From Prompt to Progression: Taming Video Diffusion Models for Seamless Attribute Transition

Supplementary Material

7. Implementation Details

During sampling, the classifier-free guidance scale ω is set to 12, and the scaling factor α is configured to [0,1], ensuring the attribute transition scale ranges from -1 to 1 across frames in the video.

For quantitative evaluation on CAT-Bench, we generate five videos per prompt, yielding a total of 600 videos. Scores are averaged across all generated videos, and the default resolution settings of each model are used to ensure consistency. Similarly, for TC-Bench-T2V, we generate five videos per prompt and report the TCR and TC-Score using GPT-4 [1] for assertion-based evaluations. For competing methods, we reference the scores reported in the original paper [7]. All models can be run on a single 24 GB NVIDIA RTX A5000.

For the user study, we combined our generated videos with those produced by three other baseline methods using the same backbone, ensuring the order was randomized to eliminate bias. A total of 38 participants were asked to evaluate the videos and select the best one based on five criteria: Transition Correctness, Transition Smoothness, Video Consistency, Motion Accuracy, and Overall Performance.

8. Details of CAT-Bench

We define 8 categories of attributes for transitions: four focused on human characteristics (age, beard, makeup, and hair), two on subject attributes (color and material), and two on background features (light conditions and weather). Each category includes 15 samples, resulting in a diverse dataset. For a comprehensive evaluation, each sample includes a prompt pair for multi-prompt generation methods and a single prompt for single-prompt generation. We ensure that each sample incorporates both an attribute transition and consistent motion, allowing the evaluation to cover not only the successful completion of attribute transitions but also the preservation of motion dynamics. This ensures that the generated videos are not merely interpolations between static images without motion.

To construct the benchmark, we first define the prompt pairs and then use GPT-4 to generate a single descriptive prompt that encapsulates the attribute transition within a single prompt and facilitates fair and consistent evaluation across different methods. Table 5 presents some examples from our CAT-Bench.

9. Discussion on Video Quality

We employ metrics from VBench [15] to evaluate overall video quality across five key dimensions. First, we assess *Consistency*. For human and object attribute transitions, we use background consistency to ensure that elements aside from the attribute remain stable across frames, while for background attribute transitions, we employ subject consistency. Next, we evaluate *Temporal Flickering* to measure the overall temporal stability of the video. We then assess *Motion Smoothness* to ensure that the generated movements are fluid and natural. As a complete static video could score well on previous metrics, we use *Dynamic Degree*, which checks whether the video has sufficient dynamism. Finally, we evaluate *Imaging Quality* on a frame-by-frame basis to ensure high visual fidelity throughout the video. Table 6 presents the evaluation results for video quality. The proposed approach achieves comparable scores to the backbone model, VideoCrafter2, across metrics such as Consistency, Imaging Quality, Temporal Flickering, and Motion Smoothness, indicating that our method maintains video quality while generating attribute transitions. Additionally, our approach achieves the highest score in Dynamic Degree among approaches using VideoCrafter2 as the backbone, demonstrating that the motion described in the prompts is effectively captured. Although AnimateDiff attains the highest Dynamic Degree, its low Motion Smoothness suggests that the high Dynamic score results from frame inconsistencies rather than genuine motion dynamics. In contrast, our method strikes a balance, ensuring both smooth motion with fluidity.

10. Discussion on Scale Factor α

Unlike previous approaches that rely on prompt interpolation or leverage LLMs to define intermediate attribute states, our method not only facilitates attribute transitions but also enables control over the intensity of these transitions through the scaling factor, α. As illustrated in Figure 8, setting α to [0,1], corresponding to a transition range from -1 to 1 across the video, results in a well-balanced progression from sunny to rainy. On the other hand, increasing α to [0,2], representing a range from -2 to 2, intensifies the transition. The rain in the final frame becomes heavier, while the sunlight in the initial frame shines more brightly, amplifying the visual impact of the attribute change. The flexibility highlights the adaptability of our approach in controlling transition magnitude to suit different scenarios.

Table 5. Examples of the Prompts in Our CAT-Bench. The underline highlights the transitioning attribute in the prompt pair for multi-prompt generation.

