TALC: Time-Aligned Captions for Multi-Scene Text-to-Video Generation

Hritik Bansal[1] **Yonatan Bitton**[2†] **Michal Yarom**[2†]
Idan Szpektor[2*] **Aditya Grover**[1*] **Kai-Wei Chang**[1*]
[1]University of California Los Angeles [2]Google Research
https://github.com/Hritikbansal/talc

Abstract

Most of these text-to-video (T2V) generative models often produce single-scene video clips that depict an entity performing a particular action (e.g., 'a red panda climbing a tree'). However, it is pertinent to generate multi-scene videos since they are ubiquitous in the real-world (e.g., 'a red panda climbing a tree' followed by 'the red panda sleeps on the top of the tree'). To generate multi-scene videos from the pretrained T2V model, we introduce a simple and effective **T**ime-**Al**igned **C**aptions (TALC) framework. Specifically, we enhance the text-conditioning mechanism in the T2V architecture to recognize the temporal alignment between the video scenes and scene descriptions. For instance, we condition the visual features of the earlier and later scenes of the generated video with the representations of the first scene description (e.g., 'a red panda climbing a tree') and second scene description (e.g., 'the red panda sleeps on the top of the tree'), respectively. As a result, we show that the T2V model can generate multi-scene videos that adhere to the multi-scene text descriptions and be visually consistent (e.g., entity and background). Further, we finetune the pretrained T2V model with multi-scene video-text data using the TALC framework. We show that the TALC-finetuned model outperforms the baseline by achieving a relative gain of 29% in the overall score, which averages visual consistency and text adherence using human evaluation.

1 Introduction

The ability to generate videos that simulate the physical world has been a long-standing goal of artificial intelligence [1, 47, 10, 36]. In this regard, text-to-video (T2V) models have seen rapid advancements by pretraining on internet-scale datasets of images, videos, and texts [9, 6]. Previous work [20, 18, 28, 49, 8, 7] primarily focus on training conditional denoising diffusion probabilistic models [22] on paired video-text data [4, 54]. After training, these models allow for video generation by sampling from the trained diffusion model, conditioned on a text prompt. However, most of the open-models such as ModelScope[49] VideoCrafter [13, 14], OpenSora [57] are trained with single-scene video-text dataset [4, 50], which is widely available and easy to acquire. However, real-world scenarios often require the generation of multi-scene videos from multi-scene descriptions (e.g., *Scene1:* 'A koala is napping on a tree.' *Scene2:* 'The koala eats leaves on the tree.'). In such cases, the generated video should accurately depict the events in their temporal order (e.g., *Scene2* follows *Scene1*) while maintaining visual consistency, meaning that backgrounds and entities should remain consistent across scenes. While high-performance T2V models such as Sora [36] might be able to generate multi-scene videos, we point out that they are closed-source models trained with massive compute resources and lack sufficient details on the model design, training protocol, and

[†] Equal Contribution. [*] Equal Advising. Contact hbansal@ucla.edu,yonatanbitton1@gmail.com.

Workshop on Video-Language Models at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

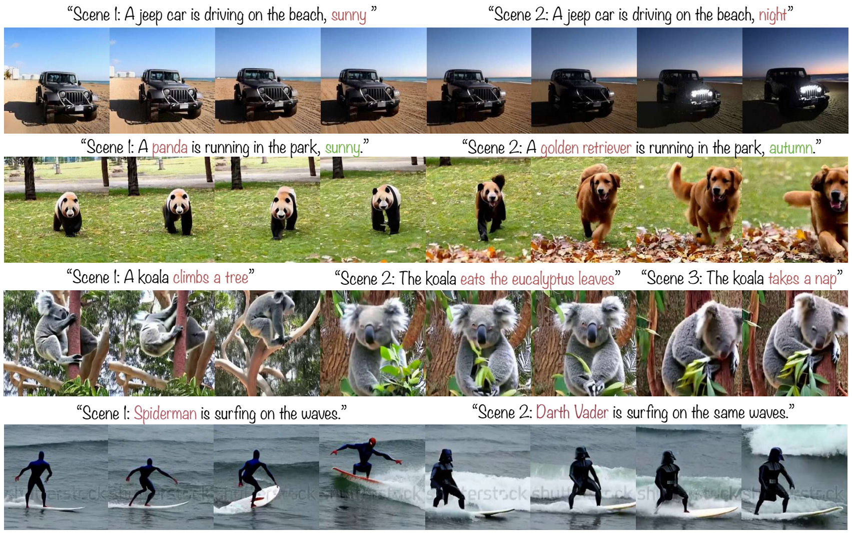

"Scene 1: A jeep car is driving on the beach, sunny "
"Scene 2: A jeep car is driving on the beach, night"
"Scene 1: A panda is running in the park, sunny."
"Scene 2: A golden retriever is running in the park, autumn."
"Scene 1: A koala climbs a tree"
"Scene 2: The koala eats the eucalyptus leaves"
"Scene 3: The koala takes a nap"
"Scene 1: Spiderman is surfing on the waves."
"Scene 2: Darth Vader is surfing on the same waves."

Figure 1: **Examples of multi-scene video generation using the TALC framework.** Our proposed method, TALC, enhances scene-level alignment between text and video, facilitating the generation of videos with seamless transitions between textual descriptions while preserving visual consistency. The first two rows are generated using Lumiere [6] and the last two rows use ModelScope [49].

datasets. In this work, we present a complementary approach and tackle the challenge of effectively leveraging the capabilities of base T2V models for multi-scene video generation.

The multi-scene T2V generation differs from long video synthesis where the goal is to either interpolate (few frames to many frames) [18] or create continuing patterns of the single event in the generated video [8]. Prior work like Phenaki [45, 28] use transformers [44, 3] to generate video frames for a given scene autoregressively. However, it is hard for their model to generate multiple scenes reliably as the context length increases with the history of text descriptions and visual tokens [56] of the previous generated videos (e.g., generating *Scene 4* conditioned on the *Scene1, 2, 3* videos and descriptions). Other works [40] utilize a latent diffusion model [41] to generate video frames autoregressively by conditioning on the entire history of generated videos and scene descriptions. However, the approach is (a) slow due to repeated sampling, (b) generates only one frame per scene description, and (c) shown to work with only limited cartoon characters [29, 55] instead of wide range of visual concepts in the real-world. In this work, our goal is to generate multi-scene videos in the end-to-end manner using a diffusion T2V generative model. Prior work like VideoDirectorGPT [30, 31] generates multi-scene videos by utilizing knowledge of the entity, background, and their movements from large language models [2]. However, these videos are generated independently for each scene before being merged.

To remedy these challenges, we propose TALC (**T**ime-**AL**igned **C**aptions), a simple and effective framework to generate consistent and faithful multi-scene videos. In particular, our approach conditions the T2V generative model with the knowledge of the temporal alignment between the parts of the multi-scene video and multi-scene descriptions (Figure 3). Specifically, TALC conditions the visual representations of earlier video frames on the embeddings of the earlier scene description, and likewise, it conditions the representations of later video frames on the embeddings of the later scene description in the temporal dimension. Additionally, the temporal modules in the T2V diffusion architecture allows information sharing between video frames (the first half and the second half) to maintain visual consistency. Therefore, TALC enhances the scene-level text-video alignment while providing the scene descriptions to the diffusion model all at once (Figure 1).

Prior methods like FreeNoise [38] propose a motion injection strategy to address these challenges. However, this method is quite sophisticated and difficult to control due to diverse hyperparameters, such as time-specific motion injection, prompt interpolation, and injection in specific cross-attention layers. In contrast, our approach eliminates these complexities and introduces a straightforward mechanism that significantly boosts performance. Unlike previous work, we also demonstrate that finetuning with TALC using real-world multi-scene data can enhance generation capabilities.

Specifically, we propose a pipeline to curate multi-scene video data and subsequently finetune the T2V model on this multi-scene data using TALC (§3.2).

