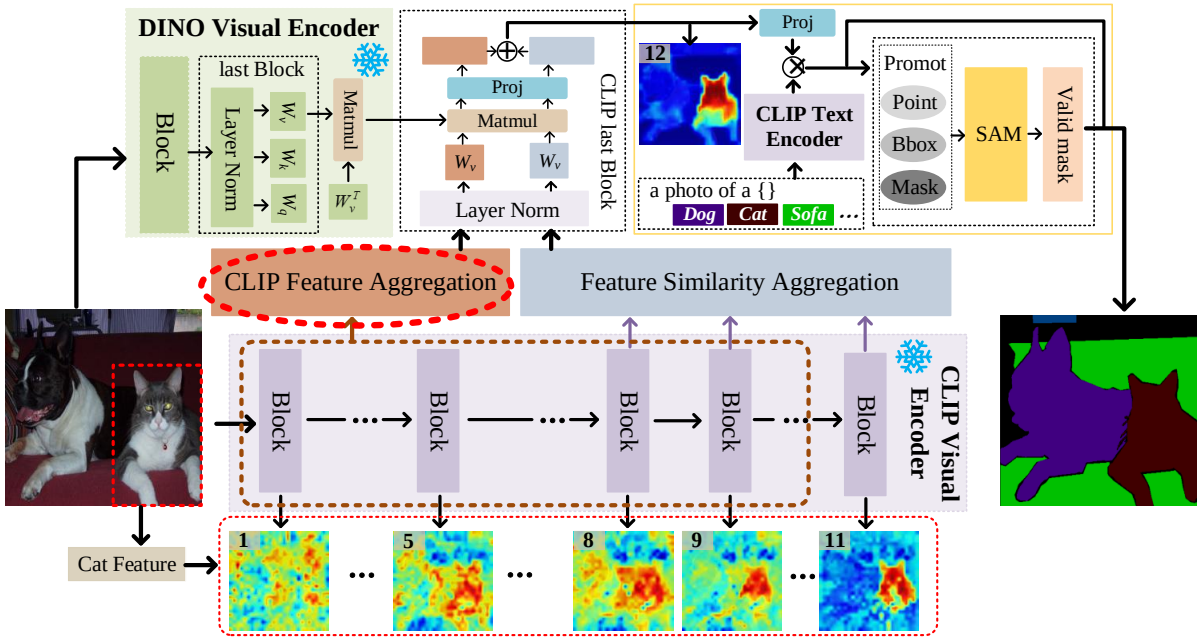

Figure 1. Overall architecture of the proposed S2CLIP. The framework consists of hierarchical progressive feature extraction from CLIP, feature fusion between CLIP and DINO, and SAM post-processing stage.

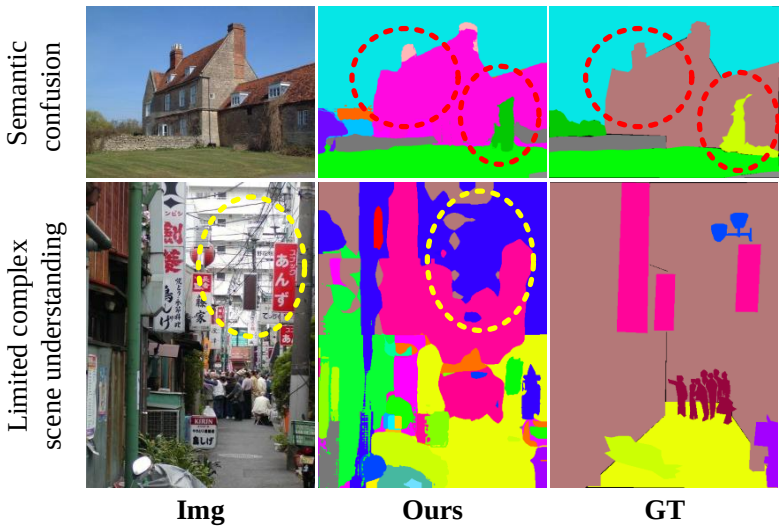

Figure 2. Failure Cases of S2CLIP. The upper section demonstrates semantic category confusion in S2CLIP, while the lower section illustrates that S2CLIP's understanding of extremely complex scenes requires further improvement. These limitations stem from two primary factors: (1) S2CLIP's use of simple CLIP text templates fails to adequately capture subtle semantic variations and state descriptions of objects in complex scenes, and (2) S2CLIP lacks the capability to explicitly model inter-object relationships.