Multi-Prompt	Single-Prompt
"A young man is rowing a boat" → *"An elderly man is rowing a boat"*	*"A man is rowing a boat, gradually aging from young to old over time."*
"A woman without makeup is laughing at the party. " → *"A woman with makeup is laughing at the party. "*	*"A woman is laughing at the party, gradually transforming from having no makeup to wearing makeup."*
"A man without a beard is applauding. " → *"A man with a beard is applauding. "*	*"A man is applauding, gradually transitioning from being clean-shaven to having a beard."*
"A woman with short hair is riding a horse." → *"A woman with long hair is riding a horse."*	*"A woman is riding a horse, with her hair gradually transitioning from short to long."*
"A white dog is running in the field." → *"A gray dog is running in the field."*	*"A dog is running in the field, gradually transitioning from white to gray."*
"A knit shirt is floating in the wind." → *"A silk shirt is floating in the wind."*	*"A shirt is floating in the wind, gradually transitioning from knit to silk."*
"A boat is drifting on the river in a dark light." → *"A boat is drifting on the river in a bright light."*	*"A boat is drifting on the river as the light transitions gradually from dark to bright."*
"A hot air balloon is flying on a sunny day." → *"A hot air balloon is flying on a cloudy day."*	*"A hot air balloon is flying across the sky as the weather transitions from sunny to cloudy."*

Table 6. Quantitative Evaluation for Overall Video Quality. The highest scores are highlighted in blue and the second-highest scores are highlighted in magenta.

	Consistency↑	Imaging Quality↑	Temporal Flickering↑	Motion Smoothness↑	Dynamic Degree↑
AnimateDiff	0.8044	0.4848	0.7007	0.7309	**1.0000**
ModelScope	0.9204	0.6075	0.9528	0.9651	0.5083
Latte	0.9178	0.5600	0.9352	0.9523	0.7808
Free-Bloom	0.8422	**0.7086**	0.8987	0.9091	0.4833
VideoTetris	0.8905	0.5670	0.9337	0.9535	0.6999
VideoCrafter2	**0.9639**	0.6549	**0.9595**	0.9774	0.4750
Gen-L	0.9527	0.6429	0.9525	0.9776	0.5749
FreeNoise	0.9556	0.6628	0.9547	0.9748	0.3888
Ours	0.9537	0.6648	0.9573	**0.9780**	0.7083

While our method performs well with different α values, excessively large values can still lead to the distance effect. For instance, when α is set to [0,5], the video begins to exhibit significant artifacts and distortions, which occurs because driving the latent too far from the decision boundary leads to unnatural changes and inconsistencies, as mentioned earlier in the main paper.

11. Discussion on Compatibility

We demonstrate the compatibility of our method by applying it to OpenSora [28], a transformer-based model. As shown in Table 7, we evaluate both OpenSora and our method, using OpenSora as the base model, on CAT-Bench. Our evaluation considers both attribute transition quality and overall video quality. The results indicate that our approach successfully enables OpenSora to generate smooth attribute transitions while maintaining high video quality. This experiment highlights the flexibility of our method, demonstrating its effectiveness even with transformer-based architectures.

Table 7. evaluation on compatibility

	OpenSora	Ours
Motion Smoothness ↑	0.9723	**0.9809**
Dynamic Degree ↑	0.4622	**0.7194**
Temporal Flickering↑	0.9542	**0.9566**
Imaging Quality ↑	0.6632	**0.6751**
Background Consistency ↑	**0.9701**	0.9602
Wholistic Transition Score ↑	0.0035	**0.1622**
Frame-wise Transition Score ↑	0.0002	**0.0193**

Figure 8. Effect of Scaling Factor (α) on Attribute Transitions. Increasing α intensifies the transition (e.g., heavier rain, brighter sun), but excessively large values (e.g., $\alpha = 5$) introduce artifacts.

Figure 9. Video quality comparison with reduced sampling steps. Despite using only 80% of the sampling steps, the generated videos maintain similar visual quality and smooth attribute transitions.

12. Discussion on Inference Time

We evaluate the inference time of our method and compare it to the base model, as shown in Table 8. Our method requires inference time slightly higher than the base model, which is a reasonable trade-off for enabling smooth attribute transitions. Additionally, by reducing the sampling steps to 80%, we achieve an inference time of 349.23 seconds, which is nearly on par with the base model while still maintaining smooth transitions. Furthermore, Figure 9 illustrates that reducing the sampling steps does not significantly degrade video quality, demonstrating that our approach remains effective even with fewer denoising iterations.

13. Additional Results

We present additional qualitative evaluation examples in Figures 10-13 to further demonstrate the effectiveness of our method in generating smooth and accurate attribute transitions compared to baseline approaches. Additionally, we include video examples as part of the supplementary material for a more comprehensive evaluation of the performance of our approach.

Table 8. Inference time comparison between the base model and our method.