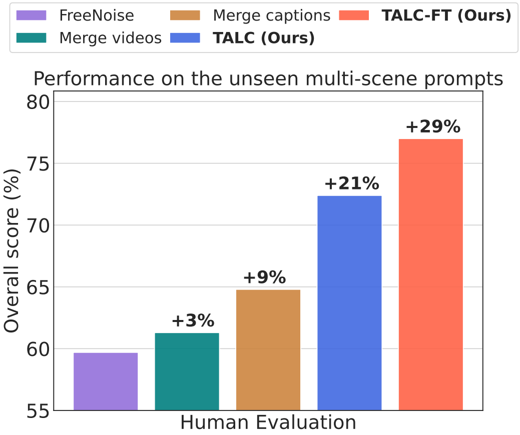


Figure 2: **Summary of the results.** We compare several baselines and our TALC framework with ModelScope [49] for generating multi-scene videos on the unseen prompts. Specifically, we study the overall score which averages the visual consistency (object and background) and text adherence (video-text alignment). We observe that using TALC with the base model i.e., training-free achieves relative gains of 21% in comparison to FreeNoise [38]. In addition, we show that multi-scene finetuning with TALC (TALC-FT) allows larger gains, achieving relative gains upto 29%.

In our experiments, we assess the visual consistency (background and entity consistency) and multi-scene script adherence of the generated videos from T2V generative models. Through our human evaluation, we find that TALC strikes an effective balance between visual consistency and text adherence, and outperforms the FreeNoise by achieving relative gains of 21% points on the overall score. This score represents the average of visual consistency and text adherence scores (Figure 2). Furthermore, we construct a multi-scene text-video dataset from real-world videos and fine-tune the T2V generative model using TALC. We show that finetuning with TALC outperforms FreeNoise by achieving relative gains of 29% on the overall score (Figure 2). Further, we perform automatic evaluation for scalable judgements (§5.4) and provide qualitative examples to highlight the benefits of our approach (§5.5). We present the related work in Appendix §A.

2 Preliminaries

2.1 Diffusion Models for Text-to-Video Generation

Diffusion models [22, 33] $p_\theta(x)$ are a class of generative models that learn data distribution $p_{data}(x)$. Due to their flexible design, we can train their class-conditional versions to learn class-conditional data distributions $p_{data}(x|y)$ where y is the conditioning variable. We assume a dataset $\mathcal{S} \subset \mathcal{V} \times \mathcal{T}$ consisting of pairs of (V_j, T_j) where $V_j \in \mathbb{R}^{L\times 3\times H\times W}$ is a raw video consisting of 3 RGB channels, L frames, H height, W width, and T_j is a text caption. We use $\mathcal{V}$ and $\mathcal{T}$ to denote the domain of videos and text, respectively. The aim of T2V generative modeling is to learn the conditional distribution of the videos conditioned on the text $p_{\mathcal{S}}(V_j|T_j)$. In this work, we consider diffusion-based generative models that learn the data distribution via iterative denoising of the input video $z_j \in \mathbb{R}^{L\times C\times H'\times W'}$. Here, z_j can either represent the input video in the raw pixel space V_j [6] or it can represent the latent representation of the video $z_j = \mathcal{E}(V_j)$ for the latent diffusion models [41] where $\mathcal{E}$ is an encoder network [27].

Given z_j, diffused variable $z_{\tau,j} = \alpha_\tau z_j + \beta_\tau\epsilon$ are constructed where $\epsilon \sim \mathcal{N}(0, I)$ where α_τ and β_τ are sampled from the noise scheduler p_τ [16]. Finally, we train a denoiser network f_θ [42, 37] that inputs the diffused variable z_τ and embeddings of the text caption to predict the target vector y where y can be the original noise ϵ, which minimizes the denoising objective [22]:

$$\mathbb{E}_{(V_j,T_j)\in S,\tau\sim p_\tau,\epsilon\sim N(0,I)}\left[||\epsilon - f_\theta(\tau, z_{\tau,j}, h_j)||_2^2\right] \tag{1}$$

where $h_j = \mathcal{H}(T_j) \in \mathbb{R}^d$ is the embedding of the text caption T_j where $\mathcal{H}$ is the text embedding model [39] and d is the dimension size.

2.2 Text Conditioning Mechanism

To ensure the effective textual controllability of video generation, the structure of the denoiser networks is equipped with a cross-attention mechanism [49, 18]. Specifically, it conditions the visual content $z_\tau \in \mathbb{R}^{L\times C\times H'\times W'}$ on the text. To do so, we first *repeat* the text embeddings of the text caption $r_j = R(h_j) \in \mathbb{R}^{L\times d}$ where R is a function that repeats the input text embedding h_j for L times in the temporal dimension. Intuitively, the repeat operation represents that the L frames of the video z_j are semantically aligned with the textual description T_j or its text embedding r_j.

These repeated text embeddings r_j are inputs to the spatial attention block as the key and value in the multi-head attention block. The cross-attention enables the intermediate visual features to capture the semantic information that facilitates an alignment between the language and vision embeddings.

$$z'_{\tau,j} = CA_{f_\theta}(Q = z_{\tau,j}; K = r_j; V = r_j) \tag{2}$$

where CA_{f_θ} is the cross attention mechanism with Q, K, V as the query, key, and value, respectively, in the spatial blocks of the denoiser network. Additionally, $z'_{\tau,j}$ is the intermediate representation that is informed with the visual and textual content of the data. In addition to the spatial blocks, the denoiser network also consists temporal blocks that aggregate features across video frames which are useful for maintaining visual consistency in the video.

2.3 Multi-Scene Text-to-Video Generation

In this work, we aim to generate multi-scene video $X = \{x_1, x_2, \ldots, x_n\}$ from multi-scene descriptions $Y = \{y_1, y_2, \ldots, y_n\}$ where n are the number of sentences and each sentence y_j is a scene description for scene j. Additionally, the index j also defines the temporal order of events in the multi-scene script i.e., we want the events described in the scene j to be depicted earlier than the events described in the scene k where $k > j$. Further, we want the parts of the entire generated video X, given by x_j, to have high video-text semantic alignment with the corresponding scene description y_j. In addition, we expect the appearance of the objects to remain consistent throughout the video unless a change is specified in the description.

3 TALC: Time-Aligned Captions for Multi-Scene T2V Generation

3.1 Approach

Most of the existing T2V generative models [49, 13, 6] are trained with large-scale short video-text datasets (10 seconds - 30 seconds) such as WebVid-10M [4]. Here, each instance of the dataset consists of a video and a human-written video description. These videos either lack the depiction of multiple events, or the video descriptions do not cover the broad set of events in the video, instead focusing on the major event shown. As a result, the pretrained T2V generative models only synthesize single video scenes depicting individual events.

We introduce TALC, a novel and effective framework to generate multi-scene videos from diffusion T2V generative models based on the scene descriptions. Our approach focuses on the role of text conditioning mechanism that is widely used in the modern T2V generative models (§2.2). Specifically, we take inspiration from the fact that the parts of the generated video x_j should depict the events described in the scene description y_j. To achieve this, we ensure that the representations for the part of the generated video aggregates language features from the scene description y_j.

Consider that we want to generate a multi-scene video $X \in \mathbb{R}^{L\times 3\times H\times W}$ from the scene descriptions $y_j \in Y$, using a T2V generative model f_θ. Furthermore, we assume that individual video segments x_j are allocated L/n frames within the entire video X. Let $z_X = [z_{x_1}; z_{x_2}; \ldots; z_{x_n}] \in \mathbb{R}^{L\times C\times H'\times W'}$ represent the representation for the entire video X, and $z_{x_j} \in \mathbb{R}^{(L/n)\times C\times H'\times W'}$ for the j^{th} part of the video that are concatenated in the temporal dimension. In addition, consider $r_Y = \{r_{y_1}, \ldots, r_{y_n}\}$ be the set of text embeddings for the multi-scene description Y and y_j be an individual scene description. In the TALC framework, the Eq. 2 is changed to:

$$z'_{\tau,x_j} = CA_{f_\theta}(Q = z_{\tau,x_j}, K = r_{y_j}, V = r_{y_j}) \tag{3}$$

$$z'_{\tau,X} = [z'_{x_1}; z'_{x_2}; \ldots; z'_{x_n}] \tag{4}$$

Here, τ represents the timestamp in the diffusion modeling setup, which is applied during training as well as inference. We illustrate the framework in Figure 3. While TALC aims to equip the generative model with the ability to depict all the events in the multi-scene descriptions, the visual consistency is ensured by the temporal modules (attentions and convolution blocks) in the denoiser network. By design, our approach can be applied to the base T2V model without any further training.

