Masked Cross-Modal Modeling - Generative

Combines MIM and MLM by jointly masking both image patches and text tokens, and reconstructing them conditioned on the unmasked parts.

Loss Formulation

$$\mathcal{L}_{\text{MCM}} = -\frac{1}{B}\sum_{i=1}^{B}\left[\log f_\theta(x_i^I \mid \hat{x}_i^I, \hat{x}_i^T) + \log f_\phi(x_i^T \mid \hat{x}_i^I, \hat{x}_i^T)\right]$$

where x_i^I, $\hat{x}_i^I$ are masked/unmasked image patches and x_i^T, $\hat{x}_i^T$ are masked/unmasked tokens.