Method	Sampling Steps	Inference Time (sec)
Base Model	100%	336.36
Ours (Full Steps)	100%	453.71
Ours (Reduced Steps)	80%	349.23

14. Longer Video with Attribute Transition

The proposed method is compatible with any baseline model utilizing classifier-free guidance. Additionally, it can seamlessly integrate with other sampling techniques for generating extended videos, enabling the creation of longer and more dynamic video sequences. To demonstrate its versatility, we integrate Temporal Co-Denoising from [25] with our method and generate 64 frames per video. The results in Figure 14 highlight the capability of our method to combine effectively with longer video generation techniques, producing consistent videos with smooth and accurate attribute transitions. We provide video examples alongside the supplementary material document.

Figure 10. More Generation Results for the Specified Attribute Transition.

Figure 11. More Generation Results for the Specified Attribute Transition.

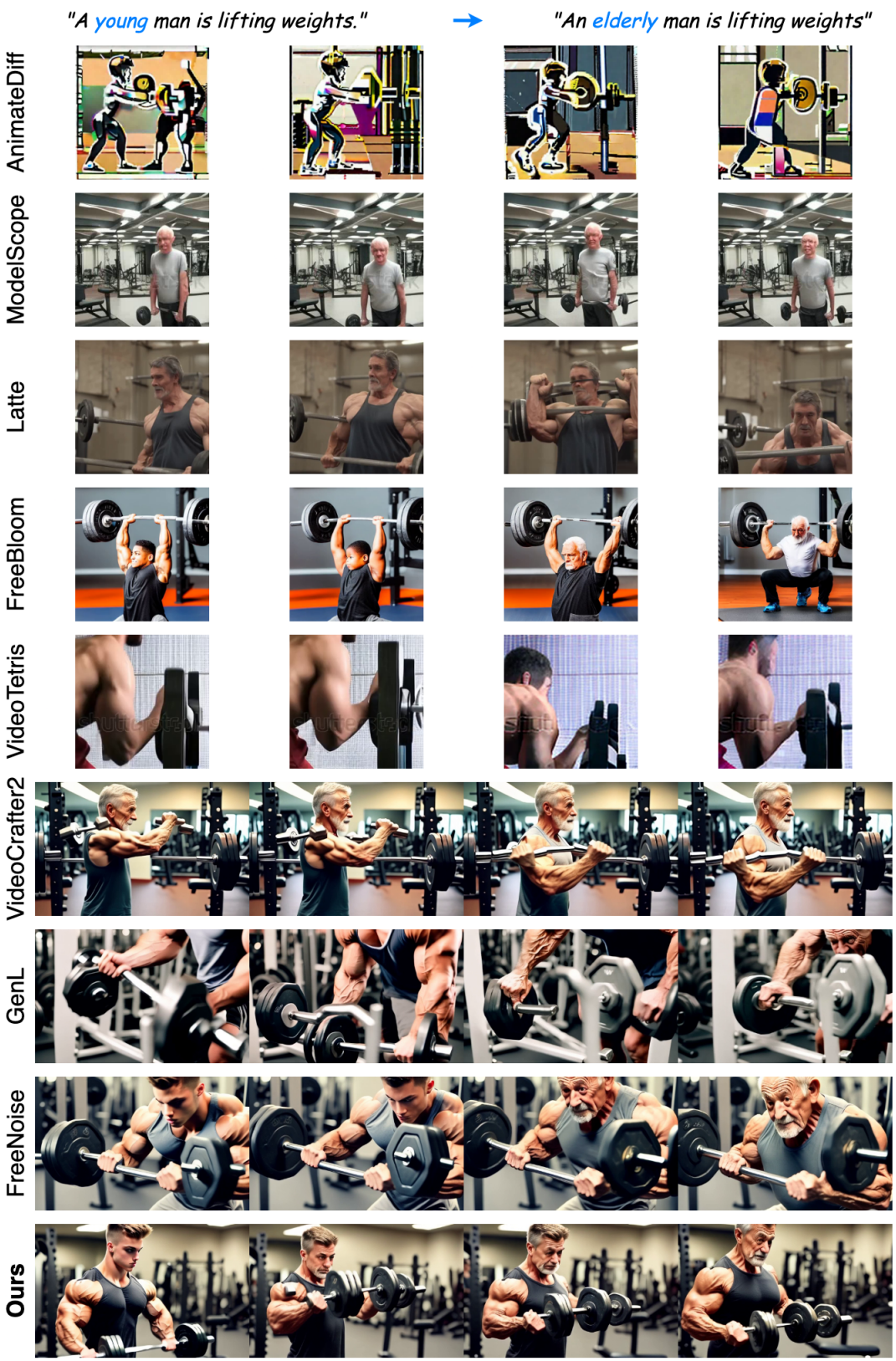

Figure 12. More Generation Results for the Specified Attribute Transition.