3.2 Multi-Scene Video-Text Data Generation

While our approach generates better multi-scene videos, the text adherence capabilities of the pretrained T2V generative model are limited. This is due to the lack of multi-scene video-text data during its pretraining. To this end, we create a real-world multi-scene video-text dataset to allow further training of the pretrained T2V models. Specifically, we leverage the capability of the multimodal foundation model, Gemini-Pro-Vision [43], to generate high-quality synthetic data for enhanced video-text training [5]. Formally, we start with a video-text dataset $\mathcal{M} = \mathcal{A} \times \mathcal{B}$ consisting of pairs of (A_i, B_i) where A_i is a raw video and B_i is the corresponding video description from the dataset. Subsequently, we utilize PySceneDetect library [1] to generate continuous video scenes from $A_i = \{A_{i,1}, A_{i,2}, \ldots, A_{i,m}\}$ where m is the number of scene cuts in the video. Further, we sample the middle video frame $F_{i,j}$ as a representative of the semantic content in the video scene $A_{i,j}$. Finally, we input all the video frames $F_i = \{F_{i,1}, \ldots, F_{i,m}\}$ for a single video A_i and the entire video caption B_i to Gemini-Pro-Vision. Specifically, the model is prompted to generate high-quality captions for each of the frames $F_{i,j}$ such they form a coherent narrative guided by the common caption B_i. We provide the prompt provided to the multimodal model in Appendix §E. In Figure 4 we provide an instance for the multi-scene video-text data generation. We highlight that higher-quality multi-scene datasets would enhance the performance of the models with TALC framework. We provide the more details in Appendix C. [2]

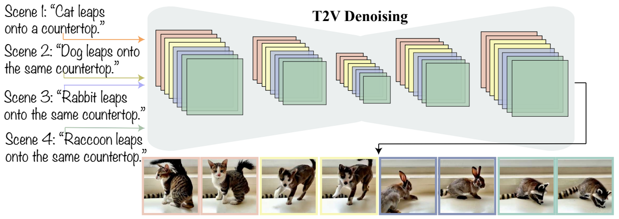

Scene 1: "Cat leaps onto a countertop."
Scene 2: "Dog leaps onto the same countertop."
Scene 3: "Rabbit leaps onto the same countertop."
Scene 4: "Raccoon leaps onto the same countertop."
T2V Denoising

Figure 3: **The architecture of Time-Aligned Captions (TALC).** During the generation process of the video, the initial half of the video frames are conditioned on the embeddings of the description of scene 1 (r_{y_1}), while the subsequent video frames are conditioned on the embeddings of the description of scene 2 (r_{y_2}).

4 Evaluation

4.1 Metrics

The ability to assess the quality of the generated multi-scene videos is a challenging task itself. As humans, we can judge the multi-scene videos across diverse perceptual dimensions [25] that the existing automatic methods often fails to capture [11]. Following [30], we focus on the visual consistency of the generated video, and text adherence capabilities of the T2V models.

Visual Consistency. This metric aims to assess the (entity or background) consistency between the frames of the multi-scene videos. Here, the **entity consistency** aims to test whether the entities in the multi-scene video are consistent across the video frames. For instance, the appearance of an animal should not change without a change described in the text description. In addition, the **background**

[1] https://github.com/Breakthrough/PySceneDetect

[2] We note that the parts of this pipeline are also utilized in a concurrent work, ShareGPTVideo [15]. However, their approach is focused on long-form video generation pretraining, while we are focus on multi-scene video generation for finetuning.

Table 1: **Human evaluation results.** We present the human evaluation results for the overall score for several baselines (e.g., FreeNoise, merge videos, and merge captions) and TALC framework for the ModelScope generative model. Specifically, we find that the TALC with the base model outperforms the FreeNoise by achieving the relative gain of 21% on the overall score. In addition, finetuning ModelScope with TALC framework enables better multi-scene video generation by achieving the highest overall score. Overall, TALC strikes a good balance between visual consistency and text adherence.

	Overall score	**Visual consistency**	**Text adherence**
FreeNoise [38]	59.7	77.0	42.5
Merge videos	61.3 (+2.6%)	55.0	**67.5**
Merge captions	64.8 (+8.5%)	**96.5**	33.0
Finetuning w/ Merge captions	61.1 (+2.3%)	80.0	42.3
TALC	72.4 (+21%)	92.3	52.5
Finetuning w/ TALC	**76.8** (+29%)	86.4	67.2

consistency aims to test whether the background of the multi-scene video remains consistent across the video frames. For instance, the room should not change without a change in text description.

Text Adherence. This metric aims to test whether the generated video adheres to the multi-scene text description. For instance, the events and actions described in the text script should be presented in the video accurately, and in the correct temporal order.

In our experiments, we compute the visual consistency and text adherence with the human and automatic evaluators. Further, we compute the overall score, which is the average of the visual consistency and text adherence scores. In addition, we also assess the visual quality of the generated videos using human evaluation to understand whether the video contains any flimsy frames, shaky images, or undesirable artifacts (Appendix D).

4.2 Task Prompts

Here, we curate task prompts for diverse scenarios to holistically assess the quality of the videos.

Single character in multiple visual contexts (S1). In this scenario, we instruct an LLM, GPT-4 [2], to create a coherent script consisting of four scenes. Each scene features a specific animal character performing diverse activities in every scene. This task assesses the capability of the T2V model to generate consistent appearance of the entity and its background while adhering to the different actions (or events) described in the multi-scene text script.

Different characters in a specific visual context (S2). In this scenario, we instruct a language model, GPT-4, to create a coherent script consisting of four scenes. Each scene features different animal characters engaging in the same activity in every scene [45]. This task assesses the capability of the T2V model to generate consistent appearance of the background while adhering to the appearance of the different characters in the multi-scene text script.

Multi-scene captions from real videos (S3). Here, we aim to assess the ability of the model to generate multi-scene videos for open-ended prompts that are derived from real-world videos. This task also assesses the ability of the T2V model to generate consistent appearances of the entity and its background while adhering to multi-scene descriptions. Specifically, we use our multi-scene video-text data generation pipeline (§3.2) to create such prompts for the real videos from the test splits of the video-text datasets. In total, we generate 100 prompts in this scenario. We present a sample of the various task prompts in the Appendix §F. We present the details about the human and automatic evaluators in §B.

4.3 Evaluation Setup

We compute the performance of the baselines and TALC by averaging the scores assigned to videos generated for two, three, and four scenes. Additionally, we report on visual consistency by averaging the performance across the entity and background consistency metrics. Here, the entity consistency scores are calculated for the task prompts S1 and S3 (since S2 aims to change the characters across scenes), and the background consistency and text adherence scores are computed for all the task prompts. We also evaluate the impact of TALC-based finetuning for single-scenes in Appendix §M.

5 Experiments

5.1 Text-to-Video Generative Models

We perform most of our experiments on ModelScope [49] due to its easy-of-access and adoption in prior works [30]. In addition, we also include Lumiere-T2V, a model that leverages space-time U-Net denoising networks to generate high-quality videos. In this work, we include early experiments with Lumiere to showcase the flexibility of the TALC approach on diverse models. We perform human evaluation for ModelScope, and automatic evaluation for both ModelScope and Lumiere.

Base model with TALC. As described in §3.1, our approach modifies the traditional text-conditioning mechanism to be aware of the alignment between text descriptions and individual video scenes. By design, the TALC framework can be applied to the base T2V model during inference, without any multi-scene finetuning. Here, we generate 16 frames per scene from ModelScope and 80 frames per scene from Lumiere. We provide more details in Appendix §J.

Finetuning with TALC. Since the base model is pretrained with single-scene data, we aim to show the usefulness of TALC framework when we have access to the multi-scene video-text data. To this end, we finetune ModelScope on the multi-scene video-text data (§3.2) with TALC framework. As a pertinent baseline, we also finetune the ModelScope without TALC framework by naively merging the scene-specific captions in the raw text space. In this setting, we finetune the T2V model with 8 frames per scene and the maximum number of scenes in an instance is set to 4. We provide further details on the finetuning setup in Appendix §L. The inference settings are identical to the prior method of generating videos from the base model without finetuning.

In this section, we present the results for the baselines and TALC framework averaged over a diverse task prompts and multiple scenes using automatic evaluation (§5.4) and human evaluation (§5.3). Finally, we provide qualitative examples for the multi-scene generated videos to showcase the usefulness of our approach (§5.5).

