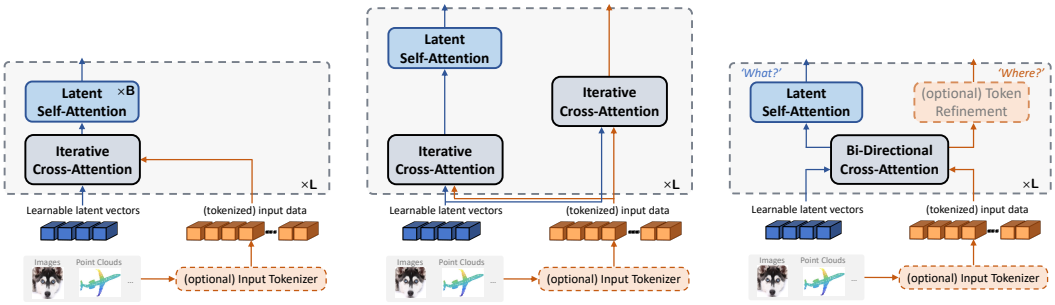
BiXT – Rebuttal Visuals

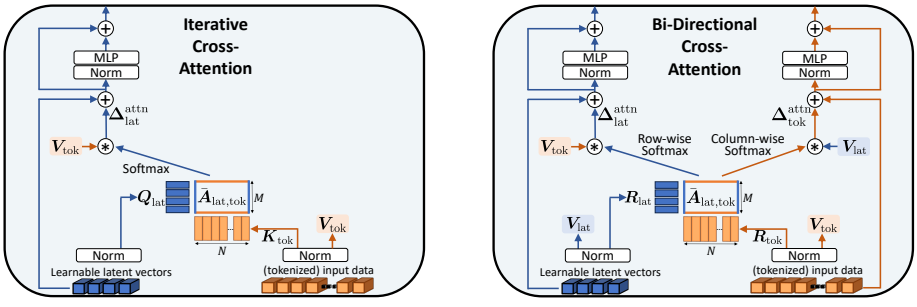
Anonymous Author(s)
Affiliation
Address
email

1 Figures



(a) iterative: $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ (b) sequential: $2 \times \boldsymbol{A} \in \mathbb{R}^{M \times N}$ (c) bi-directional: $\boldsymbol{A} \in \mathbb{R}^{M \times N}$

Figure 1: **Transitioning from iterative to bi-directional attention.** (a) Perceiver-like iterative attention, creating a bottleneck and small effective working memory; (b) Naïve sequential attention 'unblocking' the bottleneck and extending working memory, but still markedly less efficient than: (c) Bi-directional cross-attention used in BiXT, combining efficient linear scaling with competitive performance across various tasks. Note that iterative attention attends to the (unrefined) input at every layer, while sequential and bi-directional attend to variants of the input refined by the previous layer. The Perceiver-like setup additionally uses multiple self-attention layers to refine between each iterative cross-attention ($\times B$) in each architectural layer, whereas sequential and bi-directional variants only use *one* self-attention operation per architectural layer. Architectures are then built by stacking L layers.



(a) iterative attention (b) sequential attention

Figure 2: **Detailed structure of attention blocks.** (a) Perceiver-like iterative attention, creating a bottleneck and small effective working memory; (b) Bi-directional cross-attention used in BiXT.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.