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

Ling Lo
National Yang Ming Chiao Tung University
linglo.ee08@nycu.edu.tw

Kelvin C.K. Chan
Google DeepMind
kelvinckchan@google.com

Wen-Huang Cheng
National Taiwan University
wenhuang@csie.ntu.edu.tw

Ming-Hsuan Yang
UC Merced
mhyang@ucmerced.edu

Abstract

Existing models often struggle with complex temporal changes, particularly when generating videos with gradual attribute transitions. The most common prompt interpolation approach for motion transitions often fails to handle gradual attribute transitions, where inconsistencies tend to become more pronounced. In this work, we propose a simple yet effective method to extend existing models for smooth and consistent attribute transitions, through introducing frame-wise guidance during the denoising process. Our approach constructs a data-specific transitional direction for each noisy latent, guiding the gradual shift from initial to final attributes frame by frame while preserving the motion dynamics of the video. Moreover, we present the Controlled-Attribute-Transition Benchmark (CAT-Bench), which integrates both attribute and motion dynamics, to comprehensively evaluate the performance of different models. We further propose two metrics to assess the accuracy and smoothness of attribute transitions. Experimental results demonstrate that our approach performs favorably against existing baselines, achieving visual fidelity, maintaining alignment with text prompts, and delivering seamless attribute transitions.

1. Introduction

Diffusion models have made remarkable advancements in text-to-video generation [5, 9, 10, 13, 15, 17, 25, 26]. Compared to image generation, the additional time dimension in videos provides an opportunity to create more dynamic and engaging content. Recent developments in large-scale datasets and architecture designs have led to high-quality, powerful text-to-video (T2V) models that generate videos with remarkable coherence and realism. Despite their success in producing temporally consistent video sequences, these models often struggle with complex variation over the time dimension, such as varying motions and attribute transitions. Real-world videos, on the other hand, are composed of sequences featuring different motions, color changes, varying light conditions, and evolving backgrounds over time. The limitation inevitably restricts the potential of video generation in practical applications, where the ability to handle dynamic changes is essential.

Figure 1. Example of Video Generation with Attribute Transitions Using the Same Base Model. The base model generates static appearances throughout the video. Prompt interpolation leads to inconsistencies, such as abrupt changes in the buildings, while our method ensures smoother and more consistent attribute transitions.

As shown in Figure 1, when attempting to generate a transition from *"A bustling cityscape at midday, bathed in bright sunlight."* to *"A nighttime cityscape under heavy rain."* in a video, the base T2V model fails to produce meaningful attribute changes, resulting in a static appearance throughout the sequence. To address temporal variations, recent studies propose additional networks to facilitate multi-prompt video generation [8, 23], utilizing multiple descriptions that capture the desired transitions over

time. However, they require costly retraining and rely on large, complex text-video datasets. To mitigate training cost, training-free methods [19, 24] are developed, generally employing prompt interpolation to improve transition fluidity. Nevertheless, they are primarily designed for dynamic motions, and the generation of smooth and gradual attribute transitions remains underexplored. Attribute transitions typically involve a more gradual process that unfolds over several frames. Consequently, inconsistencies from these methods become more pronounced, making them less effective for scenarios that demand smooth and seamless attribute changes. As observed in Figure 1, prompt interpolation struggles to maintain object consistency, causing elements like buildings to distort across frames.

In this work, we focus on extending the base model to generate progressive attribute transitions without additional training. Our approach begins by crafting a ***transitional direction*** representing the transition from the initial attribute to the final attribute[1]. Using the crafted transitional direction, we provide frame-wise guidance to the video throughout the denoising process to obtain smooth transitions of the desired attribute states. To stabilize generation, we propose ***neutral prompt anchoring***, which anchors the generation to a neutral prompt while applying the transitional direction. The neutral prompt mitigates inconsistencies in prompt interpolation, where discrepancies between the first and last prompts can lead to abrupt changes, ensuring a seamless transition throughout the video. Our method is training-free and achieves smooth, controlled attribute transitions without sacrificing motion dynamics and video quality.

To evaluate attribute transitions in video generation, we introduce a new benchmark, ***Controlled-Attribute-Transition Benchmark (CAT-Bench)***, designed to measure the smoothness and accuracy of attribute changes throughout the video. Specifically, we design prompts across eight categories, encompassing human, object, and background attribute transitions. Each prompt contains an attribute transition alongside subject motion, creating a complex scene that captures both attribute and motion dynamics. In addition, we develop two metrics: 1) *Wholistic Transition Score* and 2) *Frame-wise Transition Score* to assess the effectiveness and smoothness of attribute transitions in videos. We extensively benchmark multiple baselines, including single-prompt T2V generation and multi-prompt T2V generation. Our experiments show the effectiveness of our method in generating videos that smoothly align with the provided text prompts. The contributions of this work are:

- We propose a simple yet effective approach for generating videos with smooth and gradual attribute transitions. Our training-free method integrates attribute changes without compromising the motion dynamics of the video.
- We introduce a new benchmark designed to evaluate the attribute transition in video generation. In addition, we propose two novel metrics to quantitatively evaluate the smoothness and fidelity of attribute transitions.
- Our experimental results demonstrate the effectiveness of the proposed approach. Moreover, we show that the proposed metrics are well aligned with human perception.

