Appendix

1 The Design of Key and Value in Self-Attention and Fine-tuned Parameters

Let [;] denote the concat operation. We consider using these embeddings as key and value: (1) v^i: original spatial self-attention in T2I models. (2) v^k, which is our key-frame attention. We select four different key frames. (3) $[v^m; v^i]$, where $m = Round(\frac{N}{2})$. (4) $[v^1; v^{i-1}]$. (5) $[v^1; v^{i-1}; v^{i+1}]$, which includes bi-directional information. (6) $[v^1; v^i; v^{i-1}; v^{i+1}]$. As shown in Figure 1, key-frame attention shows the highest temporal consistency, implying that utilizing a key frame to propagate throughout videos is useful. As shown in Table 1, selecting a key-frame as key and value achieves high temporal consistency performance, which is consistent with the qualitative results. There is no significant difference in different key frame selections. In addition, adding the current frame features v^i shows less temporal inconsistency because the v^i contains different information between frames. For example, the color of the car turned red in $[v^m; v^i], [v^1; v^i; v^{i-1}; v^{i+1}]$ following v^i (column 2).

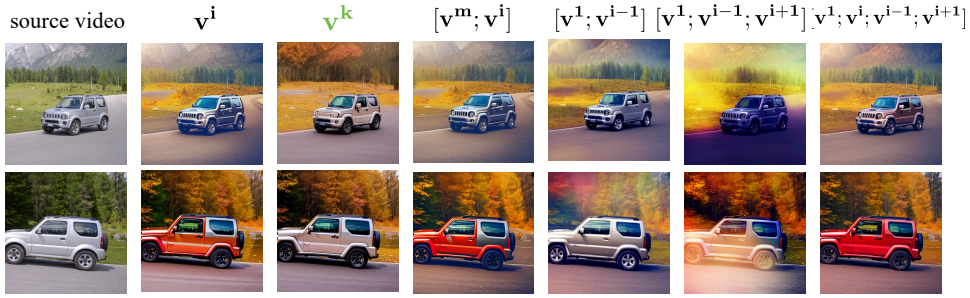

Figure 1 Comparisons with different designs of key and value in self-attention. The green color marked our choice.

Table 1 Quantitative results about different choices of key and value in self-attention.

Method	CLIP-text↑	CLIP-temp↑	SSIM ↑
v^i	0.263	0.905	0.635
$[v^m; v^i]$	0.260	0.939	0.642
$[v^1; v^{i-1}]$	0.264	0.953	0.639
$[v^1; v^{i-1}; v^{i+1}]$	0.261	0.941	0.637
$[v^1; v^i; v^{i-1}; v^{i+1}]$	0.261	0.955	0.648
$v^k, k = 1$	0.263	0.954	0.655
$v^k, k = 3$	0.263	**0.961**	0.654
$v^k, k = 5$	0.261	0.958	**0.657**
$v^k, k = 7$	0.260	0.958	0.650

2 The Way to Initialization and the Incorporation of Local and Global Positions for Introducing Temporal Attention

As shown in Figure 2(a), using pretrain spatial self-attention weights as initialization achieves better performance. Next, we explore following potential locations to incorporate temporal attention in transformer blocks: (1) before self-attention. (2) with self-attention. (3) after self-attention. (4)after cross-attention. (5) after FNN. As shown in Figure 2(b), before self-attention and with self-attention result in the best temporal consistency. This is because the input of these two locations is the same as spatial self-attention, which serves as the initial weight of temporal attention. Notably, with self-attention shows higher text

Figure 2 Ablation studies of (a) the way to initialize and the incorporation of (b) local positions and (c) global positions for introducing temporal attention. The green color marked our choice.

alignment, making it our final choice. Moreover, we find the after FNN location yields the worst temporal consistency and should be avoided.

