Most scene graph parsers use a two-stage pipeline to detect visual relationships: the first stage detects entities, and the second predicts the predicate for each