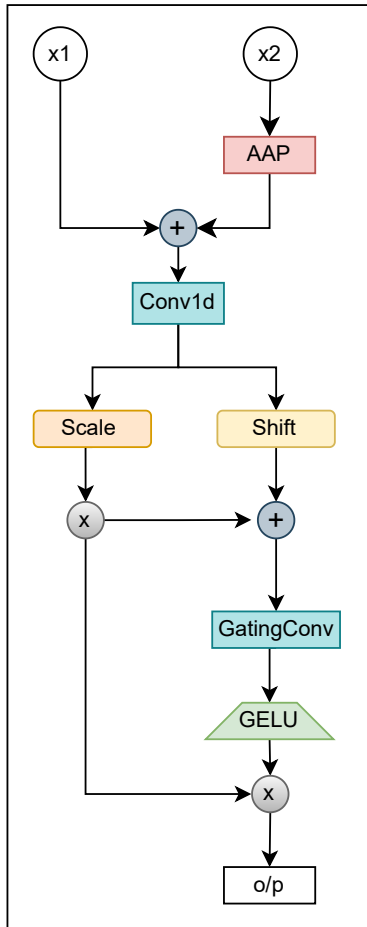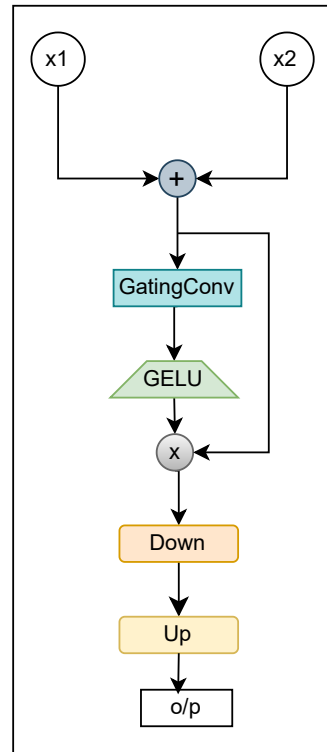**Figure X: Illustration of the proposed Hierarchical Cache-Modulated Attention (HCMA) model. (a) Overall architecture of theproposed HCMA, showing the multi-stage encoder-decoder pipeline for crowd flow prediction. (b) Internal structure of the Hierarchical Cache-Modulated Attention (HCMA) block, detailing cache-driven modulation, head-wise attention, and hierarchical grouping mechanisms.**

Inter Module

Intra Module