[1]For example, in Figure 1, the initial attribute is *"midday, in bright sunlight"*, and the final attribute is *"nighttime, under heavy rain"*.

2. Related Work

Single-Prompt Video Generation via Diffusion Models. Diffusion models have recently become a leading approach for T2V generation [5, 9, 10, 17, 25]. Building on advances in large-scale text-to-image (T2I) models, ModelScope [25] and VideoCrafter2 [5] incorporate additional spatial-temporal blocks for temporal dependencies. AnimateDiff [9] finetunes a pre-trained T2I model to adapt to motion patterns seamlessly. Latte [17] uses a series of transformer blocks to effectively model video distributions. Although these models can generate high-quality videos from a single prompt, they often struggle with capturing dynamic changes across the temporal dimension.

Multi-Prompt Video Generation via Diffusion Models. Since a single prompt is often insufficient to fully describe a video, Villegas *et al*. introduce Phenaki [23] that takes multiple prompts as input. Although effective, retraining the model requires substantial computational resources and large amounts of data. In contrast, GenL [24] and FreeNoise [19] propose training-free methods that leverage existing models to generate short clips using different prompts, combining them into a video with prompt interpolation for smooth transitions. However, they primarily address dynamic motions and struggle with attribute transitions, which demand a more gradual process that prompt interpolation fails to achieve. Additionally, VideoTetris [22], FreeBloom [14], and MEVG [18] use large language models (LLMs) to describe the video frame by frame. Yet for attribute transition, the subtle changes between frames are difficult to capture accurately with textual descriptions.

Score Estimates Guidance. As diffusion models can be seen as score-based generative models [21], research [6, 12] that manipulates the score estimates has proven effective in guiding the generation process post-training, enabling finer control over the resulting outputs. Liu *et al*. [16] combine multiple estimates to facilitate changes in image composition. Brack *et al*. [3] introduce semantic score guidance that scales monotonically, enabling fine-grain control over subtle semantic changes. However, these efforts have mainly focused on fine-grained control in T2I generation. Unlike static images, attribute transition in video demands continuous attribute changes across frames, presenting unique and complex challenges and remaining under-explored.

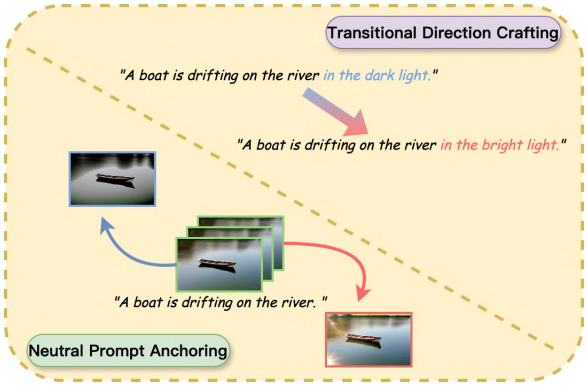


Figure 2. Overview of Our Method. We first create the transitional direction, then guide the video sampled with a neutral prompt frame by frame to achieve smooth attribute transitions while maintaining motion dynamics.

3. Method

Given two descriptive prompts, P_I and P_F, where P_I describes the scene with initial attribute state[2], and P_F specifies the scene with final attribute state[3], our goal is to generate a video that captures this attribute transition while maintaining consistent motion dynamics. We leverage a pre-trained T2V diffusion model and calibrate the sampling process during inference in a training-free manner. Figure 2 illustrates an overview of the proposed method.

The core idea is to first identify a transitional direction in the latent space, and then adjust the sampling process for the video latent, guiding each frame to follow the transition direction in a frame-wise manner. In this section, we start with a brief overview of T2V diffusion models, followed by a detailed description of our proposed method.

3.1. Preliminaries

T2V Diffusion Models. Text-to-video diffusion models aim to map a random noise vector Z_T and a text condition P to an output latent Z_0 through a denoising network ϵ_θ. During training, a forward diffusion process is first applied to transform the initial latent vector $Z_0 \in R^{F\times C\times H\times W}$ into a noisy approximation Z_t over timestep t, where F represents the number of frames in the video, and H, W and C denote the height, width, and channel, respectively. Conditioning on P, the denoising network ϵ_θ is trained to estimate the noise ϵ_t in each time step t within the noisy video latent Z_t by minimizing the objective:

$$\min_\theta E_{Z_0,\epsilon\sim N(0,1),t\sim uniform(1,T)}\|\epsilon_t - \epsilon_\theta(Z_t, P)\|_2^2. \quad (1)$$

Since the objective is noise matching over multiple levels, the learned denoising score $\epsilon_\theta(Z_t, P)$ estimates the gradient of the log-density of the distribution of the noisy data Z_t conditioning on P [21]:

$$\epsilon_\theta(Z_t, P) \approx -\sigma_t \nabla_{Z_t} \log P(Z_t|P), \quad (2)$$

where $P(Z_t|P)$ represent the distribution generated by the diffusion model conditioned on P.

