ShotAdapter: Text-to-Multi-Shot Video Generation with Diffusion Models

Ozgur Kara[1,2] Krishna Kumar Singh[2] Feng Liu[2] Duygu Ceylan[2]
James M. Rehg[1] Tobias Hinz[2]
[1]UIUC [2]Adobe

Project Webpage: https://shotadapter.github.io/

Figure 1. **ShotAdapter** is a lightweight framework that enables text-to-multi-shot video generation by fine-tuning a pre-trained text-to-video model. It allows control over number and duration of shots as well as shot content through shot-specific text prompts. The framework maintains character identity while being able to preserve backgrounds (*e.g.* 3rd row) or transition to new ones (*e.g.* 4th row), featuring distinct activities (*e.g.* playing guitar, then using laptop) and perspectives.

Abstract

Current diffusion-based text-to-video methods are limited to producing short video clips of a single shot and lack the capability to generate multi-shot videos with discrete transitions where the same character performs distinct activities across the same or different backgrounds. To address this limitation we propose a framework that includes a dataset collection pipeline and architectural extensions to video diffusion models to enable text-to-multi-shot video generation. Our approach enables generation of multi-shot videos as a single video with full attention across all frames of all shots, ensuring character and background consistency, and allows users to control the number, duration, and content of shots through shot-specific conditioning. This is achieved by incorporating a transition token into the text-to-video model to control at which frames a new shot begins and a local attention masking strategy which controls the transition token's effect and allows shot-specific prompting. To obtain training data we propose a novel data collection pipeline to construct a multi-shot video dataset from existing single-shot video datasets. Extensive experiments demonstrate that fine-tuning a pre-trained text-to-video model for a few thousand iterations is enough for the model to subsequently be able to generate multi-shot videos with shot-specific control, outperforming the baselines. You can find more details in our webpage.

1. Introduction

While diffusion models [18, 38] have shown impressive capabilities in the image domain [2, 12, 17, 33, 37, 39, 40, 51, 55, 59], extending them to video synthesis presents significant challenges due to the dynamic nature of videos. One group of researchers has explored various methods to adapt text-to-image (T2I) models for video synthesis [14,

21, 22, 41, 52]. In contrast, another group has focused on text-to-video (T2V) diffusion models [4, 20, 34] which demonstrate superior performance by processing the entire video as a unified input rather than handling each frame independently. Despite their state-of-the-art performance in video generation, all existing models are designed to generate a single continuous video, which imposes inherent limitations when attempting to generate a multi-shot video of the same character engaged in multiple distinct activities, where each shot is separated by discrete cuts in dynamic elements. This challenge becomes particularly significant when each activity requires a unique setting, even within the same background (Fig. 2, 1st column: a man writes code and then sketches diagrams on a whiteboard) or when the background needs to change while maintaining the same identity (Fig. 2, 2nd column: transitioning from a home gym to a park bench). Moreover, existing models are limited to generating videos of very short durations, and therefore lack flexibility. Because they have been trained exclusively on single-shot videos, they are ill-suited for multi-shot video synthesis. Consequently, these limitations hinder their applicability in real-world contexts, such as film production, where narratives rely on multiple shots, each featuring distinct actions or perspectives and often featuring the same character across diverse scenes and time frames.

The simplest approach to multi-shot synthesis is to generate a single-shot video by combining all shot-specific prompts into a single prompt. However, this method cannot create "cuts" (transitions between shots) and struggles to provide different backgrounds for the same character. Even within a same background, it cannot generate characters featuring different activities requiring distinct settings, as it is constrained to generating short videos (Fig. 2 (a)). An extension of this approach is to generate each shot many times individually, then concatenate the most similar shots, which requires the generation of numerous videos (Fig. 2 (b)). An improvement to this baseline entails first generating consistent keyframes of a character using a reference image, then animating these keyframes with an image-to-video (I2V) model, and finally concatenating the generated clips to produce a multi-shot video. However, this baseline is inevitably limited by the capabilities of off-the-shelf methods, as they struggle with maintaining consistency and quality. Additionally, it remains challenging to depict the character within the same background performing distinct activities (Fig. 2 (c)).