To investigate the optimal global location for adding temporal attention, we first conduct the following experiments: (1) ControlNet+UNet. (2) ControlNet. (3) UNet. (4) Encoder of UNet. (5) Decoder of UNet. As shown in Figure 2(c), incorporating temporal attention only to the ControlNet fails to preserve the background and removing it does not decrease performance (all vs UNet). This suggests that ControlNet only extracts condition-related features (e.g. pose) and discards the other features (e.g. background), while U-Net, which is used for generation task, preserves all image information. As such, we ultimately choose to add temporal attention to UNet. Additionally, the decoder location achieves better performance than the encoder. This may be because, in U-Net, the decoder contains more information than the encoder by using skip connections to incorporate features from the encoder. Next, we investigate the location in UNet by following experiments: (1) all; (2) Block 1,2; (3) Block 1,3; (4) Block 2,3; (5) Block 1,2,3, which is UNet except middle block. As shown in Figure 2(c), the Block 1,2,3 shows similar performance with all while with less parameters, which is chosen as the final design.

Table 2 Ablation studies for key components in ControlVideo.

Method	CLIP-text↑	CLIP-temp↑	SSIM ↑
Stable Diffusion	0.282	0.898	0.608
w/o temporal attention	0.264	0.936	0.724
w/o key-frame attention	0.262	0.922	0.712
w/o control	0.276	0.956	0.642
Our full version	0.266	0.964	0.738

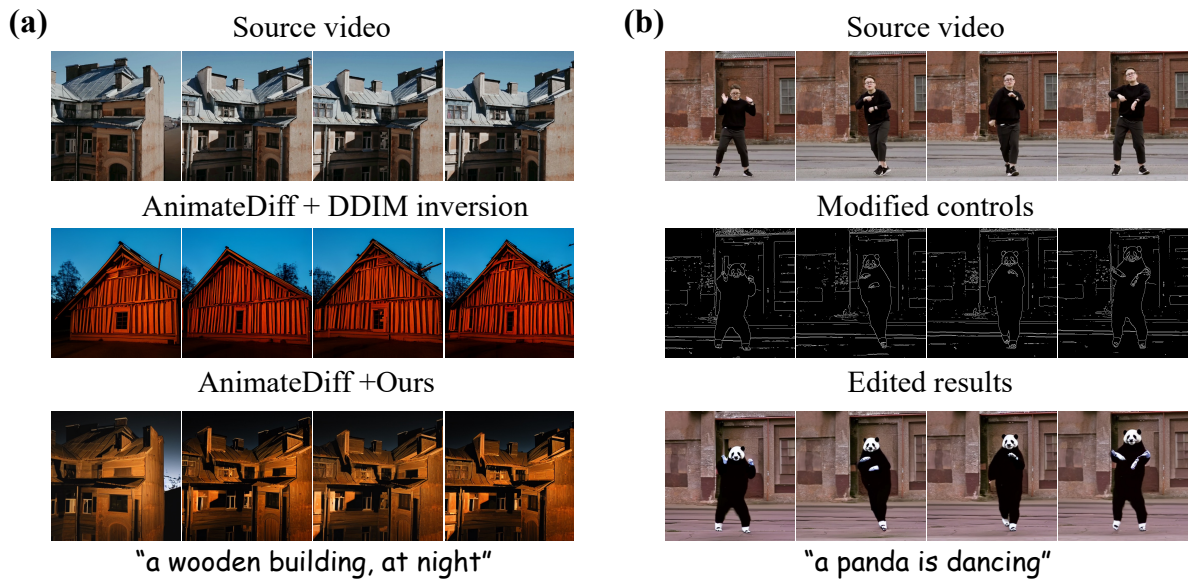

Figure 3 (a) Application of our method to text-to-video models. Compared to the baseline approach of directly using DDIM inversion for AnimateDiff [1], our designs in AnimateDiff more faithfully preserve the source content. (b) Application on the structural editing scenario by modifying the controls.

3 Ablation studies for Key Components in Controlvideo

As shown in Table 2, the quantitative results are consistent with the qualitative results in the main text. We can observe that introducing additional control mainly contributes to faithfulness a lot. The key-frame attention and temporal attention mainly contribute to temporal consistency and faithfulness.

References

1 Guo Y, Yang C, Rao A, et al. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: Proceedings of International Conference on Learning Representations, 2024