Sampling from Diffusion Model. Classifier-free guidance allows diffusion models to trade off diversity for fidelity [12]. When a conditional and an unconditional diffusion model are jointly trained, the predicted noise during inference can be guided by combining both the conditional and unconditional denoising scores:

$$\tilde{\epsilon}_\theta(Z_t, P) = (1+\omega)\cdot\epsilon_\theta(Z_t, P) - \omega\cdot\epsilon_\theta(Z_t), \quad (3)$$

where ω is the guidance scale that controls the trade-off between mode coverage as well as sample fidelity.

3.2. Transitional Direction Crafting

We aim to generate a video, where each frame shifts progressively from displaying P_I to P_F. First, we identify a transitional direction in the latent space that transitions from the P_I to P_F, which involves increasing the likelihood of P_F and decreasing the likelihood of P_I in Z_t:

$$\frac{P_\theta(P_F|Z_t)}{P_\theta(P_I|Z_t)}. \quad (4)$$

By the Bayes' Rule, $P_\theta(P_F|Z_t) = \frac{P_\theta(Z_t|P_F)\cdot P_\theta(P_F)}{P(Z_t)}$, and assume a uniform prior over P, which causes the $P_\theta(P_F)$ and $P_\theta(P_I)$ terms to cancel, we have

$$\frac{P_\theta(P_F|Z_t)}{P_\theta(P_I|Z_t)} \propto \frac{P_\theta(Z_t|P_F)}{P_\theta(Z_t|P_I)}. \quad (5)$$

The direction for the transition can be interpreted as the gradient of the log-likelihood:

$$\nabla_{Z_t}\log P(Z_t|P_F) - \nabla_{Z_t}\log P(Z_t|P_I). \quad (6)$$

Combining Equation 2 and Equation 6, we obtain the unit transitional direction in the ϵ-space:

$$D_{\epsilon_\theta}(Z_t, P_{I-F}) = \frac{\epsilon_\theta(Z_t, P_F) - \epsilon_\theta(Z_t, P_I)}{\|\epsilon_\theta(Z_t, P_F) - \epsilon_\theta(Z_t, P_I)\|_2}. \quad (7)$$

The obtained $D_{\epsilon_\theta}(Z_t, P_{I-F})$ is a local, data-dependent direction that guides the transition from P_I to P_F for Z_t.

The quality of the transitional direction depends on the precision of the denoising score. Using simple terms in P_I and P_F may introduce ambiguity, leading to unintended changes as shown in Figure 6. To reduce such issues, we use full descriptive prompts instead of isolated attributes, ensuring the denoising score captures the intended context and enhances the accuracy of the transitional direction.

[2]For example, *"A boat is drifting on the river* ***in the dark light.****"*
[3]For example, *"A boat is drifting on the river in* ***the bright light.****"*

3.3. Neutral Prompt Anchoring

To achieve smooth attribute transitions in video generation, we manipulate the video latents during the denoising process at inference time. By guiding each frame along the transitional direction, we ensure a gradual and seamless change of attributes across the entire sequence.

We leverage the transitional direction D in the ϵ-space and adjust the denoising score in Equation 3 for each frame. Notably, the transitional direction is crafted with a holistic view of the video, providing a consistent manipulation scale across all frames. Moreover, the direction of each frame can be treated independently, allowing fine-grained control over attribute adjustments throughout the video sequence. A straightforward approach is to incrementally add frame-dependent scales of D to the score associated with P_I, $\epsilon_\theta(Z_t, P_I)$. The refined denoising score of the jth frame of the video can be formulated as:

$$\hat{\epsilon}^j_\theta(Z_t, P_{I-F}) = \epsilon^j_\theta(Z_t, P_I) + \alpha_j \cdot D^j_{\epsilon_\theta}(Z_t, P_{I-F}), \quad (8)$$

where $\hat{\epsilon}^j_\theta(Z_t, P_{I-F})$ represents the denoising score of the jth frame with specified attribute transition and the scale factor α_j changes uniformly over the sequence of frames.

However, as shown in Figure 5, inconsistencies are observed in the synthesized video, where the final frame differs significantly from the first due to the distance effect [20]: the transitional direction defines a hyperplane separating the initial and final attributes. Denoising scores near this boundary effectively yield smooth and controlled transitions. However, moving too far from the hyperplane can amplify changes, causing unwanted distortions.

