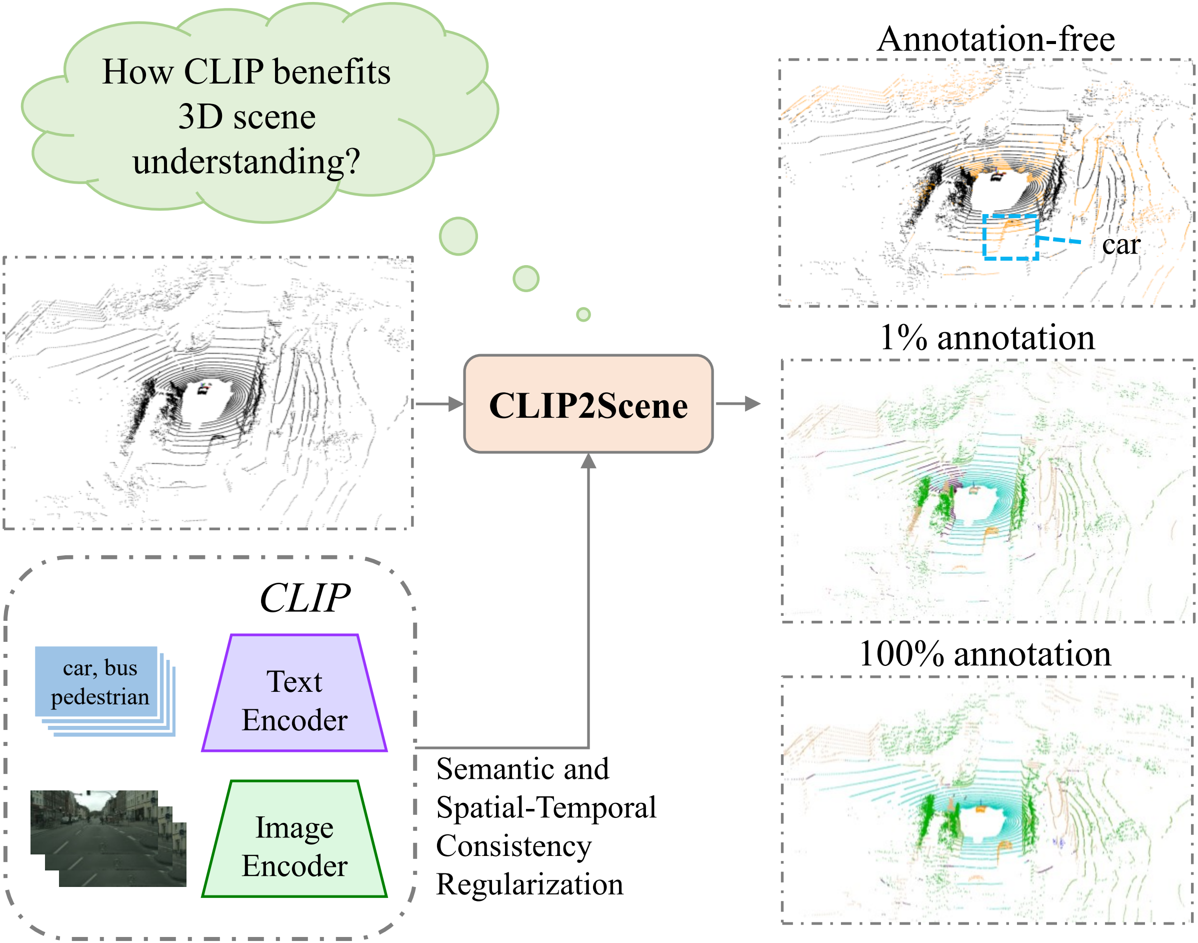
Semantic-driven Cross-modal Contrastive Learning


How CLIP benefits 3D scene understanding?
CLIP2Scene
Annotation-free
car
1% annotation
100% annotation
CLIP
car, bus
pedestrian
Text Encoder
Image Encoder
Semantic and Spatial-Temporal Consistency Regularization