5.2 Baselines

Merge Captions In this setup, we create a single caption by merging all the multi-scene descriptions. While this approach mentions the temporal sequence of the events in a single prompt, the T2V model does not understand the temporal boundaries between the two events. Specifically, the visual features for all the frames will aggregate information from the entire multi-scene description, at once, without any knowledge about the alignment between the scene description and its expected appearance.

Merge Videos In this setup, we generate videos for each scene description individually and merge them in the raw input space. In this process, the parts of the multi-scene video closely adhere to the scene descriptions, leading to high text fidelity. However, since the generated videos do not have access to all the multi-scene descriptions (e.g., the video for Scene 2 is not informed about Scene 1), the visual consistency across the entire video is quite poor.

FreeNoise We also consider a more sophisticated approach, FreeNoise [38]. Specifically, it reschedules a sequence of noises for long-range correlation and perform temporal attention over them by window-based fusion. In addition, it includes motion injection method to support the generation of videos conditioned on multiple text prompts. Due to its complexity, this method involves setting various hyperparameters, including the denoising step at which motion injection should be activated, the cross-attention layers where prompts are injected, and the frames between which the prompt representations should be interpolated. In this work, we evaluate the publicly available implementation with VideoCrafter [13] on our unseen prompts.[3] While Gen-L-Video [48] has been shown to perform worse than FreeNoise, we conduct a qualitative analysis to emphasize the robustness of our method against it. (§5.5).[4]

[3] https://github.com/AILab-CVC/FreeNoise

[4] Other approaches such as VideoDirectorGPT [30] and Phenaki [45] are not publicly available.

5.3 Human Evaluation

TALC achieves the best performance in human evaluation. We compare the performance of the baselines and TALC framework using human evaluation in Table 1. We find that TALC applied to ModelScope (MS) outperforms FreeNoise by achieving a relative gain of 21% on the overall score. In addition, finetuning MS with TALC framework increases the video generation capability, leading to relative gain of 27% over FreeNoise in the overall score. The poor performance of FreeNoise can be attributed to the inability to introduce new content and limited motions as the multiple scenes (> 2) are requested. We provide some qualitative examples for FreeNoise to highlight its limitations in Appendix N. In addition, we find that using TALC framework in the base model outperforms the merging captions and merging video methods with the base model by 7.6 points and 11.1 points, respectively, on the overall score. Further, we note that TALC-finetuned model outperforms the merging captions and merging video methods with the base model by 12 points and 15.5 points, respectively, on the overall score.

Additionally, we observe that the merging captions with the base model achieves the highest visual consistency score of 96.5 points while it is the lowest for merging videos generated from the base model. Our results indicate that merging videos independently for the individual scene descriptions does not preserve the background and entity appearances across the different frames. Further, we note that the text adherence of the TALC-finetuned and TALC-base model is better than merging captions-finetuned and merging captions-base model, respectively. The high text adherence for merging videos can be attributed to its design where individual video scenes adhere to the scene-specific descriptions well. Overall, our empirical findings highlight that our simple framework can enable robust multi-scene video generation with finetuning. We present the results for the visual quality metric in Appendix D.

Table 2: **Automatic evaluation results for ModelScope.** We compare the performance of the baselines and TALC framework using the automatic evaluation for ModelScope generative model. Similar to the human evaluation, we observe that our simple approach achieves the best overall score with the base model (TALC) and finetuned model (Finetuning w/ TALC). We abbreviate visual consistency as VC, and text adherence as TA.

	Overall score	**VC**	**TA**
Merge captions	61.7	**91.0**	32.4
Merge videos	67.5 (+9.4%)	65.0	**70.0**
Finetuning w/ Merge captions	57.3 (−7.1%)	77.0	37.5
TALC	68.6 (+11.2%)	89.9	47.2
Finetuning w/ TALC	**75.6** (+22.5%)	89.0	62.3

Table 3: **Automatic evaluation results for Lumiere.** We present the results for the comparison between the baselines and TALC for the Lumiere video generative model. Specifically, our results indicate that TALC achieves the highest overall score, by achieving 7.1% relative gains over merge captions on the overall score. This indicates that TALC is a flexible strategy that can be applied to diverse video generative models. We abbreviate visual consistency as VC, and text adherence as TA.

	Overall score	**VC**	**TA**
Merge captions	64.4	94.7	34
Merge videos	66.5 (+3.2%)	68.0	**65.0**
TALC	**69.0** (+7.1%)	**97.8**	40.0

5.4 Automatic Evaluation

We compare the performance of the baselines with the TALC framework for ModelScope and Lumiere using the automatic evaluation in Table 2 and 3, respectively.

TALC outperforms the baselines without any finetuning. In Table 2, we find that the overall score, average of visual consistency and text adherence, of the multi-scene videos generated using the base ModelScope with TALC (68.6 points), outperforms the overall score achieved by the videos generated using merging captions (61.7 points) and merging videos (67.5 points) with the base ModelScope. In addition, we observe that the text adherence using TALC outperforms merging captions by 14.8 points, while the text adherence is the highest with a score of 70 points using merging videos. In Table 3, we observe similar trends for the Lumiere T2V generative model. Specifically, we find that the overall score for TALC outperforms merging captions and merging videos by 4 points and 2 points, respectively. In addition, we observe that merging captions and TALC achieve a high visual consistency score while merging videos independently has poor visual consistency. Further, we find that TALC outperforms merging captions by 5 points on text adherence, while merging videos achieves the highest text adherence 65 points. This highlights that the model more easily generates multi-scene videos that adhere to individual text scripts, whereas adherence to the text diminishes when the model is given descriptions of multiple scenes all at once.

Finetuning with TALC achieves the best performance. In Table 2, we find that finetuning with TALC achieves the highest overall score of 75.6 points in comparison to all the baselines. Specifically, we observe that the visual consistency does not change much with finetuning using the TALC method (89.9 points vs 89 points). Interestingly, we observe that finetuning with merging captions reduces the visual consistency by a large margin of 14 points. This can be attributed to the lack of knowledge about the natural alignment between video scenes and individual scene descriptions, which gets lost during the merging of captions. Additionally, we find that the text adherence of the TALC-finetuned model is 15.1 points more than the text adherence of the TALC-base model. This highlights that finetuning a T2V model with multi-scene data helps the most with its text adherence capability.

Fine-grained Results. To perform fine-grained analysis of the performance, we assess the visual consistency and text adherence scores for the baselines and TALC framework across diverse task prompts and number of scenes on ModelScope. We present their results in Appendix §I. In our analysis, we find that finetuning with TALC achieves the highest overall score over the baselines across all the scenarios. In addition, we notice that the highest performance is achieved in the scenario that consist of the different entities in a specific visual context. Further, we observe that the performance of the all the methods reduces when the task prompts get more complex i.e., multi-scene captions from real videos. In addition, we observe that finetuning with TALC achieves the highest overall score over the baselines across all the number of scenes. Specifically, we observe that the performance of the merging captions and TALC framework reduces as the number of scenes being generated increases. Overall, we show that the TALC strikes a good balance between visual consistency and text adherence to generate high-quality multi-scene videos.

5.5 Qualitative Analysis

We provide qualitative examples of generating multi-scene videos using TALC, FreeNoise and Gen-L-Video in Appendix Figure 10 and Figure 11. Our analysis demonstrates that all methods are capable of generating multi-scene videos that exhibit a high degree of text adherence. However, the primary distinction lies in the quality of the videos. FreeNoise is capable of producing videos characterized by superior quality and visual coherence. Nonetheless, the motion within these videos is notably limited, resulting in a relatively static scene. Gen-L-Video can generate videos that incorporate motion. However, visual consistency is not maintained throughout the video. For instance, in the provided example depicting a man engaged in both surfing and skiing, the man's appearance undergoes noticeable changes across the video. Overall, we observe that TALC is capable of generating realistic multi-scene videos that not only exhibit high textual adherence but also maintain visual consistency. Furthermore, the transitions between scenes are smooth and natural.

6 Conclusion

We introduced TALC, a simple and effective method for improving the text-to-video (T2V) models for multi-scene generation. Specifically, it incorporates the knowledge of the natural alignment between the video segments and the scene-specific descriptions. Further, we show that TALC-finetuned T2V model achieve high visual consistency and text adherence while the baselines suffer from one or both of the metrics. Given its design, our framework can be easily adapted into any diffusion-based T2V model. An important future direction will be to scale the amount of multi-scene video-text data and deploy TALC framework during pretraining of the T2V models.