As such, we employ a neutral prompt to keep the denoising score near the boundary throughout the denoising process. Specifically, given P_I and P_F, we construct a neutral prompt P_N[4] by extracting the common elements from both prompts that describe only the motion in the video without referencing the attributes we intend to change. We designate the middle frame to have a neutral attribute. The transition can be achieved by adjusting the denoising score along or against the transitional direction for the following and preceding frames. Equation 8 can thus be rewritten as:

$$\hat{\epsilon}^j_\theta(Z_t, P_{I\text{-}F}) = \begin{cases} \epsilon^j_\theta(Z_t, P_N) - \alpha_{M\text{-}j} \cdot D^j_{\epsilon_\theta}(Z_t, P_{I\text{-}F}), & j < M \\ \epsilon^j_\theta(Z_t, P_N), & j = M \\ \epsilon^j_\theta(Z_t, P_N) + \alpha_{j\text{-}M} \cdot D^j_{\epsilon_\theta}(Z_t, P_{I\text{-}F}), & j > M \end{cases}, \quad (9)$$

where M is the middle frame of the video. By anchoring the middle frame with a neutral prompt, which is devoid of dominant attributes and thus near the boundary, we keep the transition near the hyperplane, effectively mitigating the distance effect.

[4]For example, *"A boat is drifting on the river."* in Figure 2.

3.4. Overall Sampling Flow

As early stages of the denoising process mainly establish geometric, such as the layout of the video [2, 4, 11, 19], in the first τ denoising steps, we employ the neutral prompt and sample from diffusion model using classifier-free guidance in Equation 3 to construct a general video structure for enhanced consistency. After completing τ denoising steps, we guide the sampling process using the transition direction to achieve smooth attribute transitions. It is worth noting that classifier-free guidance can still be applied to enhance the quality of the generated samples:

$$\tilde{\epsilon}_\theta(Z_t, P_{I-F}) = (1+\omega) \cdot \hat{\epsilon}_\theta(Z_t, P_{I-F}) - \omega \cdot \epsilon_\theta(Z_t). \quad (10)$$

4. CAT-Bench

To assess model performances in generating videos with attribute transitions, we construct a new benchmark, Controlled-Attribute-Transition Benchmark (CAT-Bench), to evaluate the fidelity of models in handling attribute transitions while maintaining motion consistency and fluidity.

4.1. Benchmark Construction

We define **eight** attribute categories for transitions: four focused on *human attributes* (age, beard, makeup, and hair), two on *subject attributes* (color and material), and two on *background attributes* (light conditions and weather conditions). We generate prompt pairs to capture attribute changes in videos. The first prompt describes the initial frame, while the second outlines the final frame. Each pair is carefully designed to incorporate both motion and attribute transitions. To maintain consistent motion throughout the video, both prompts are identical except for the part specifying the evolving attribute, allowing each prompt pair to convey ongoing motion and attribute changes seamlessly.

We use GPT-4 [1] to generate multiple samples across all attribute categories, producing 15 prompt pairs per category, resulting in a total of 120 prompts. To facilitate a comprehensive assessment, we also use GPT-4 to generate a single prompt describing the full attribute transition for each prompt pair, allowing us to compare and analyze different generation approaches thoroughly.

4.2. Evaluation Metrics

We introduce two metrics to assess attribute transition quality: *Wholistic Transition Score* and *Frame-wise Transition Score*. *Wholistic Transition Score* evaluates the overall correctness of the attribute transition, determining whether the change from the first to the last frame follows the intended progression. The score is computed using a CLIP-based directional similarity between the transformation of visual features from the initial to the final frame and the semantic

change described by the prompts:

$$\frac{\langle E_I(F^l) - E_I(F^f), E_T(P^s) - E_T(P^f)\rangle}{\|E_I(F^l) - E_I(F^f)\|\|E_T(P^s) - E_T(P^f)\|}, \quad (11)$$

where E_I denotes the CLIP image encoder and E_T represents the CLIP text encoder. F^f and F^l are the visual features of the first and last frames, respectively, while P^f and P^l are the features of the first and second prompts.

In contrast, *Frame-wise Attribute Transition* assesses the smoothness and gradual nature of the transition, ensuring that each frame changes consistently without abrupt shifts. We compute the directional similarity between every consecutive frame in the video to determine if each pair aligns with the direction specified in the prompt pair:

$$\frac{1}{N-1}\sum_{i=1}^{N-1}\frac{\langle E_I(F^{i+1}) - E_I(F^i), E_T(P^s) - E_T(P^f)\rangle}{\|E_I(F^{i+1}) - E_I(F^i)\|\|E_T(P^s) - E_T(P^f)\|}, \quad (12)$$

thereby verifying the smoothness of the transition.

5. Experimental Results

We conduct experiments to evaluate the performance of our method quantitatively and qualitatively. The proposed CAT-Bench and the source code will be available to the public.

5.1. Experimental Setup

Our method is built on the open-source T2V foundation model, VideoCrafter2 [5], which serves as the backbone of our implementation. In our experiments, we set the total number of denoising time steps T to 50, consistent with the backbone model. The number of frames F in each generated video is set to 32. The parameter τ, which controls the sampling process, is set to 5.

Baselines. We evaluate our method alongside various state-of-the-art baselines, including single-prompt and multiprompt approaches. The single-prompt baselines consist of AnimateDiff [9], ModelScope [25], Latte [17], and VideoCrafter2 [5], which are foundational T2V models. The multi-prompt baselines include FreeBloom [14], Gen-L [24], Freenoise [19], and VideoTetris [22]. Similar to our method, Gen-L and Freenoise are training-free approaches and also use VideoCrafter2 as their backbone.