Given the limitations of potential baselines, we introduce `ShotAdapter`, a simple yet powerful, model-agnostic framework for controllable text-to-multi-shot video (T2MSV) generation. `ShotAdapter` transforms a single-shot T2V generator into a T2MSV generator with minimal fine-tuning, requiring only five thousand iterations (less than 1% of the total pre-training iterations), on a multi-shot video dataset. This transformation is made possible through a novel attention-layer masking strategy and a special learnable token, called "transition token", which signals the transition between shots. It enables multi-shot video generation where the camera perspective can shift abruptly between shots (Fig. 1, 2nd row) within a single background with the character performing different activities, or the background itself can change (Fig. 1, 4th row), all while preserving the character's identity. Notably, `ShotAdapter` offers users precise control over the content of each generated shot through shot-specific text prompts, along with the flexibility to specify the number and duration of shots. Additionally, the use of a unified input for the T2V model ensures consistency across shots with minimal fine-tuning on an appropriate multi-shot video dataset.

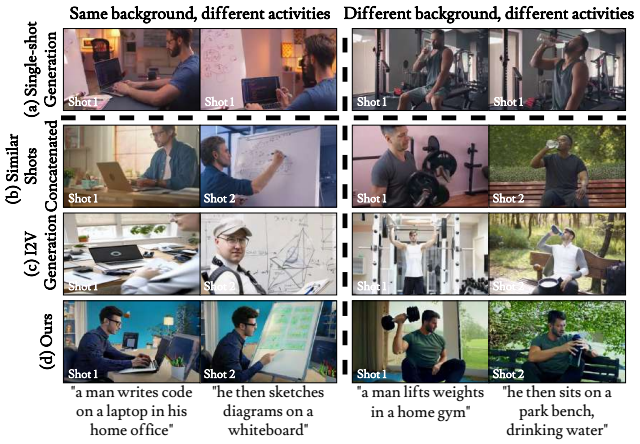


Figure 2. ***Comparison with baselines.*** Each row displays frames from generated 2-shot videos, guided by shot-specific text prompts. The left column shows results with same background and different activities, while the right column presents results both with different backgrounds and activities.

To address the lack of training datasets, we propose two novel pipelines to construct a multi-shot video dataset from a single-shot video dataset through pre- and post-processing steps, which eliminates the need to collect actual multi-shot video. Recognizing the lack of an evaluation standard for T2MSV generation, we propose an evaluation pipeline and several baseline comparisons to assess synthesis results based on identity and background consistency, text alignment, and shot-duration precision. In summary, our main contributions are as follows:

- We propose a model-agnostic and computationally efficient framework that transforms a T2V model into a T2MSV generator with minimal fine-tuning, ensuring the preservation of character identity across all generated shots.
- Our approach allows user-defined control over the number of shots, their duration, and the content of each shot through text prompts, leveraging a novel "transition token" coupled with a localized attention masking strategy.
- We propose two novel data collection strategies to con-

struct a multi-shot video dataset derived from single-shot video dataset, along with pre- and post-processing steps.
- We believe we are the first to frame the T2MSV generation task as a challenge for the research community. Therefore, we introduce an evaluation pipeline to assess performance in this domain and will release a validation dataset to promote standardized evaluation.

2. Related Works

Text guided video synthesis Text-guided video synthesis is a rapidly evolving field. Early methods adapted pre-trained text-to-image (T2I) diffusion models [38] for video synthesis by modifying architecture, such as adding temporal layers [52], altering latent inputs [22], and introducing temporally correlated noise for frame consistency [13]. However, T2I models struggle with frame coherence due to their lack of temporal processing. Recent approaches, such as T2V models using unified video inputs and Diffusion Transformer (DiT) frameworks, address this. OpenAI's SORA [4] and Gen-Tron [7] enhance DiT with large datasets and advanced text-conditioning. Commercial models like Luma Dream Machine [24], MiniMax [29], and Kling AI [1] achieve coherence through extensive training. MovieGen [34] enables instruction-based editing and personalized generation. Open-source contributions have advanced T2V methods, such as W.A.L.T [15], which uses a two-stage algorithm for training on image and video datasets, and Latte [28], which employs Transformer blocks. RIVER [9] and PyramidFlow [20] use autoregressive generation with flow-matching, while OpenSora [60] and OpenSora-Plan [23] aim to replicate SORA more efficiently. In contrast to prior works that focus on single-shot video generation, we introduce a framework for fine-tuning T2V models, transforming them into T2MSV generators.

