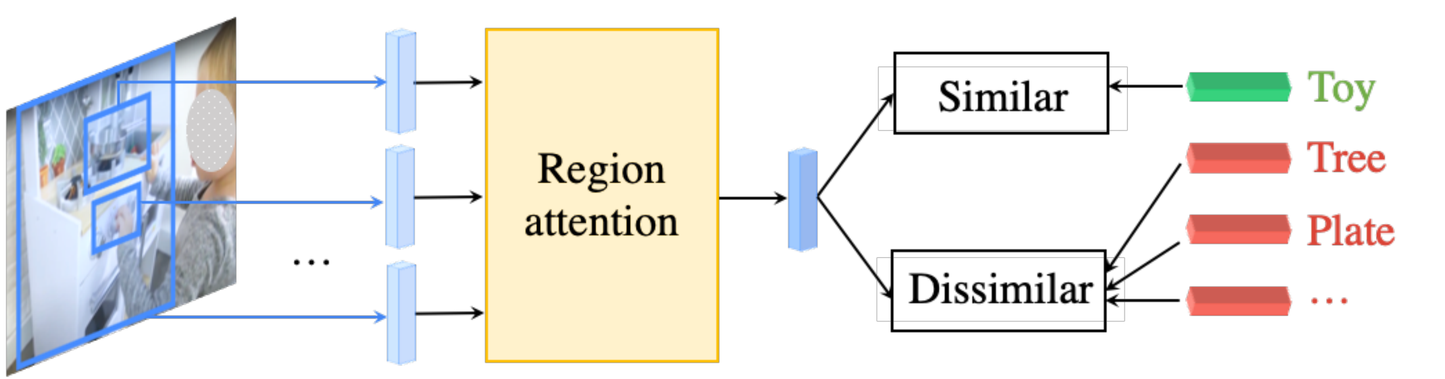
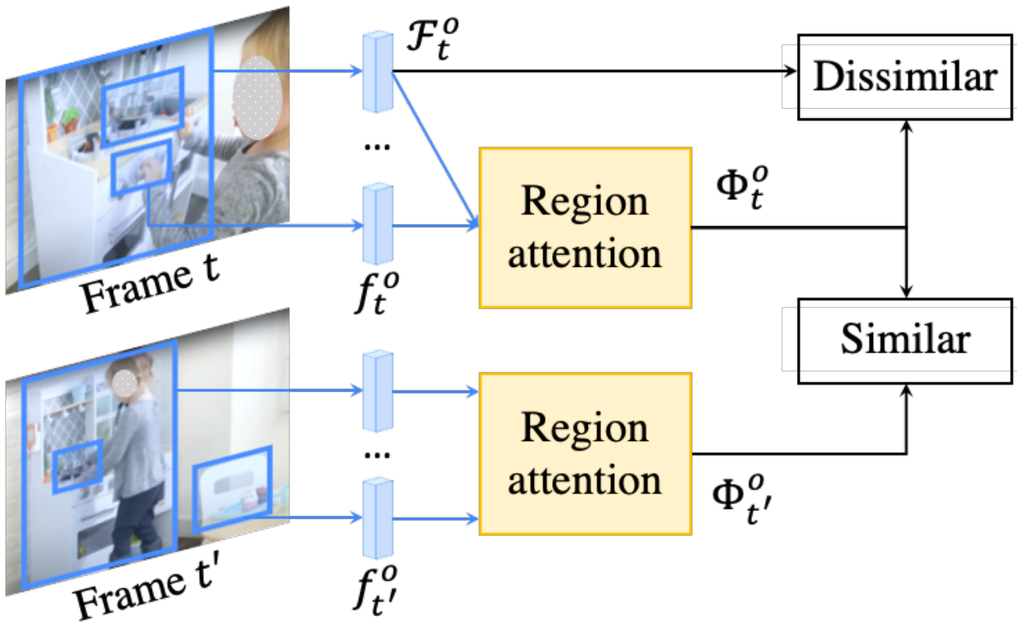


(a) Weakly supervised language-embedding alignment loss



(b) Self-supervised temporal contrastive loss