7 Acknowledgement

We would like to thank Ashima Suvarna for providing feedback on the draft. Hritik Bansal is supported in part by AFOSR MURI grant FA9550-22-1-0380.

References

[1] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Towards high resolution video generation with progressive growing of sliced wasserstein gans. *arXiv preprint arXiv:1810.02419*, 2018. 1, 14

[2] 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. 2, 6

[3] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 2

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 1, 4

[5] Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. Videocon: Robust video-language alignment via contrast captions. *arXiv preprint arXiv:2311.10111*, 2023. 5, 14

[6] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 1, 2, 3, 4

[7] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 14

[8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 1, 2

[9] T Brooks, B Peebles, C Homes, W DePue, Y Guo, L Jing, D Schnurr, J Taylor, T Luhman, E Luhman, et al. Video generation models as world simulators, 2024. 1

[10] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024. 1, 14

[11] Emanuele Bugliarello, H Hernan Moraldo, Ruben Villegas, Mohammad Babaeizadeh, Mohammad Taghi Saffar, Han Zhang, Dumitru Erhan, Vittorio Ferrari, Pieter-Jan Kindermans, and Paul Voigtlaender. Storybench: A multifaceted benchmark for continuous story visualization. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[12] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 14

[13] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 1, 4, 7

[14] 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. *arXiv preprint arXiv:2401.09047*, 2024. 1

[15] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 5

[16] Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023. 3

[17] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 14

[18] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 1, 2, 4

[19] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024. 14

[20] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[21] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 14

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[23] Susung Hong, Junyoung Seo, Heeseong Shin, Sunghwan Hong, and Seungryong Kim. Direct2v: Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv preprint arXiv:2305.14330*, 2023. 14

[24] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 14

[25] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023. 5, 19

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 19

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[28] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 1, 2, 14

[29] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6329–6338, 2019. 2

[30] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023. 2, 5, 7, 14

[31] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videodrafter: Content-consistent multi-scene video generation with llm. *arXiv preprint arXiv:2401.01256*, 2024. 2

[32] Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: General-purpose video diffusion transformers via mask modeling. *arXiv preprint arXiv:2305.13311*, 2023. 14

[33] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022. 3

[34] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 14

[35] OpenAI. Gpt-4v(ision) system card, 2023b. https://openai.com/research/gpt-4v-system-card, 2023. 14

[36] OpenAI. Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators, 2024. 1

[37] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[38] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 2, 3, 6, 7, 20

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[40] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2493–2502, 2023. 2

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3

[43] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 5

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[45] 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 description. *arXiv preprint arXiv:2210.02399*, 2022. 2, 6, 7, 14

[46] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022. 14

[47] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016. 1, 14

[48] 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. 7, 14

[49] 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, 3, 4, 7, 14

[50] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. 1

[51] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 14

[52] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 14

[53] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 14

[54] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022. 1

[55] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023. 2

[56] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469, 2023. 2

[57] Zangwei Zheng, Xiangyu Peng, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024. 1

A Related Work

Text-to-Video Generative Modeling: Diffusion models like Imagen Video [21] represent a significant advancement in T2V synthesis, yet generating multi-scene videos that realistically capture the complexity of the physical world [1, 47, 10] remains challenging. Recent research has attempted longer video generation, but limitations persist. Phenaki [45] targets arbitrary-length videos with a focus on temporal coherence, though its evaluation is constrained by the lack of publicly available code. VideoDirectorGPT [30], DirecT2V [23], and Free-bloom [24] employ zero-shot approaches for multi-scene generation, contrasting with our fine-tuned method for enhanced performance. Gen-L-Video [48] uses iterative denoising to create consistent videos by aggregating overlapping short clips, but it typically generates videos with a maximum of two scenes, whereas our method supports up to four distinct scenes. StreamingT2V [19] employs an auto-regressive, streaming-based approach for continuous video generation. While these methods contribute to long video generation, they often struggle with maintaining visual consistency and adherence to multi-scene textual descriptions, which are critical for storytelling.

Multi-Scene Video Generation: Efforts like Phenaki [45] and Stable Video Diffusion [7] push the boundaries of text-driven generation by scaling latent diffusion models. Dreamix [34] and Pix2Video [12] utilize diffusion models for video editing and animation, while methods like MCVD [46] and VDT [32] focus on improving temporal consistency in longer videos. Despite these advances, generating multi-scene videos that accurately reflect complex narratives with high visual fidelity remains difficult, as shown by ongoing research in projects like VideoPoet [28], ModelScope [49], and Make-A-Scene [17]. TALC addresses the challenges of multi-scene video generation by enhancing visual consistency and text adherence across scenes. Unlike zero-shot methods, TALC leverages fine-tuning on a curated multi-scene video dataset. We also propose a comprehensive evaluation protocol, combining human evaluation and automated assessments using GPT-4V, to ensure robust narrative coherence in the generated videos.

B Evaluator

Human Evaluation. Here, we use the annotators from Amazon Mechanical Turk (AMT) to provide their judgements for the generated videos. Specifically, we choose the annotators that pass a preliminary qualification exam. Subsequently, they assess the multi-scene generated videos along the dimensions of entity and background consistency, text adherence, and visual quality. For each metric, the multimodal model assigns one of three possible response $\{yes = 1, partial = 0.5, no = 0\}$. For instance, yes for the entity consistency metric implies that the video frames sampled from the generated video have consistent appearance of the entity described in the multi-scene script. We present the screenshot of the UI in Appendix §H.

Automatic Evaluation. Here, we utilize the capability of a large multimodal model, GPT-4-Vision [35], to reason over multiple image sequences. First, we sample four video frames, uniformly, from each scene in the generated video (e.g., 8 videos frames for two-scene video). Then, we prompt the multimodal model with the temporal sequence of video frames from different scenes and the multi-scene text description. The model is instructed to judge the generated videos for visual consistency and text adherence, similar to the human evaluation. In this work, we do not utilize any existing video-text alignment models [52, 5] for evaluating text adherence as they are trained on single-scene video-text datasets. We find that the agreement between the automatic evaluation and human evaluation is 77$\%$. We present the automatic evaluation prompt in Appendix §G.

C Multi-scene video data sources

To construct a multi-scene video-text dataset, we utilize existing dataset that include natural (real) videos and associated high-quality human-written captions that summarize the entire video. Specifically, we choose MSR-VTT [53] and VaTeX [51]. Most of the videos in MSR-VTT are 10-30 seconds long while VaTeX consists 10 seconds long videos. In addition, each video in MSR-VTT and VaTeX consists 20 captions and 10 captions, respectively, out of which one is selected at random for multi-scene data generation. As described above, a single video is cut into multiple video segments using Pyscene library. In our experiments, we retain the first four video segments and discard any

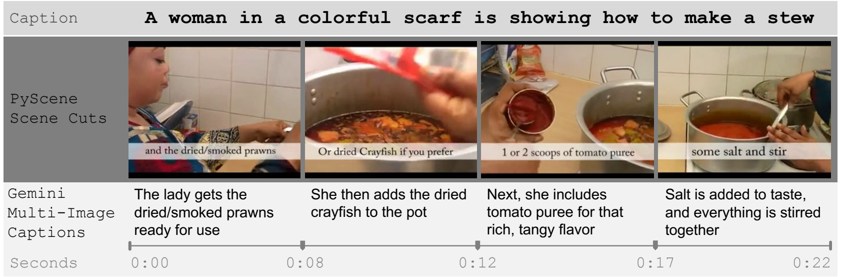

Caption
A woman in a colorful scarf is showing how to make a stew
PyScene Scene Cuts
and the dried/smoked prawns
Or dried Crayfish if you prefer
1 or 2 scoops of tomato puree
some salt and stir
Gemini Multi-Image Captions
The lady gets the dried/smoked prawns ready for use
She then adds the dried crayfish to the pot
Next, she includes tomato puree for that rich, tangy flavor
Salt is added to taste, and everything is stirred together
Seconds
0:00
0:08
0:12
0:17
0:22

Figure 4: **Our approach for generating time-aligned video captions.** The process begins with PyScene cuts identifying the boundaries of distinct scenes within a video. Keyframes are then selected from the median of each scene. These frames are processed collectively through the Gemini model to produce multi-image captions that maintain narrative continuity by contextualizing each scene within the video's overall sequence.

additional segments if the library generates more than four. Since the accuracy of the multi-scene captioning and the computational demands during finetuning are influenced by the number of scenes, we opt to limit the scene count to four for our experiments. However, future work could employ similar methodologies to scale the number of scenes, given more computing power and advanced multi-scene captioning models. We provide the data statistics for the final multi-scene data in Appendix §K.