Benchmarks. In addition to our CAT-Bench, we evaluate the effectiveness of attribute transition using TC-Bench-T2V [7]. TC-Bench is a benchmark designed to assess temporal compositionality in video generation, featuring different types of object state changes. For our evaluation, we focus on the attribute and background shifting subsets to assess whether the transition can be fulfilled effectively. While TC-Bench provides insight on various object state transitions, it lacks scenes with motion dynamics in prompts involving attribute and background transition. The limitation underscores the importance of our CAT-Bench, which addresses the gap by incorporating motions to evaluate attribute transitions in complex scenes.

5.2. Qualitative Evaluation

Figure 3 showcases the generation results from baseline approaches and the proposed method. Results from AnimateDiff, ModelScope, Latte, and VideoCrafter2 show they struggle to handle attribute transitions effectively when guided by a single prompt. Without retraining on relevant data, it is challenging to interpret or execute the attribute transitions specified in the prompts, resulting in subpar performance. On the other hand, FreeBloom and VideoTetris utilize LLMs to generate one descriptive prompt per frame. However, because subtle attribute state differences between consecutive frames are difficult to convey accurately through text, their results exhibit abrupt attribute shifts or unwanted changes. GenL and FreeNoise use prompt interpolation to achieve transitions between motions, but their approach often fails to produce intermediate attribute states reliably, indicating that interpolation does not always translate well into intermediate attributes. Additionally, because the first and last frames are generated from different prompts, the frames are likely to appear as disjointed images, introducing further inconsistency. In contrast, our proposed framework generates videos with smooth and accurate attribute transitions, ensuring consistency and a gradual progression across frames. Moreover, our method achieves the desired attribute transitions while maintaining video quality comparable to the base model. Quantitative evaluations of video quality are reported in the supplementary materials, along with additional qualitative results.

5.3. Quantitative Evaluation

CAT-Bench. Table 1 presents the Wholistic Transition Score evaluation results on CAT-Bench. The table shows that single-prompt methods achieve lower scores compared to multi-prompt-based ones, indicating that single-prompt approaches struggle to produce accurate attribute transitions. The observation aligns with the qualitative evaluation results. Although multi-prompt approaches generally perform better, prompt interpolation approaches, Gen-L and FreeNoise, show suboptimal results for transitions involving *beard*, which are described as *with beard* and *without beard* in prompts. The results indicate that interpolating between binary attributes specified using "with" and"without" fails to generate intermediate attribute states effectively.

Table 2 reports the Frame-wise Transition Score, which measures the smoothness of attribute transitions. It reveals that while some multi-prompt methods achieve correct wholistic transitions, they lack smoothness, as indi-