Image/Video story generation Personalized identity-preserving image generation presents challenges in maintaining consistent identity across diverse settings. Techniques like DreamBooth [39] and Textual Inversion [12] achieve this through time-consuming test-time fine-tuning. Some line of work [27, 51, 53] use adapter models whereas other works [25, 49, 54, 56] incorporate ArcFace [10] or keypoints on face features for conditioning. In image story generation, StoryMaker [62] ensures stylistic consistency, while ConsiStory [44] enhances subject consistency through a shared attention block and feature injection. DreamStory [16] leverages Large Language Models, and StoryDiffusion [61] introduces an attention mechanism for consistent character representation. In video story generation, DreamBooth [39], adapted in Tune-A-Video [52], requires fine-tuning for each input video, limiting scalability and restricting outputs to short, single-shot videos based on reference images. Recognizing the lack of a unified solution for multi-shot video generation, one of our baselines combines StoryMaker [62] with an I2V model, though this integration shows limited performance. Our approach achieves superior multi-shot video generation results compared to this baseline.

3. Methodology

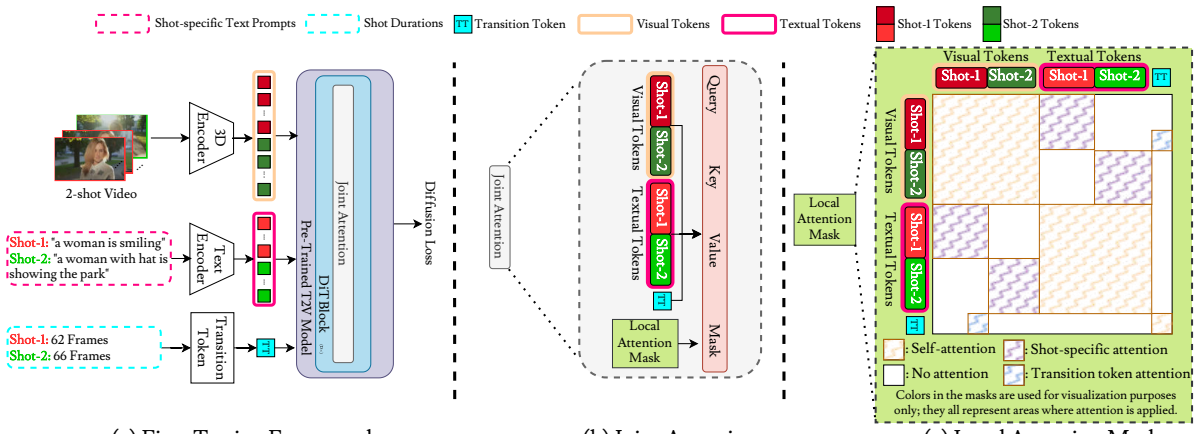
3.1. Preliminaries

Diffusion formulation Diffusion models [18, 38, 42] generate realistic data by learning to reverse a noise-adding process. During training, data is progressively noised over multiple time steps, and the model learns to denoise it step by step, starting from pure noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. A noisy version $\mathbf{x}_i^{(t)}$ at timestep t is obtained by adding scaled noise $\mathbf{x}_i^{(t)} = \sqrt{\alpha_t}\,\mathbf{x}_i + \sqrt{1-\alpha_t}\,\epsilon$, where α_t controls the noise level. The training loss $\mathcal{L}_{\text{diffusion}}(\theta) = \mathbb{E}_{\mathbf{x}_i,\epsilon,t}\left\|\epsilon - \epsilon_\theta(\mathbf{x}_i^{(t)}, t)\right\|^2$ minimizes the difference between the true noise ϵ and the model's prediction $\epsilon_\theta(\mathbf{x}_i^{(t)}, t)$ This loss trains the model to predict the noise, enabling it to reconstruct the original data through reverse diffusion.

