ 5.0)

- **3D-to-Language Generation**

Multimodal generation framework	Frozen 3D encoder	CIDEr score
X-InstructBLIP	PB w/ULIP	132.2
X-InstructBLIP	PB w/ULIP-2	**160.5**(↑ 28.3)

Our Github repo includes:

- **Pre-trained models**
- **Pre-train datasets**