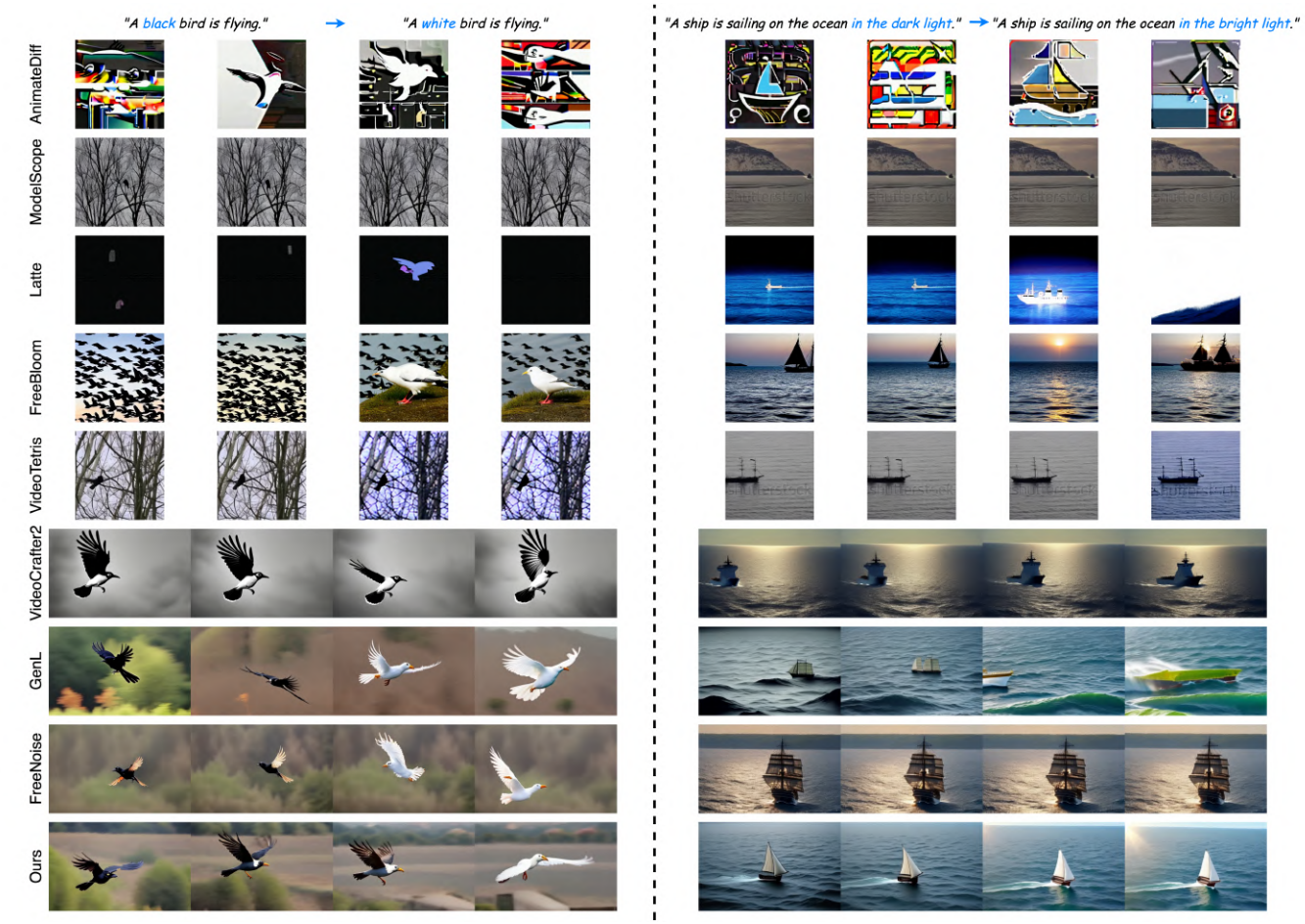


Figure 3. Generation Results for the Specified Attribute Transition. AnimateDiff, ModelScope, Latte, and VideoCrafter2 are foundational text-to-video (T2V) models. FreeBloom and VideoTetris use LLMs to generate multiple detailed prompts as input and generate videos. Gen-L, FreeNoise, and our method are training-free approaches and utilize VideoCrafter2 as the backbone.

Table 1. Evaluation Results of the Wholistic Transition Score on CAT-Bench. The highest scores are highlighted in blue.

	Age	Beard	Makeup	Hair	Color	Material	Light	Weather	Overall
Single-Prompt Generation									
AnimateDiff	0.0143	0.0012	0.0133	0.0098	0.0001	-0.0156	-0.0021	0.0207	0.0082
ModelScope	0. 0201	-0.0170	-0.0025	-0.0137	0.0285	0.0009	0.0186	-0.0047	0.0042
Latte	0.0079	0.0015	-0.0006	-0.0218	0.0210	0.0012	0.0003	0.0046	0.0019
VideoCrafter2	0.0002	0.0019	0.0178	0.0061	-0.0063	-0.0169	-0.0084	0.0041	0.0022
Multi-Prompt Generation									
Free-Bloom	0.1561	0.0255	0.0718	0.2244	0.0031	0.0162	0.0746	0.1987	0.1077
VideoTetris	0.0061	0.0210	-0.0074	0.0295	0.0222	-0.0008	0.0272	-0.0052	0.0134
Gen-L	0.1902	0.0015	0.0891	0.1881	0.0439	-0.0033	0.0769	0.2264	0.1166
FreeNoise	0.1085	0.0003	0.0586	0.1142	0.0104	-0.0037	0.0297	0.0834	0.0578
Ours	**0.2257**	**0.0332**	**0.1228**	**0.2440**	**0.0538**	**0.0284**	**0.1127**	**0.2480**	**0.1486**

cated by several negative values. The results suggest that the frame-wise transitions occasionally move against the desired direction, disrupting the smoothness of the video. Regarding correctness and smoothness, Tables 1 and Table 2 show that our method demonstrates improved performance on both criteria compared to existing approaches.

TC-Bench-T2V. Table 3 shows the evaluation results on TC-Bench-T2V. TCR and TC-Score are assertion-based evaluations conducted by LLMs, measuring transition com-

Table 2. Evaluation Results of the Frame-wise Transition Score on CAT-Bench. The highest scores are highlighted in blue.

	Age	Beard	Makeup	Hair	Color	Material	Light	Weather	Overall
Single-Prompt Generation									
AnimateDiff	-0.0002	0.0007	0.0011	0.0001	0.0002	0.0009	-0.0002	0.0021	0.0004
ModelScope	0.0010	0.0003	0.0001	-0.0016	0.0020	0.0001	0.0028	-0.0009	0.0001
Latte	0.0009	-0.0009	-0.0001	-0.0033	0.0007	-0.0001	-0.0001	0.0019	-0.0002
VideoCrafter2	0.0005	0.0011	0.0018	0.0001	-0.0012	-0.0014	-0.0013	0.0010	0.0003
Multi-Prompt Generation									
Free-Bloom	0.0119	0.0026	0.0047	-0.0153	0.0002	-0.0012	0.0071	-0.0169	-0.0020
VideoTetris	0.0013	-0.0002	0.0012	0.0024	0.0022	0.0009	0.0011	0.0005	0.0012
Gen-L	0.0262	-0.0010	0.0120	0.0246	0.0056	0.0001	0.0095	0.0271	0.0135
FreeNoise	0.0156	0.0005	0.0066	0.0153	0.0009	-0.0012	0.0033	0.0086	0.0066
Ours	**0.0321**	**0.0044**	**0.0237**	**0.0382**	**0.0073**	**0.0043**	**0.0157**	**0.0306**	**0.0201**

Table 3. Quantitative Evaluation on TC-Bench-T2V. We report assertion-based TCR and TC-Score proposed in the benchmark. The highest score is highlighted in blue.