Latent video diffusion models Given an input video sample $\mathbf{x}_i \sim p_{data}(\mathbf{x})$ with dimensions $\mathbf{x}_i \in \mathbb{R}^{F\times C\times H\times W}$, the process begins by encoding the video using a 3D encoder $\mathcal{E}$, yielding a latent representation $\mathbf{z}_i \in \mathbb{R}^{F'\times C'\times H'\times W'}$, where $F' = F/f_t$, $H' = H/f_s$, and $W' = W/f_s$, with f_s and f_t being the spatial and temporal compression ratios, respectively. The latent video representation is further transformed into a sequence of N tokens for a DiT-based model, denoted as $\{\bar{\mathbf{x}}_i^n\}_{n=1}^{n=N}$, where each token $\bar{\mathbf{x}}_i^n \in \mathbb{R}^D$ has a hidden dimension D. Patchification is performed along the width, height, and frame depth, patch sizes represented by f_{p_w}, f_{p_h}, and f_{p_f}, respectively.

3.2. Text-to-Multi-Shot Video Generation

Inline with the definitions in previous works [6, 45, 57], a **shot** is defined as the smallest segment of a video—a single, uninterrupted clip with continuous motion. Each shot is characterized by dynamic elements, including the foreground object, background setting, object actions, and camera movement. A **multi-shot video** is defined as a video composed of multiple individual shots, each being separated by a "cut" [45], which is an instantaneous change in the dynamic elements. More formally, an N-shot video V can be represented as $V = \{s_1^{K_1}, \ldots, s_N^{K_N}\}$, where each shot $s_i^{K_i}$ is a set of K_i frames, $s_i^{K_i} = \{f_1, \ldots, f_{K_i}\}$, with each frame $f_j \in \mathbb{R}^{C\times H\times W}$ representing an image with channel (C), height (H), and width (W). Our work focuses on multi-shot videos featuring a single foreground object, specifically humans, as they are often the main characters and present challenges in maintaining identity consistency for real-world applications. We aim to *preserve the identity of the foreground object across shots*, even as it performs



(a) Fine-Tuning Framework (b) Joint Attention (c) Local Attention Mask

Figure 3. ***Fine-tuning framework with transition token and local attention masking.*** (a) `ShotAdapter` fine-tunes a pre-trained T2V model by incorporating "transition tokens" (highlighted in light blue). We use $n-1$ transition tokens, initialized as learnable parameters, alongside an n-shot video with shot-specific prompts, which are fed through the pre-trained T2V model. (b) The model processes the concatenated input token sequence, guided by a "local attention mask" through joint attention layers within DiT blocks. (c) The local attention mask is structured to ensure that transition tokens interact only with the visual frames where transitions occur, while each textual token interacts exclusively with its corresponding visual tokens.

different activities. Furthermore, when the background is designated as a fixed location, we require background consistency throughout the entire video. More formally, our task is to generate an N-shot video given a set of input conditions $\{C_1, C_2, ...\}$ where each C_i denotes shot-specific conditions, *e.g.* $C_i = \{\text{shot caption}, \text{shot duration}, \cdots\}$.

`ShotAdapter` introduces model-agnostic novel extensions to transform pre-trained T2V diffusion models into T2MSV generators with minimal fine-tuning along with a multi-shot video dataset curation method. Our method allows users to control the number and duration of each shot as well as the content through shot-specific text prompts.

