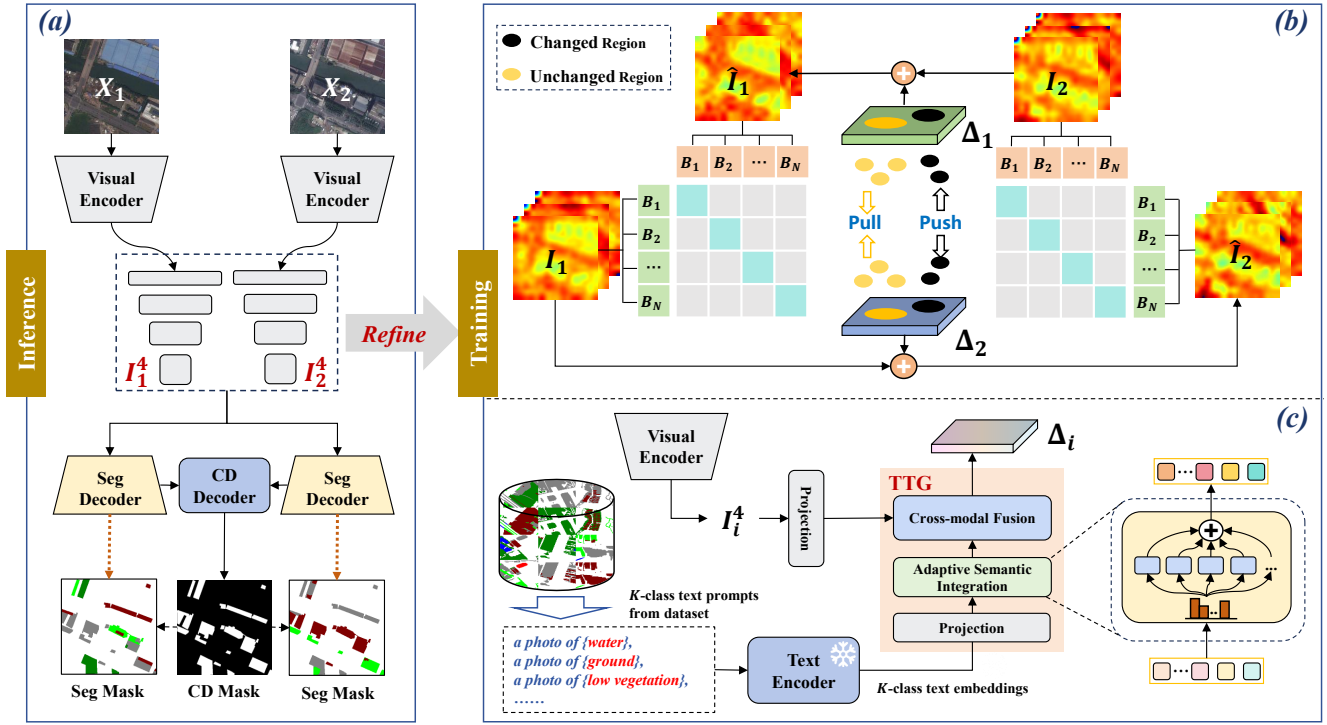

Figure 2. Overview of the proposed TaCo. (a) Structure and inference pipeline based on the siamese encoder–decoder. (b) Spatio-temporal semantic joint constraint on high-level features via reconstruction and transition losses. (c) Text-guided Transition Generator that fuses class-level text embeddings with stage-4 visual tokens to construct transition features Δ_i.