D Visual quality of the generated videos.

Table 4: **Human evaluation results on the visual quality of the generated videos from ModelScope.** We observe that the visual quality of the generated videos are close to each other for the base model. However, finetuning the model with merging captions reduces the video quality by a large margin while TALC-finetuned model retains the video quality.

Method	Quality
FreeNoise	80
Merge captions	80.5
Merge videos	86.5
Finetuning w/ Merge captions	63.4
TALC	84.5
Finetuning w/ TALC	83.3

We compare the visual quality of the generated videos using human evaluation in Table 4. We find that the visual quality of videos generated from the base model ranges from $80 - 86.5$ using the baselines and TALC framework. However, we observe that the visual quality of generated videos is quite poor for the model finetuned with merging captions with a score of 63.4 points. This highlights that finetuning a T2V model with multi-scene video-text data by naively merging the scene-specific descriptions in the raw text space leads to undesirable artifacts in the generated video. Finally, we find that the TALC-finetuned model (83.3) achieves a video quality score similar to that of the TALC-base model (84.5), indicating that our finetuning data preserves the visual quality observed during the model's pretraining.

E Prompt for Multi-Scene Caption Generation

We present the prompt used to generate multi-scene captions using large multimodal model, Gemini-Pro-Vision, in Figure 5. In particular, we utilize Gemini-Pro-Vision since it can reason over multiple image sequences. Specifically, we provide the multimodal model with a video frame from each of the segmented videos and the single caption for the entire video present in the original video-text datasets.

Your task is to create captions for a series of images, each taken from different video scenes. For every image shown, craft a 7-10 word caption (7-10 words) that precisely describes what's visible, while also linking these captions into a fluid, engaging story. A common caption will be given to help guide your narrative, ensuring a smooth transition between scenes for a cohesive story flow. Remember to not hallucinate in your responses.

Common Caption: {caption}

Figure 5: Prompt to generate multi-scene caption using large multimodal models.

F Task Prompts

F.1 Single character in multiple visual contexts

We present the prompt used to generate multi-scene text descriptions for single character in multiple visual contexts from GPT-4 in Figure 6.

Create four concise continuous movie scenes (7-10 words) focusing on a specific real-world character. The scenes should form a cohesive narrative.

Guidelines:

Choice of Character: Select a real-world animal as the focal point of your scenes.
Scene Description: Clearly describe the setting, actions, and any notable elements in each scene.
Connection: Ensure that the scenes are logically connected, with the second scene following on from the first.
Brevity and Precision: Keep descriptions short yet vividly detailed.

Example:

Character: polar bear
Scene 1: A polar bear navigates through a icy landscape.
Scene 2: The polar bear hunts seals near a crack in the ice.
Scene 3: The polar bear feasts on the seal.
Scene 4: The polar bear curls up for a nap.

Now it's your turn.

Figure 6: GPT-4 Prompt to generate multi-scene prompts for single character in multiple visual contexts.

F.2 Different characters in a specific visual context

We present the prompt used to generate multi-scene text descriptions for different characters in a specific visual context from GPT-4 in Figure 7.

Create four concise scene descriptions (7-10 words) where different characters perform identical action/events.

Choice of Characters: Select four real-world animals as the focal point of the individual scenes.
Background Consistency: Ensure that the background is consistent in both the scenes.
Brevity and Precision: Keep descriptions short yet vividly detailed.

Example:
Characters: teddy bear, panda, grizzly bear, polar bear
Scene 1: A teddy bear swims under water.
Scene 2: A panda swims under the same water.
Scene 3: A panda swims under the same water.
Scene 4: A panda swims under the same water.

Now it's your turn.

Figure 7: GPT-4 Prompt to generate multi-scene prompts for different characters in a specific visual context.

G Automatic Evaluation Prompt

We present the prompt used to perform automatic evaluation of the multi-scene generated videos using large multimodal model, GPT-4-Vision, in Figure 8. We utilize GPT-4-Vision for automatic evaluation since it can reason over multiple images. Specifically, we provide the multimodal model with four video frames for each scene in the generated video. The model has to provide its judgments based on the entity consistency, background consistency, and text adherence of the video frames.

You are a capable video evaluator. You will be shown a text script with two-scene descriptions where the events/actions . Video generating AI models receive this text script as input and asked to generate relevant videos. You will be provided with eight video frames from the generated video. Your task is to answer the following questions for the generated video.
1. Entity Consistency: Throughout the video, are entities consistent? (e.g., clothes do not change without a change described in the text script)
2. Background Consistency: Throughout the video, is the background consistent? (e.g., the room does not change described in the text script)
3. Text Adherence: Does the video adhere to the script? (e.g., are events/actions described in the script shown in the video accurately and in the correct temporal order)

Respond with NO, PARTIALLY, and YES for each category at the end. Do not provide any additional explanations.

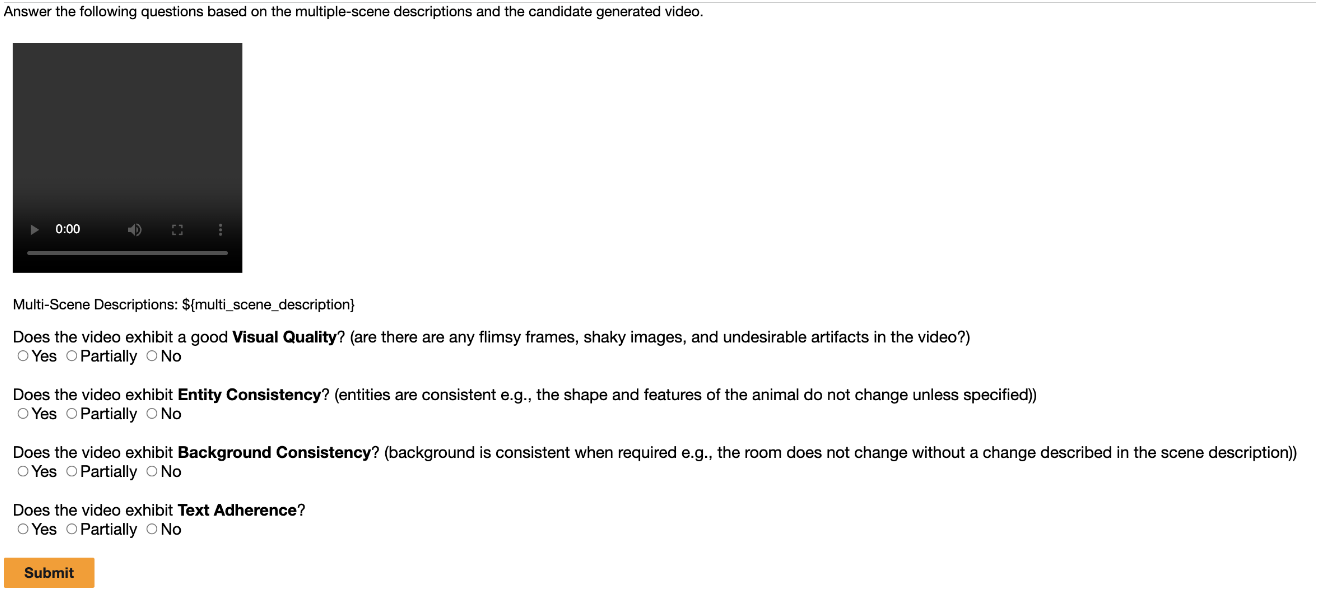
Two-scene descriptions:

Scene 1: {scene1}
Scene 2: {scene2}

Figure 8: Prompt used to perform automatic evaluation of the multi-scene generated videos. We use this prompt when the number of scenes in the task prompt is two.

H Human Evaluation Screenshot

We present the screenshot for the human evaluation in Figure 9. Specifically, we ask the annotators to judge the visual quality, entity consistency, background consistency, and text adherence of the multi-scene generated videos across diverse task prompts and number of scenes.



Answer the following questions based on the multiple-scene descriptions and the candidate generated video.

