Any2Point: Empowering Any-modality Large Models for Efficient 3D Understanding

Yiwen Tang[1,2]*, Ray Zhang[3]*, Jiaming Liu[4]*, Zoey Guo[3]*,
Bin Zhao[1,2]†, Zhigang Wang[1], Peng Gao[1], Hongsheng Li[3],
Dong Wang[1]†, Xuelong Li[5]

[1] Shanghai Artificial Intelligence Laboratory [2] Northwestern Polytechnical University
[3] The Chinese University of Hong Kong [4] Peking University [5] TeleAI

stutangyw@gmail.com

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

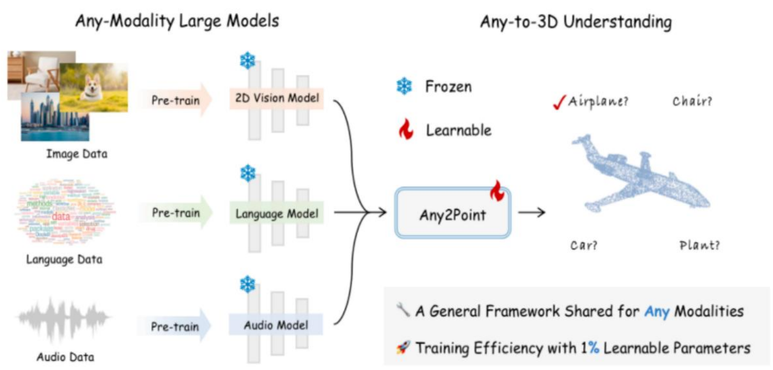
Paper code

ECCV 2024 EUROPEAN CONFERENCE ON COMPUTER VISION MILANO

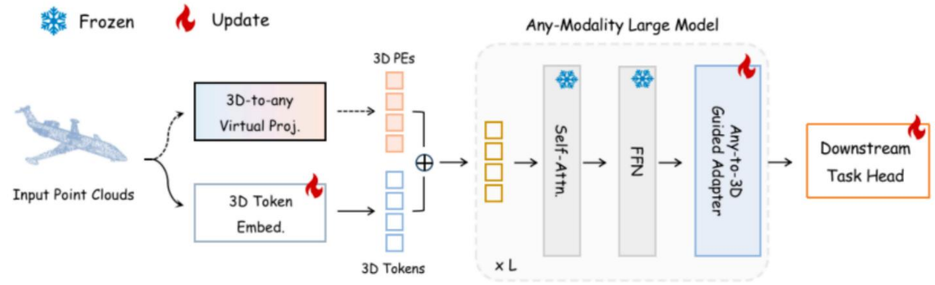
Motivation

- The major drawbacks of existing 2D-to-3D works include the loss of spatial information during data modality transformation and the substantial computational and data engineering costs associated with cross-modality knowledge distillation.
- Can we develop a general any-to-3D paradigm that empowers any-modality large models for efficient and effective point cloud understanding?

