We sincerely appreciate the reviewer's recognition of our work and the increased score. Here's our detailed response for simplicity of linear interpolation:

1. Interpolation Strategy: We adopt learnable linear interpolation to generate intermediate features between key frames. Although this approach is straightforward, it performs effectively in our system by creating smooth transitions between visual features. As shown in Figure 4(b)(line 436), the model performs poorly without interpolation, demonstrating the importance of the interpolation operation. In Figure 4(c)(line 446), applying interpolation twice on the features leads to deteriorated results, indicating that repeated interpolation causes information loss from the original features. In Figure 5(a) and (b)(line 890), we compare various interpolation strategies, including constant, linear, and quadratic interpolation, along with different initialization directions (increasing and decreasing). The quadratic interpolation shows no improvement in performance, suggesting that more complex interpolation methods are unnecessary. Moreover, the results indicate that the interpolation effectiveness remains consistent regardless of whether the interpolation values initially increase or decrease, showing its independence from action-related gradient changes. Furthermore, the learnable design enables more flexible and effective feature transitions that can adapt to different actions.

2. Feature Diversity and Integration: In our task, each action corresponds to unique visual features. The interpolation strategy helps create diverse and distinct features during training, which is crucial for model performance, as demonstrated by Super SloMo[1]. The effectiveness of our Integration approach stems from the synergistic combination of multiple components. We employ cross-attention mechanisms similar to IP-Adapter[2] for visual feature fusion, which proves more effective than direct concatenation. Our ablation studies in Table 6 (line 465) demonstrate that the holistic integration of architectural design choices and feature fusion mechanisms leads to strong results.

3. Computational Efficiency: Linear interpolation offers practical advantages in terms of implementation simplicity and computational efficiency, making it a reasonable choice given its role in the broader architecture.

In summary, exploring more sophisticated and effective interpolation strategies is an promising direction for our future research.

[1] Jiang, H., Sun, D., Jampani, V., Yang, M. H., Learned-Miller, E., & Kautz, J. (2018). Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 9000-9008).

[2] Ye, H., Zhang, J., Liu, S., Han, X., & Yang, W. (2023). Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721.