2nd Cross-Attention Block

Mask Token → Token-to-Image Attention → Image-to-Token Attention → Token-to-Image Attention

Confidence Map

$exp($ ⬤ $\times \boldsymbol{\beta})$

Attention Matrix ⊗

Attention Matrix ⊗