Overview



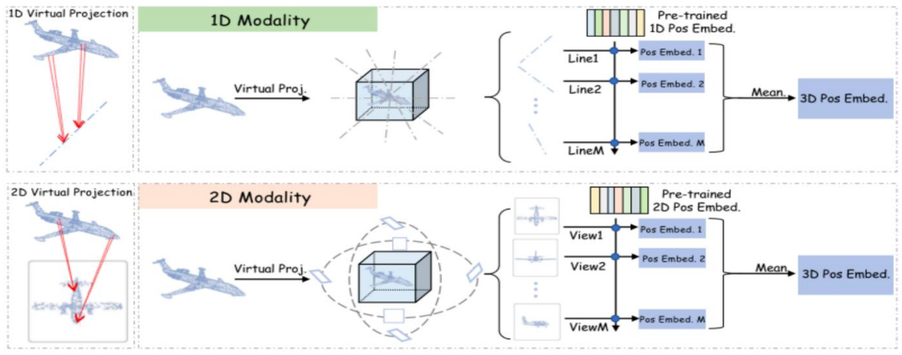
Pipeline



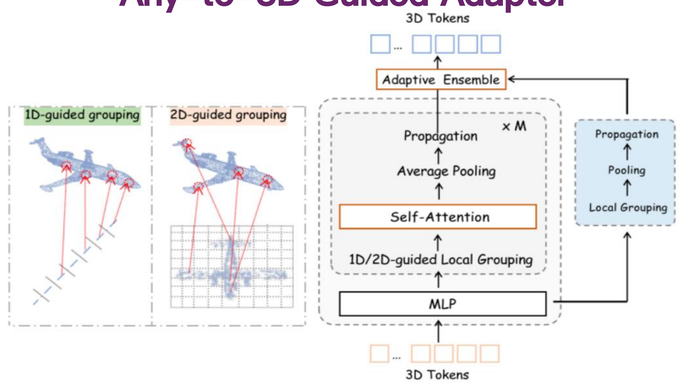
Contributions

- **To enable a general any-to-3D transferring framework**, we propose Any2Point, which empowers any-modality pre-trained large models (e.g., 2D vision, language, and audio) for efficient 3D understanding.
- We introduce two techniques, i.e., 3D-to-any virtual projection and any-to-3D guided adapter, to effectively overcome the issues within current methods, such as 3D geometry loss and excessive resource cost.

3D-to-Any Virtual Projection



Any-to-3D Guided Adapter



Method	Pre-train	#Param(M)	SCAN.(%)	MN.(%)
Point-NN	N/A	0.0	64.9	81.8
PointNet	N/A	3.5	68.0	89.2
PointNet++	N/A	1.5	77.9	90.7
DGCNN	N/A	1.8	78.1	92.9
PointMLP	N/A	12.6	85.4	94.1
Point-PN	N/A	0.8	87.1	93.8
PointNeXt	N/A	1.4	87.7	94.0
Point-BERT	3D	22.1	83.1	92.7
w/ Point-PEFT	3D	0.6	85.0	93.4
Point-MAE	3D	22.1	85.2	93.2
Point-M2AE	3D	15.3	86.4	93.4
P2P-HorNet	2D	1.2	89.3	94.0†
ACT	3D+2D	22.1	88.2	93.7
w/ IDPT	3D+2D	1.7	87.7	94.0†
I2P-MAE	3D+2D	12.9	90.1	93.7
ReCon	3D+2D+Language	43.6	90.6	94.1
Any2Point	Audio	**0.8**	**87.0**	**92.7**
	2D	**0.8**	**87.7**	**93.2**
	Language	**0.9**	**91.9**	**94.3**

Visualization

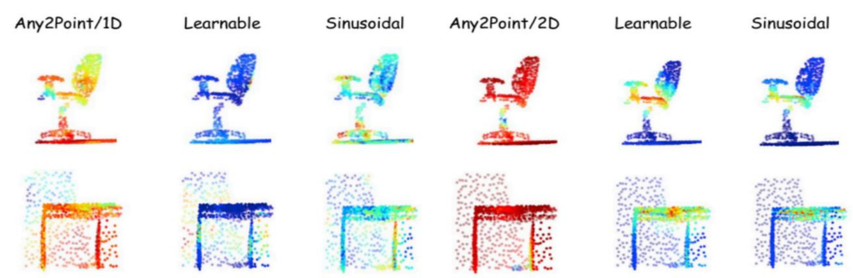


Fig. 5: Visualization of Different Positional Encoding Methods. For the 1D/2D modalities, we visualize the attention scores of the **[CLS]** token to other point cloud tokens, utilizing sinusoidal positional encoding, learnable positional encoding, and 3D-to-any Virtual Projection. **The red color indicates higher values.**

Conclusion

Our extensive experiments across various tasks demonstrate the superior performance and efficiency of Any2Point compared to previous SOTA 3D pre-trained models, achieving remarkable results with only a fraction of the trainable parameters.