1 APPENDIX

1.1 MODIFICATIONS TO SPATIAL EXPERTS

As spatial experts are trained on independent images, directly applying them to video denoising lacks inter-frame constraints. This can lead to inconsistency and flickering between video frames, increasing the burden on temporal experts. To make spatial experts more suitable for the video tasks, we have modified their attention mechanism. Specifically, we use the key and value of the first frame in each refinement segment to replace the keys and values of all other frames, thereby enhancing content consistency within the segment.

$$\text{Attention}_i = \text{Softmax}\left(\frac{Q_i {K_0}^T}{\sqrt{d}}\right) \cdot V_0 \tag{1}$$

where Q_i denotes the query of frame i in one refinement segment, and K_0, V_0 represent the key and value of frame 0, respectively.

1.2 SLIDING WINDOW TEMPORAL ATTENTION IN CONTROL EXPERTS

In ConFiner-Long, coherent transition between video segments is primarily achieved through staggered refinement. However, in some cases, control expert may generate two frames with significant differences at the junction of adjacent structures. It greatly increases the burden on the temporal expert during the refinement stage, resulting in videos that exhibits unreasonable changes. To alleviate the burden on the temporal expert and improve the overall coherence of the video, we incorporate a sliding window into the temporal attention for the control expert during structure generation. (Only the temporal attention utilizes the sliding window, while other parts are computed normally.)

In our ConFiner-Long setup, the length of a single structure is 16 frames, and we set the sliding window size to 16 frames with a stride of 8 frames. The computation process for a single window in the sliding window attention is as follows:

$$\text{Attn}^j_{i:i+16} = \text{Softmax}\left(\frac{Q_{i:i+16} K^T_{i:i+16}}{\sqrt{d}}\right) V_{i:i+16} \tag{2}$$

where $\text{Attn}^j_{i:i+16}$ represents the attention values of window j.

After applying sliding window attention, each frame corresponds to more than one attention value. Therefore, we need to perform weighted fusion on the attention values from different sliding windows to obtain the final attention value. The fusion process is as follows:

$$\text{Attn}_i = \sum_j \text{Attn}^j_i \cdot \left(\frac{8 - \lfloor |i - c^j| \rfloor}{\sum_j (8 - \lfloor |i - c^j| \rfloor)}\right) \tag{3}$$

where Attn_i denotes the final attention value of frame i, j indicates all the windows that include frame i, Attn^j_i stands for the attention value of frame i in window j, c^j refers to the index of the central position of window j, $|.|$ and $\lfloor . \rfloor$ represent taking the absolute value and the floor function, respectively.