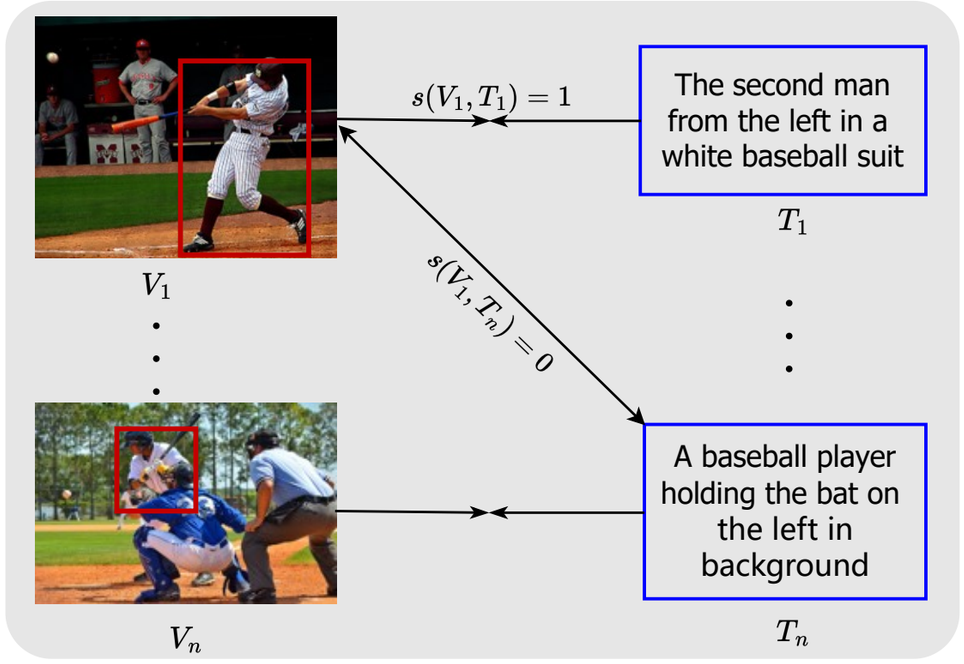
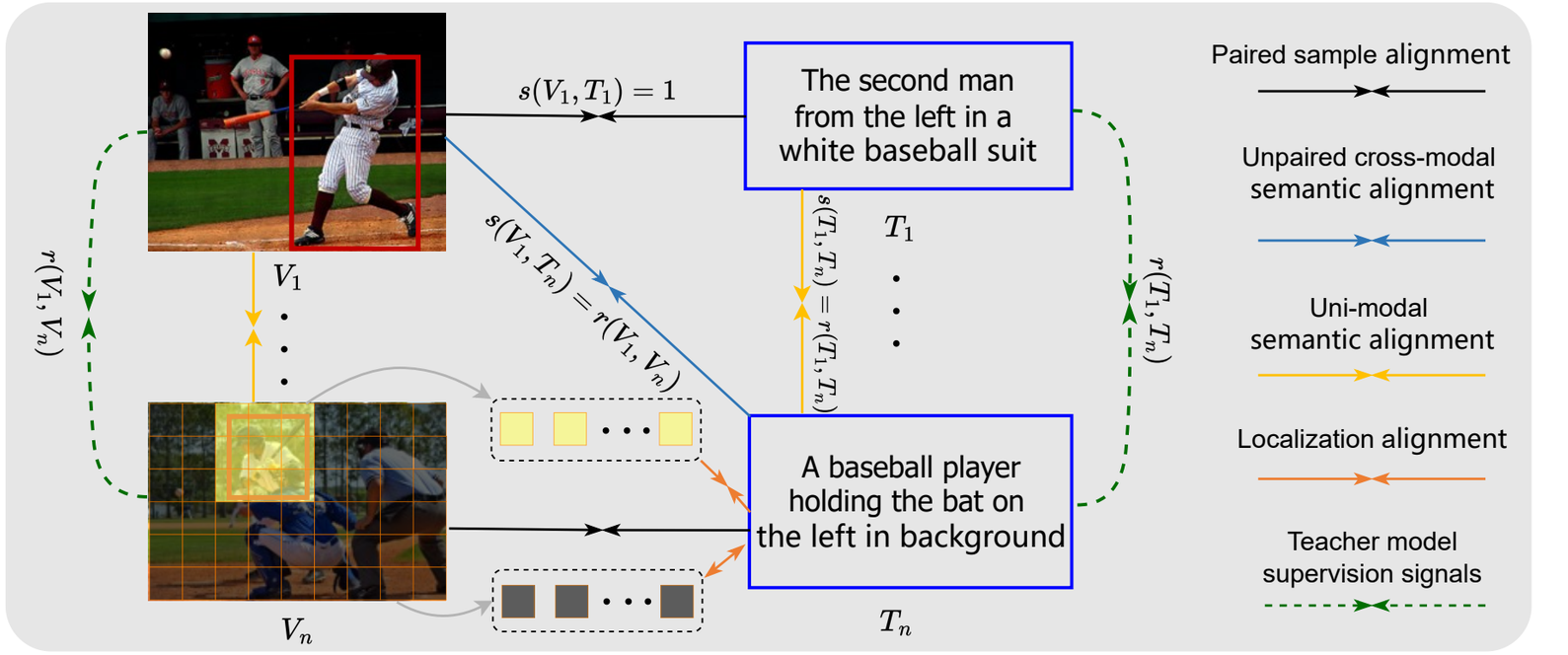
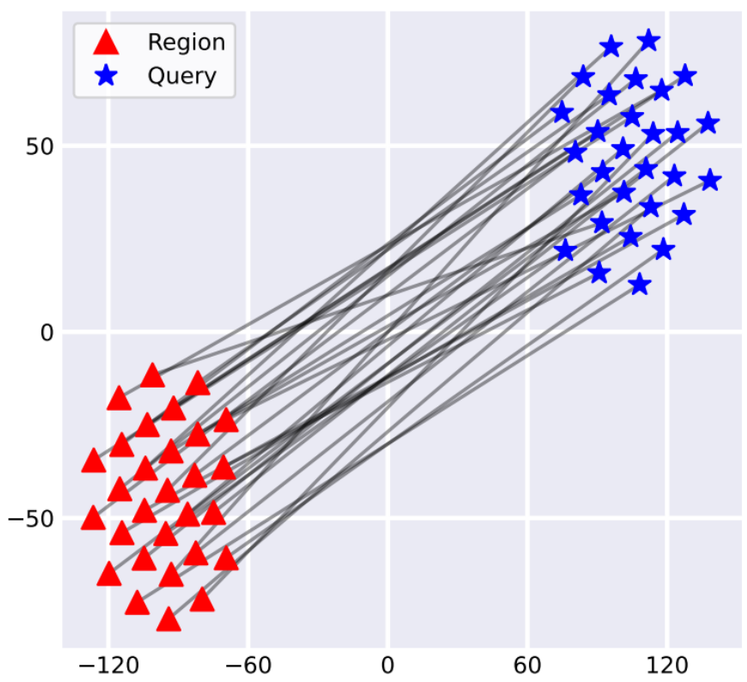