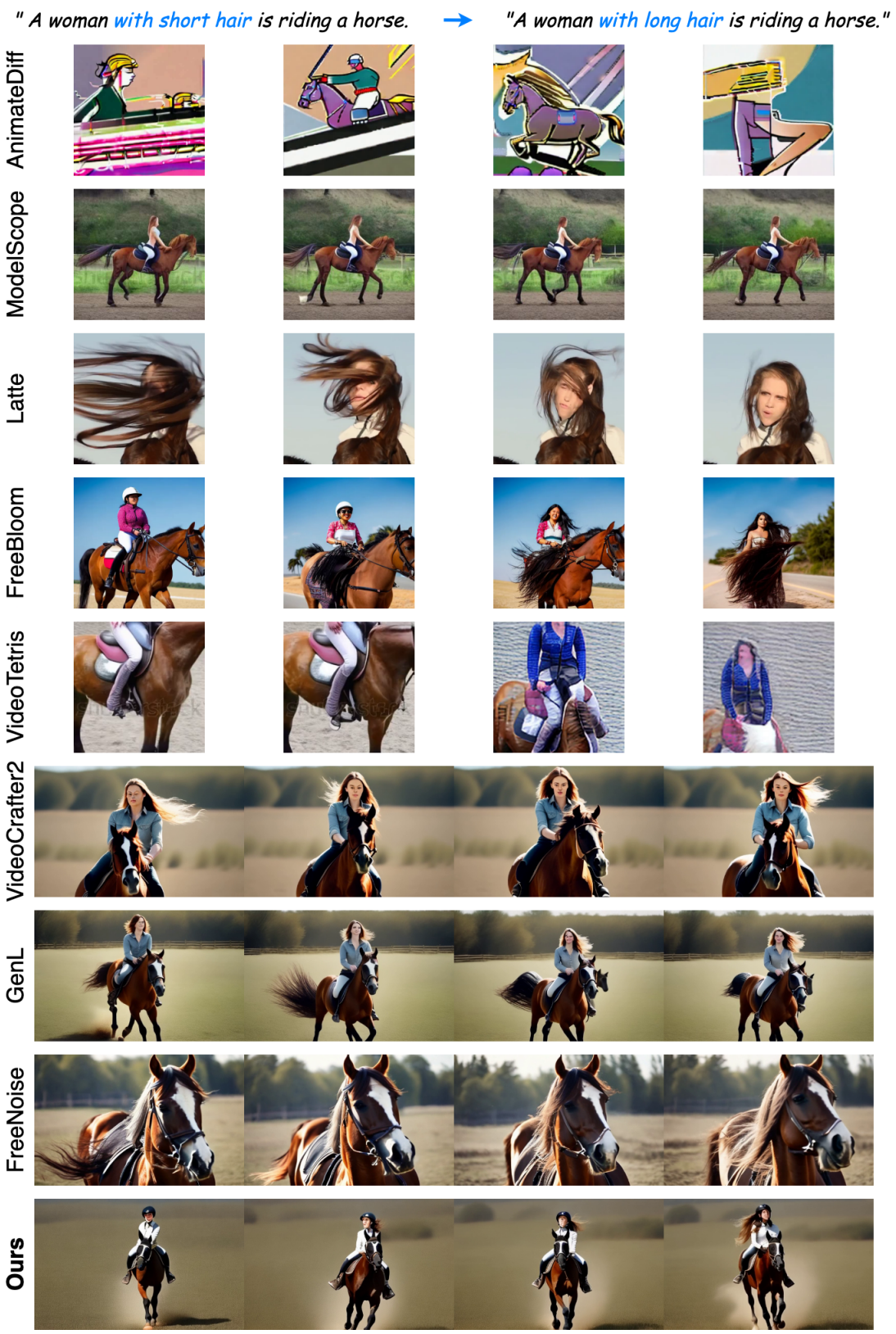

Figure 13. More Generation Results for the Specified Attribute Transition.

"A kite is flying on a sunny day." -> "A kite is flying on a rainy day."

"A hot air balloon is flying in a dark light." -> "A hot air balloon is flying in a bright light."

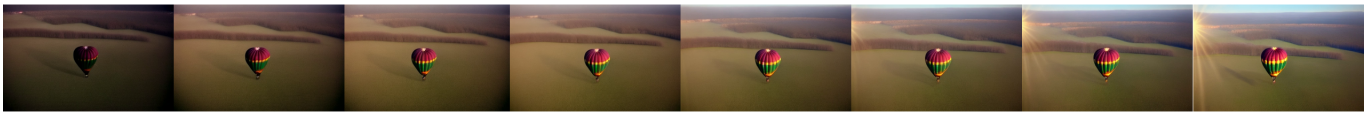

Figure 14. Generation Results of Longer Video.