Multi-Scene Descriptions: ${multi_scene_description}

Does the video exhibit a good **Visual Quality**? (are there are any flimsy frames, shaky images, and undesirable artifacts in the video?)
○ Yes ○ Partially ○ No

Does the video exhibit **Entity Consistency**? (entities are consistent e.g., the shape and features of the animal do not change unless specified))
○ Yes ○ Partially ○ No

Does the video exhibit **Background Consistency**? (background is consistent when required e.g., the room does not change without a change described in the scene description))
○ Yes ○ Partially ○ No

Does the video exhibit **Text Adherence**?
○ Yes ○ Partially ○ No

Submit

Figure 9: Human Annotation Layout

I Fine-grained Analysis

We present the automatic evaluation results across task prompts (§I.1) and number of scenes (§I.2).

I.1 Task Prompts

We compare the performance of the baselines and TALC framework across different task prompts in Table 5. We find that the TALC-finetuned model achieves the highest overall score over all the baselines. Specifically, we find that the TALC framework achieves a high visual consistency with scores close to the merging captions baseline. Further, we observe that the TALC framework achieves a higher text adherence in comparison to the merging captions, with or without finetuning, across all the task prompts.

Table 5: **Automatic evaluation results across task prompts.** Here, S1 refers to the single character in multiple visual contexts. S2 refers to the different characters in a specific visual context. S3 refers to the multi-scene captions from real videos. We abbreviate Finetuning as F.T., Visual consistency as V.C., Text adherence as T.A.

Method	S1			S2			S3		
	V.C.	T.A.	Overall	V.C.	T.A.	Overall	V.C.	T.A.	Overall
Merge captions	95.9	43.8	69.9	99.5	21.5	60.5	81.9	32.0	56.9
Merge videos	69.4	71.5	70.5	71.2	82.3	76.7	57.9	58.7	58.3
F.T. w/ merge captions	94.2	57.1	75.7	94.9	31.2	63.1	50.8	24.3	37.5
TALC	93.6	48.3	71.0	99.3	57.3	78.3	81.4	36.1	58.8
F.T. w/ TALC	93.3	62.6	**77.9**	98.9	76.3	**87.6**	79.7	47.9	**63.8**

I.2 Number of Scenes

We compare the performance of the baselines and TALC framework across different number of generated scenes in Table 6. We find that the TALC-finetuned model outperforms all the baselines in this setup. In addition, we find that the visual consistency of the TALC framework does not change much with the number of the scenes. However, we find that the text adherence of the baselines and TALC framework reduces with the number of generated scenes. The text adherence scores of the merging videos does not change with the number of scenes as it generates the videos for the individual scenes independently.

Table 6: **Automatic evaluation results across different number of scenes in the task prompts.** We abbreviate Finetuning as F.T., visual consistency as V.C., and Text Adherence as T.A.

	# scenes = 2			# scenes = 3			# scenes = 4		
	V.C.	T.A.	Overall	V.C.	T.A.	Overall	V.C.	T.A.	Overall
Merge captions	93.2	34.4	63.8	92.5	33.0	62.7	87.4	29.9	58.6
Merge videos	66.7	69.9	68.3	65.2	71.9	68.6	63.5	70.7	67.1
F.T. w/ merge captions	87.7	45.7	66.7	83.2	39.5	61.4	60.0	27.4	43.7
TALC	92.6	54.4	73.5	89.8	48.0	68.9	87.3	39.3	63.3
F.T. w/ TALC	88.5	66.6	**77.5**	90.7	64.1	**77.4**	87.8	56.1	**71.9**

J Inference Details

We provide the details for sampling multi-scene videos from the ModelScope and Lumiere T2V models in Table 7 and Table 8, respectively.

Table 7: Sampling setup for ModelScope T2V model.

Resolution	256×256
Number of video frames per scene	16
Guidance scale	12
Sampling steps	100
Noise scheduler	DPMSolverMultiStepScheduler

add FreeNoise details too

K Multi-Scene Data Statistics

We provide the details for the multi-scene video-text dataset in Table 9.

Table 8: Sampling setup for Lumiere T2V model.

Resolution	1024×1024
Number of video frames per scene	80
Guidance scale	8
Sampling steps	256
Noise scheduler	DPMSolverMultiStepScheduler

Table 9: Multi Scene Video-Text Data Statistics

Number of entire videos	7107
% of single scene	27.3%
% of two scenes	25.4%
% of three scenes	31.3%
% of four scenes	16.0%
Number of video scene-caption instances	20177

L Finetuning Details

We provide the details for finetuning ModelScope T2V model with TALC framework in Table 10.

Table 10: Training details for the TALC-finetuned ModelScope T2V model.

Base Model	ModelScope [5]
Trainable Module	UNet
Frozen Modules	Text Encoder, VAE
Batch size	20
Number of GPUs	5 Nvidia A6000
Resolution	256×256
Crop	CenterCrop
Learning Rate Scheduler	Constant
Peak LR	1.00E-05
Warmup steps	1000
Optimizer	Adam (0.9, 0.999, 1e-2, 1e-8) [26]
Max grad norm	1
precision	fp16
Noise scheduler	DDPM
Number of frames per video scene	8
prediction type	epsilon

M Single-Scene Video Generation

To ascertain that our model's new multi-scene generation function does not detract from its single-scene generation performance, we conducted a series of evaluations using the VBench framework [25]. VBench offers a robust analysis of various video generation aspects such as adherence to text prompts, stylistic integrity, semantic coherence, and overall aesthetic quality.

Our analysis, shown in Table 11, establishes a refined baseline: ModelScope (Single-Scene Fine-tuning), fine-tuned on single-scene video generation data. This process yielded an average score of 0.48, indicating a decrease from the base ModelScope's average score of 0.63. This suggests that the optimizations in the base model, such as integrating high-quality images, are not fully utilized in single-scene fine-tuning.

Interestingly, fine-tuning the model on multi-scene data (ModelScope - Multi-Scene Finetuning) resulted in improved performance with an average score of 0.59, surpassing the single-scene fine-tuned version. This indicates that multi-scene data enriches the model's understanding of video content, enhancing both multi-scene and single-scene video generation.

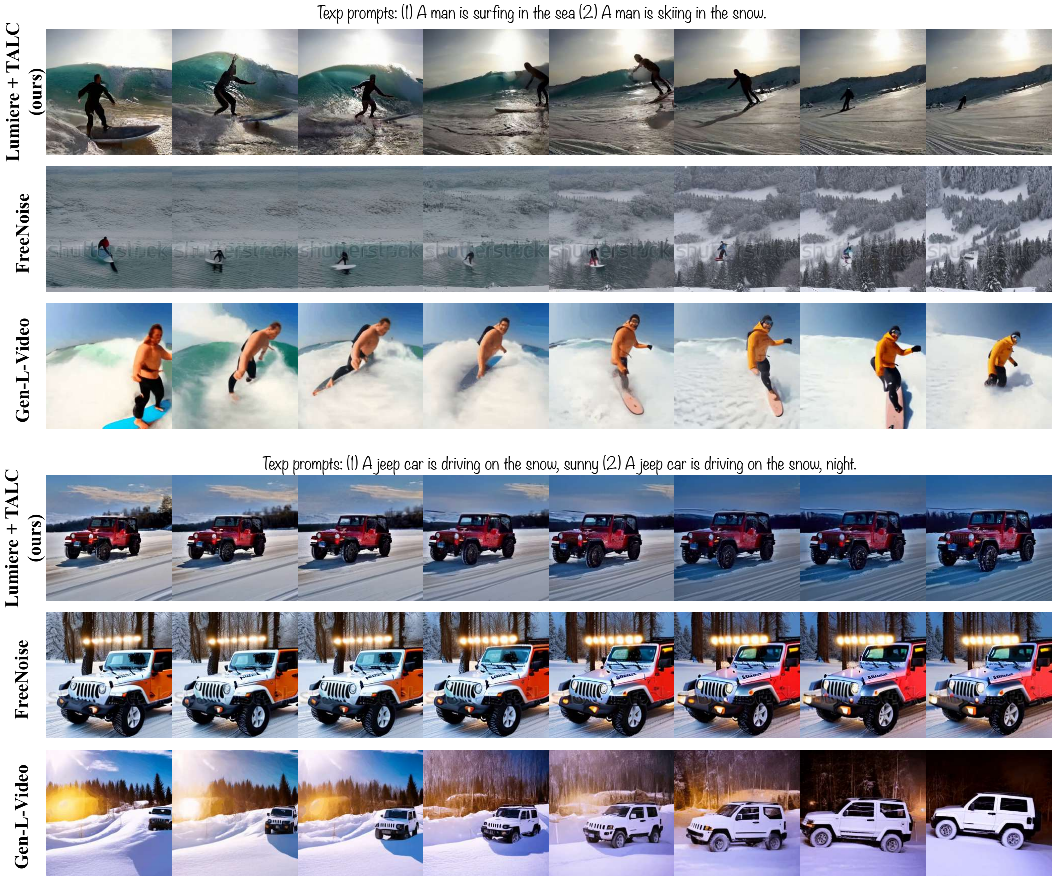

Texp prompts: (1) A man is surfing in the sea (2) A man is skiing in the snow.
Lumiere + TALC (ours)
FreeNoise
Gen-L-Video
Texp prompts: (1) A jeep car is driving on the snow, sunny (2) A jeep car is driving on the snow, night.
Lumiere + TALC (ours)
FreeNoise
Gen-L-Video