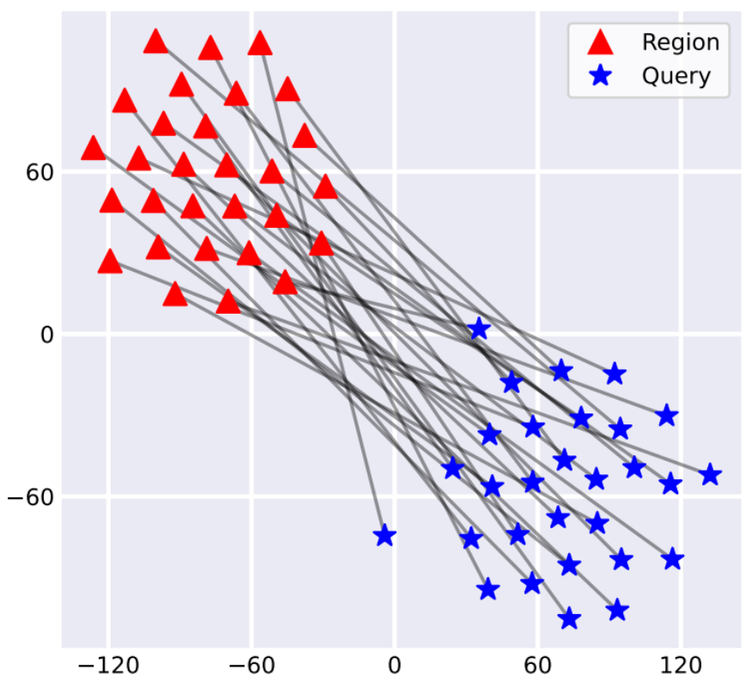
(a) Paired sample alignment



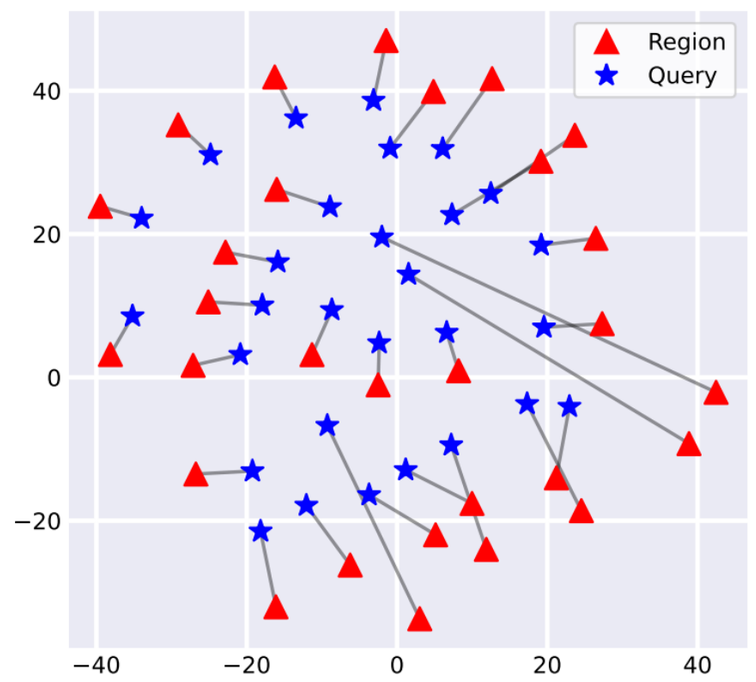
(b) Semantic alignment and localization alignment



(c) QRNet



(d) QRNet w/ CLIP



(e) QRNet w/ SALA (ours)