	Attribute		Background	
Method	TCR↑	TC-Score↑	TCR↑	TC-Score↑
ModelScope	3.52	0.5942	3.54	0.5715
VideoCrafter2	4.25	0.6166	7.06	0.6338
FreeBloom	6.32	0.6256	24.02	0.7394
TC-Bench	13.08	0.6579	35.43	0.7916
Ours	**13.47**	**0.6889**	**35.88**	**0.8043**

Table 4. User Study. Participants are required to select the best video based on five different criteria. The table reports the percentage of users who choose each approach as the best.

	VideoCrafter2	GenL	Freenoise	**Ours**
Transition Correctness	4.87%	28.91%	17.07%	**49.12%**
Transition Smoothness	7.69%	28.32%	18.53%	**45.45%**
Video Consistency	15.03%	27.62%	19.93%	**37.41%**
Motion Correctness	11.22%	29.08%	17.19%	**41.75%**
Overall Performance	5.81%	28.72%	19.27%	**46.19%**

pletion and object consistency. TC-bench-T2V evaluates various object changes, including complex state changes such as a rose transitioning from deep crimson to fully bloomed. Our approach performs well on TC-Bench-T2V, demonstrating its effectiveness in handling diverse attribute transitions while maintaining object consistency.

User Study. We conduct a user study to evaluate the results based on human subjective assessments. Participants are asked to watch videos generated by four approaches that use VideoCrafter2 as the basemodel and select the best video based on five criteria: Transition Correctness, Transition Smoothness, Video Consistency, Motion Correctness, and Overall Performance. Table 4 presents the results,

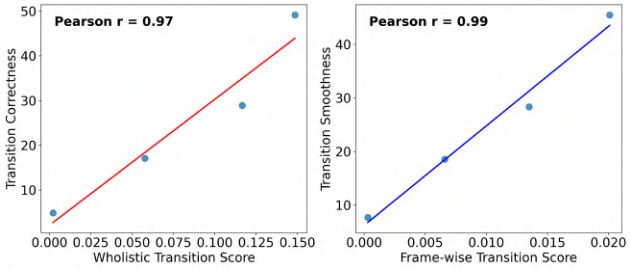


Figure 4. Correlation between Our Metrics and Human Perception. The x-axis represents the Wholistic and Frame-wise Transition Score, while the y-axis corresponds to Transition Correctness and Smoothness from the user study. The results indicate a strong correlation, validating the effectiveness of our proposed metrics.

showing that our method performs favorably compared to other approaches using the same base model. Moreover, the ranking orders for Transition Correctness and Smoothness align with the rankings in Tables 1 and Table 2, which report the Wholistic and Frame-wise Transition Scores. We illustrate the alignment in Figure 4, which demonstrates that our proposed metrics effectively correlate with human perception. The Pearson correlation coefficients (r) further confirm the strong correlation, validating the reliability of our metrics in assessing perceptual quality.

5.4. Ablation Study

Neutral Prompt. We perform an ablation study on neutral prompt anchoring to show its necessity. As depicted in Figure 5, without the neutral prompt, while the desired change is achieved, unintended alterations occur: the color of the car changes from blue to black and veers instead of continuing straight. In contrast, with neutral prompt anchoring, the appearance of the car is preserved, while the background transitions smoothly from sunny to rainy.

Full Prompt. We conduct an ablation study on the use of full prompts. As discussed in Section 3.2, the quality of transitional scenes depends on the conditional de-

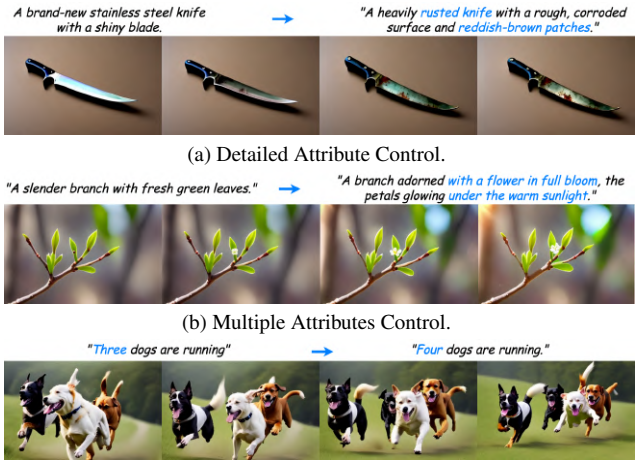
Figure 5. Ablation on Neutral Prompt Anchoring. The unanchored approach introduces severe inconsistencies, such as unwanted changes in color and motion.