3.3. Model

Model architecture We use a diffusion transformer (DiT) [32] based T2V framework similar to OpenSora [60] and MovieGen [34], shown in Fig. 3 (a). Our model integrates a 3D Variational Autoencoder (3D-VAE) for video encoding, along with a variant of joint-attention layers [11] for conditioning. Specifically, input videos are first subject to temporal and spatial encoding via the 3D encoder before being patchified. The textual condition tokens are then concatenated with these patchified visual tokens and subsequently processed by the DiT. We leverage a pre-trained T2V model that can generate 128 frames with 192×320 resolution, and fine-tune it to enable multi-shot generation. This extension is achieved by introducing a transition token and implementing local attention masking. Note that our framework also supports cross-attention based conditioning.

Transition token Inspired by the commonly employed '[EOS]' (End of Sentence) token in natural language processing, which signals the model to recognize the end of a sentence, we propose a novel, learnable embedding referred to as the "transition token" which enables the model to learn transitions between consecutive shots within a multi-shot video. Specifically, we initialize a set of learnable parameters at the start of fine-tuning, matching the hidden dimension of the input tokens. We repeat these parameters $n-1$ times—where n represents the specified number of shots—and append them to the end of the input visual and textual token sequence (Fig. 3 (a)). In the model's attention layers, we implement a masking strategy that ensures the transition token interacts only with the tokens corresponding to the frames where transitions are intended to occur (Fig. 3 (b) and (c)). This approach allows the model to focus on transition frames, enabling it to learn to generate cuts between shots, while also allowing users to specify both the number of shots and their respective durations, by simply replicating the transition token and adjusting the attention mask accordingly.

Local attention masking To enable shot-specific control, we introduce a local attention masking technique (Fig. 3 (c)). Without masking, all tokens interact with each other, diluting the impact of shot-specific information. Our proposed local attention masking strategy overcomes this limitation by restricting attention interactions to specific token groups. Specifically, the attention matrix is masked to enforce the interactions, where the transition token attends only to the tokens at the transition frames, while visual and textual tokens are restricted to self-attention. Additionally, to enhance precise and localized control of textual tokens over their corresponding shot-visual tokens, each textual token attends solely to its associated visual tokens. Consequently, the fine-tuned model is better equipped to capture fine-grained dependencies between frames, shot-