Figure 10: **Examples of videos generated by TALC with Lumiere, FreeNoise and Gen-L-Video.** The videos generated using TALC with Lumiere exhibited smooth motion, seamless transitions between text prompts, and high overall quality. In contrast, videos created with FreeNoise were of lower quality, with notably less motion. For example, in the second video, the background remained static except for darkening, and the jeep did not appear to be in motion. Videos generated with Gen-L-Video demonstrated less smooth transitions between prompts and, at times, showed inconsistencies in object representation, as seen in the first video where the man's appearance changed repeatedly.

This comparison highlights the importance of data curation and fine-tuning strategies, showing that our approach not only enables complex multi-scene narratives but also improves single-scene video generation.

N FreeNoise failure cases

In this work, we observe that the state-of-the-art method FreeNoise [38] does not perform well on our multi-scene prompts. We provide some qualitative examples to show its inability to introduce new content and limited motions as multiple scenes are requested in Figure 12, 13, and 14.

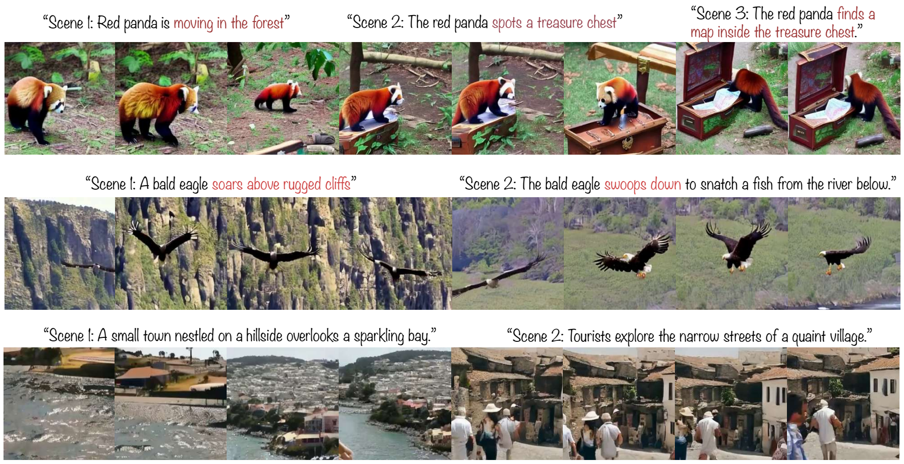

"Scene 1: Red panda is moving in the forest"
"Scene 2: The red panda spots a treasure chest"
"Scene 3: The red panda finds a map inside the treasure chest."
"Scene 1: A bald eagle soars above rugged cliffs"
"Scene 2: The bald eagle swoops down to snatch a fish from the river below."
"Scene 1: A small town nestled on a hillside overlooks a sparkling bay."
"Scene 2: Tourists explore the narrow streets of a quaint village."

Figure 11: **Qualitative examples of videos generated by TALC with ModelScope.** Videos generated with TALC allow for the creation of videos where a single object performs multiple actions while maintaining visual consistency. Furthermore, TALC facilitates the generation of videos with diverse scenes, as demonstrated in the final example.

Table 11: Single-Scene Evaluation Results using VBench, comparing the base model, when it is fine-tuned on multi-scene data, and single-scene data ('f.t.' stands for fine-tuned). Our analysis shows ModelScope (Single-Scene Finetuning) as a refined baseline with an average score of 0.48, compared to the base ModelScope's 0.63. Fine-tuning on multi-scene data (ModelScope - Multi-Scene Finetuning) yields an improved score of 0.59, highlighting the efficacy of multi-scene data in enhancing video generation performance.

Dimension	ModelScope (Base)	ModelScope (Multi-scene f.t.)	ModelScope (Single-Scene f.t.)
Appearance style	0.23	0.24	0.21
Color	0.85	0.83	0.78
Human action	0.96	0.92	0.75
Object class	0.86	0.77	0.42
Overall consistency	0.26	0.26	0.22
Spatial relationship	0.35	0.29	0.14
Subject consistency	0.90	0.83	0.75
Temporal flickering	0.97	0.89	0.86
Temporal style	0.26	0.25	0.21
Average	0.63	0.59	0.48

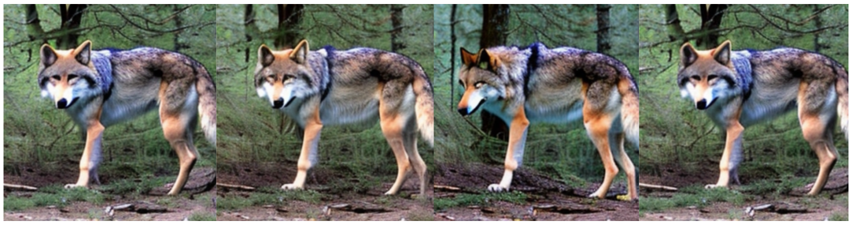


Figure 12: FreeNoise generation for the prompt: *"A wolf howls at the moon in a dense forest.; The wolf prowls stealthily through the underbrush, eyes gleaming.; The wolf catches a scent and follows it eagerly.; The wolf emerges victorious, holding a fresh kill."* **Explanation:** The generated scene shows a wolf standing in a forest, but the key actions described in the prompt are missing. The wolf does not prowl, its eyes are not gleaming, it does not follow any scent, and no kill is presented. This indicates a significant gap in the model's ability to convey the sequential and dynamic nature of the narrative.

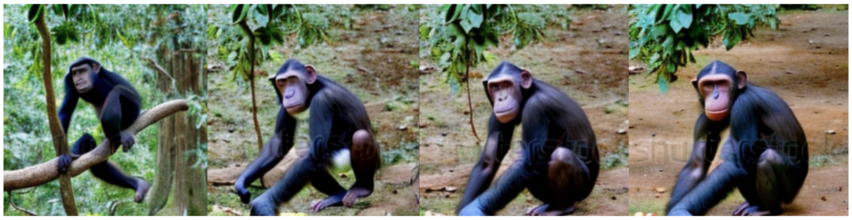


Figure 13: FreeNoise generation for the prompt: *"A chimpanzee swings effortlessly through the forest canopy.; The chimpanzee gathers fruit, chattering with companions."* **Explanation:** Instead of depicting the chimpanzee swinging through the canopy, the generated scene shows it sitting on a tree branch and then on the ground. Furthermore, the scene lacks the critical elements of the prompt, such as the presence of fruits and companions, which are essential for conveying the intended narrative.

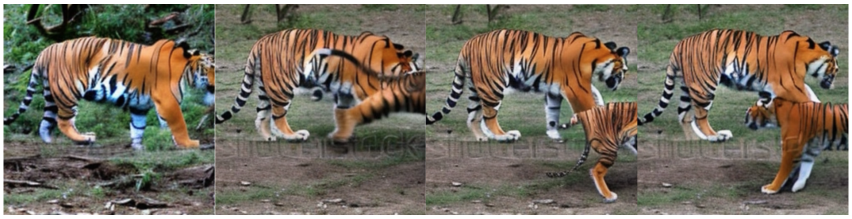


Figure 14: FreeNoise generation for the prompt: *"A tiger prowls through dense jungle undergrowth stealthily.; The tiger ambushes a deer, swift and deadly."* **Explanation:** The scene fails to show the tiger ambushing a deer as described. Instead, the tiger is depicted standing calmly, and no deer is visible. Additionally, a second tiger appears unexpectedly from the first tiger's body, abruptly changing directions, which introduces further inconsistencies and detracts from the intended narrative.