Figure 6. Ablation on Full Prompt. Simple prompts lack context, leading to ambiguous or incomplete transitions. In contrast, full prompts provide clear guidance for precise attribute transitions.

noising score, which aligns more accurately with precise prompts. As shown in Figure 6, using a simple prompt such as "*black*" and "*white*" is ambiguous, as it may refer to different entities, leading to unintended changes or no change at all due to uncertainty. In contrast, using full prompts, such as "*a woman in a black hat.*" transitioning to "*a woman in a white hat.*", provides clear contextual guidance, resulting in a more precise and controlled transition.

5.5. Broader Applicability

Beyond the attribute transitions systematically evaluated in our benchmark, we demonstrate a greater applicability of our method in Figure 7 by allowing more detailed prompts over attribute changes. First, our approach allows for precise control over transitions that go beyond basic changes, offering detailed modifications. Additionally, our method supports multiple attribute control, enabling the simultaneous adjustment of multiple attributes while ensuring smooth transitions. Finally, we present the change of object count, facilitating dynamic scene modifications while preserving motion consistency. These capabilities underscore the versatility and broader applicability of our method, making it



(a) Detailed Attribute Control.

(b) Multiple Attributes Control.

(c) Object Count Change.

Figure 7. More Applications. We showcase three enhanced capabilities of our method: (a) detailed attribute control, (b) multiple attribute control, and (c) object count change for dynamic scene modifications while maintaining consistency.

suitable for more complex, real-world scenarios.

5.6. Limitations

Our method is compatible with any classifier-free guidance method, including both UNet-based and transformer-based diffusion models, enabling smooth transitions without sacrificing video quality. Experiments are reported in the Supplemental Material. However, the effectiveness of the approach depends on the performance of the backbone model. If the backbone model struggles to represent certain attributes, creating precise transitions can become difficult. Nonetheless, the compatibility of our approach with any classifier-free guidance approach suggests that as backbone models continue to evolve, our method will benefit from these advancements, leading to even better results over time.

6. Conclusion

We present a simple yet effective approach for video generation with smooth attribute transitions. Our method introduces a transitional direction to guide the sampling process during denoising, ensuring gradual and coherent transitions. Using neutral prompt anchoring, we mitigate unwanted changes and maintain consistency. Additionally, we propose CAT-Bench, a benchmark designed to evaluate complex scenes involving both attribute transitions and motion dynamics. To assess transition quality, we introduce the Wholistic Transition Score and the Frame-wise Transition Score, which measure correctness and smoothness. Experimental results demonstrate that our framework enables seamless attribute transitions, advancing video generation capabilities. In addition, our proposed metrics align closely with human evaluations, confirming their reliability in assessing transition quality.

Acknowledgements

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project). It was also partially supported by the National Science and Technology Council (NSTC), Taiwan, under Grant NSTC-112-2628-E-002-033-MY4, and was financially supported by the Center of Data Intelligence: Technologies, Applications, and Systems, National Taiwan University (Grants 114L900901, 114L900902, and 114L900903), through the Featured Areas Research Center Program of the Higher Education Sprout Project, Ministry of Education, Taiwan.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 4

[3] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. In *NeurIPS*, 2023. 2

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *CVPR*, 2023. 4

[5] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, 2024. 1, 2, 5

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 2

[7] Weixi Feng, Jiachen Li, Michael Saxon, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. *arXiv preprint arXiv:2406.08656*, 2024. 5

[8] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *ECCV*, 2022. 1

[9] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024. 1, 2, 5

[10] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 1, 2

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023. 4

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS*, 2021. 2, 3

[13] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023. 1

[14] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. In *NeurIPS*, 2023. 2, 5

[15] Taegyeong Lee, Soyeong Kwon, and Taehwan Kim. Grid diffusion models for text-to-video generation. In *CVPR*, 2024. 1

[16] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 2

[17] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 1, 2, 5

[18] Gyeongrok Oh, Jaehwan Jeong, Sieun Kim, Wonmin Byeon, Jinkyu Kim, Sungwoong Kim, and Sangpil Kim. Mevg: Multi-event video generation with text-to-video models. In *ECCV*, 2024. 2

[19] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. In *ICLR*, 2024. 2, 4, 5

[20] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 4

[21] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *NeurIPS*, 2021. 2, 3

[22] Ye Tian, Ling Yang, Haotian Yang, Yuan Gao, Yufan Deng, Jingmin Chen, Xintao Wang, Zhaochen Yu, Xin Tao, Pengfei Wan, et al. Videotetris: Towards compositional text-to-video generation. In *NeurIPS*, 2024. 2, 5

[23] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *ICLR*, 2022. 1, 2

[24] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023. 2, 5

[25] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1, 2, 5

[26] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023. 1