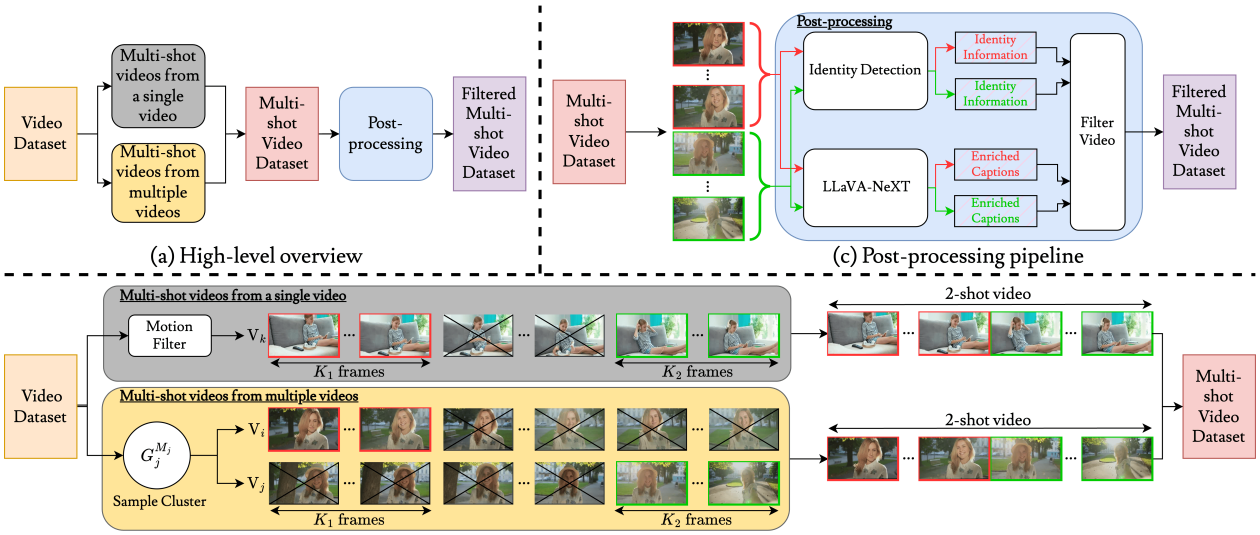


Figure 4. ***Multi-shot video dataset collection pipeline.*** A high-level overview of this pipeline is presented in (a). Our first method (gray box in (b)) samples videos with large motion, randomly splits them into n-shots with varied durations, and concatenates them into multi-shot videos. Our second method (yellow box in (b)) randomly samples n videos from pre-clustered groups containing videos of the same identities and concatenates them to form a multi-shot video. Finally, we post-process (c) the multi-shot videos to ensure identity consistency and obtain shot-specific captions using LLaVA-NeXT [58].

specific text prompts, and shot transitions, resulting in improved control over the T2MSV generation.

3.4. Data

To perform fine-tuning we require a suitable dataset consisting of multi-shot videos where each shot depicts the same identity both in same and different backgrounds, featuring different activities. We develop two methods to curate multi-shot videos given access to a large-scale video dataset (Fig. 4 (c)). The first method creates multi-shot videos from long single-shot videos that exhibit large motion while the second method combines multiple independent videos of the same human subjects to produce multi-shot videos. Post-processing is then applied to obtain shot-specific captions and ensure that the foreground identity remains consistent across shots.

Multi-shot videos from single-shot videos In this approach we generate a multi-shot video from a single-shot video with large motion by randomly trimming short sub-clips and combining them to create a new n-shot video. (Fig. 4 (a), gray box). We first obtain optical flow maps using RAFT [43] and compute the average motion in the x and y directions (t_x, t_y) along with scaling ratio s for each video (see Supplementary for details). Videos with motion metrics below specified thresholds are discarded, retaining only those with significant motion in at least one metric. Additionally, videos with fewer than 250 frames are excluded to ensure sufficient variability. Finally, we choose random sub-clips of random duration and concatenate them in a random order to generate a dataset of multi-shot videos.

Multi-shot videos from independent videos Despite focusing on large-motion videos, the previous approach encounters limitations in terms of diversity in camera perspective, motion and backgrounds, reducing the variety of generated multi-shot videos. To enhance the dataset variability we perform clustering on the single-video dataset, *i.e.* $\mathcal{D} = \{G_1^{M_1}, \cdots, G_n^{M_n}\}$, where each cluster $G_j^{M_j}$ contains M_j videos featuring the same foreground object. These clusters are constructed by choosing videos uploaded by the same user within the same time-frame (*e.g.* 3 days), and videos that match based on the similarity of their captions. we obtain approximately 550K clusters, with an average of 6 videos per cluster. Subsequently, we generate a fixed number of multi-shot videos for each cluster by randomly trimming clips from separate videos within the same cluster (Fig. 4 (a), gray-yellow). This methodology significantly enhances the dataset's richness and variability, introducing a wider array of scenes and perspectives.

Post-processing We apply several post-processing pipelines (Fig. 4 (c)) to (i) obtain shot-specific prompts and (ii) ensure identity consistency across shots. Since each shot in the multi-shot video dataset requires a specific text description, we employ the LLaVA-NeXT [58] model to generate shot-specific captions. Additionally, we apply an additional filtering stage to the multi-shot video dataset to ensure identity consistency across shots. Videos containing more than one character or featuring different identities are filtered out using YOLO [46] as a person detector in the middle frames of each shot within a multi-shot video. To verify identity consistency, we utilize embeddings from DINOv2 and applied a threshold to ensure the same

