

PAB

- 实时视频生成中的金字塔注意力广播
- <https://arxiv.org/abs/2408.12588>
- ICLR 2025

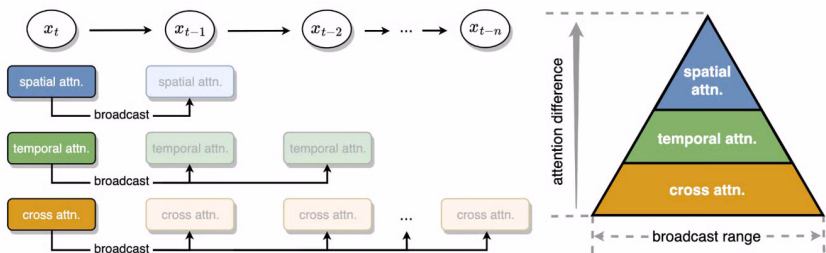


Figure 5: Overview of Pyramid Attention Broadcast. Our method (shown on the right side) which sets different broadcast ranges for three attentions based on their differences. The smaller the variation in attention, the longer the broadcast range. During runtime, we broadcast attention results to the next several steps (shown on the left side) to avoid redundant attention computations.

attention
attention
PAB

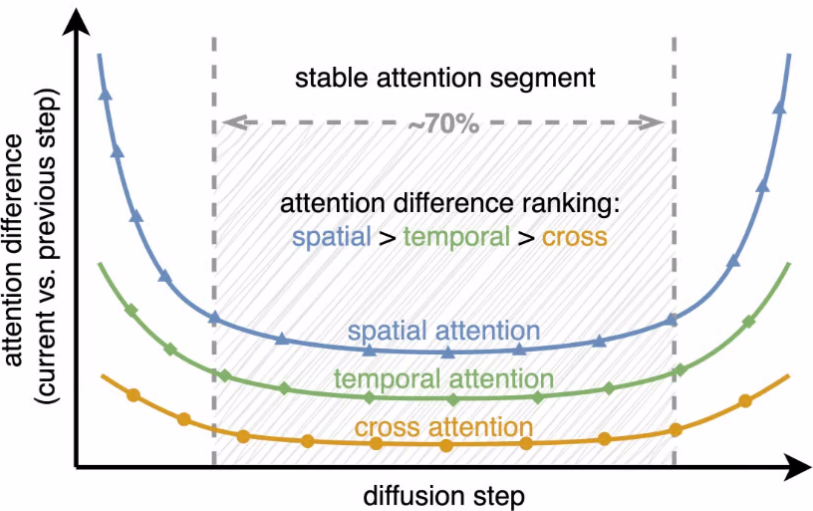


Figure 2: Comparison of the attention outputs differences between the current and previous diffusion steps. Differences are measured by Mean Square Error (MSE) and averaged across all layers for each diffusion step.

- training-free
-
-
- <https://github.com/NUS-HPC-AI-Lab/VideoSys>

Newer	Older
2025-11-04 FasterCache	2025-10-29 SVG