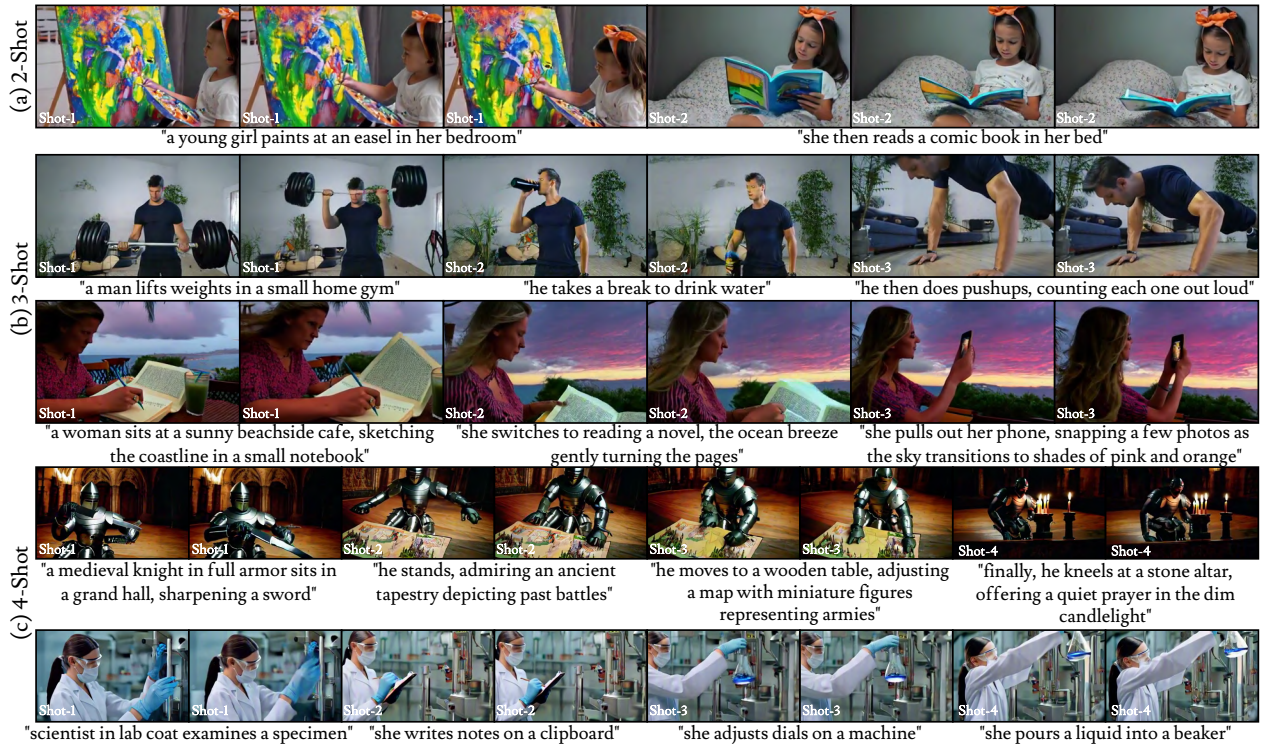


Figure 5. ***Qualitative results.*** Our approach enables multi-shot video generation depicting different actions and background guided by shot-specific prompts. In the 2nd row, the shots maintain a *consistent background* while capturing different perspectives, whereas the 3rd row depicts the same woman in *related backgrounds* that subtly change in response to the prompt. For complete videos, see Supplementary.

identity appears across all shots. This post-processing step filters out 38% of the multi-shot videos, resulting in a dataset composed of multi-shot videos with 2, 3, and 4 shots of varying durations, where a single consistent identity performs similar or different activities across diverse backgrounds and motions.

4. Experiments

Since there is no existing evaluation pipeline for the T2MSV generation task, we design a benchmark to enable standardized evaluations (see Supplementary for implementation details).

Evaluation dataset Our dataset is partitioned according to whether the background remains constant or changes and to the number of shots as specified in the text prompts. For each multi-shot configuration (2, 3, and 4 shots) we provide 8 prompts per scenario (*i.e.* background remains the same or changes) resulting in a total of 48 prompts for multi-shot settings. These prompts are designed to depict a human subject by including terms such as "a person" performing different activities in each shot using ChatGPT [31] (see Supplementary). In total, we generate 128 frames for each sample, with randomly selected shot durations.

Baselines To evaluate our approach in the absence of directly comparable methods we devise three baselines. **Random Shots (RS)** generates each shot guided by detailed text descriptions as a single-shot video multiple times (*i.e.* 48) and concatenates the randomly selected shots. **Similar Shots (SS)** improves on this by selecting shots based on DINOv2 embedding similarity across foreground objects. **Shots by Reference (SR)** first generates keyframes of consistent characters using StoryMaker [62] and then animates those into individual video shots using our model's I2V capability. Note that for all baselines, we use the original pre-trained model, providing a fair comparison by using the same underlying model architecture.

Quantitative evaluation We adapt commonly used metrics from single-shot video generation [19]: (a) *Identity Consistency* (IC), which calculates the average DINOv2 [5] embedding similarity between the segmented persons (segmented using YOLO [46]) at the middle frames of each shot; (b) *Text Alignment* (TA), which assesses the alignment of generated content with text prompts by calculating the similarity between text features and shot features extracted by ViCLIP [50], then averaging across shots; and (c) *Background Consistency* (BC), which measures similarity by segmenting the background and computing DINOv2 embedding similarity across the middle frames of each shot. Our approach outperforms all previous baselines in preserving identity for 3- and 4-shots, while achieving competitive results with SR for 2-shots, as measured by the IC metric (Table 1). As the number of shots increases, there is a decline in performance due to the propagation and accumulation of errors in generating consistent identities. Addi-

Figure 6. ***Qualitative comparison.*** We compare our approach to single-shot generation and baseline methods for (a) 2-shot, (b) 3-shot, and (c) 4-shot videos. Our model (last row) enables local control over shot content while Single-Shot Generation fails to feature distinct activities (1st row). Random Shots (2nd row) and Similar Shots (3rd row) struggle to preserve character identity, especially facial features (cropped below frames in red bounding boxes), which our method effectively maintains. Shots by Reference (4th row) improves identity consistency to some extent but falls short in maintaining both identity (*e.g.* shot 4 in (c)) and background coherence (*e.g.* shot 3 in (b)), where our model demonstrates superior performance. **For the full videos, please see Supplementary.**

tionally, when the background changes (*diff bg*), a modest overall decrease in performance is observed. We report background consistency for samples where the background is intended to remain constant (*same bg*), where our model outperforms all other approaches across every shot by a large margin. In terms of text alignment, RS and SS generally perform better, as expected, since they generate each shot individually, effectively serving as an upper bound for this score. However, our approach achieves competitive results, demonstrating that it preserves identity and background consistency more effectively without significantly compromising text alignment. For additional comparisons to SEINE [8], MEVG [30], FreeNoise [36] and Gen-L-Video [48], see Supplementary.

User study To complement the quantitative metrics, we conduct a user study on the Prolific [35] platform with 75 participants. Each participant views two videos simultaneously, selected from a pool of 10 randomly chosen videos from the generated results, with one video always generated by `ShotAdapter`. Participants are then asked to choose their preferred video based on identity consistency (IC), background consistency (BC), and text alignment (TA). Our approach achieves superior results in a 1-to-1 comparison with baselines (Table 1) in identity and background consistency. In terms of text alignment, it achieves a slight improvement over SS and RS, while outperforming SR by a substantial margin, confirming the trends observed in the quantitative metrics.

Ablation study We conduct two ablation studies: (i) removing the transition token while retaining the local attention mask (Table 1 `ShotAdapter` w/o TT), and (ii) fine-tuning the model exclusively on 2-shot videos (Table 1 `ShotAdapter` w/ 2-shots). Including transition tokens yields a slight improvement in IC, BC, and TA, as it assists the model in generating cuts, thereby enhancing localized control over shot transitions. Although fine-tuning on only 2-shot videos reduces the dataset size, results on 3- and 4-shot videos reveal the model's generalizability, maintaining better identity and background consistency overall than the baselines, despite a slight performance decrease compared to the final model.

Qualitative evaluation Fig. 1 and Fig. 5 show our model's T2MSV generation capabilities across 2, 3, and 4-shot videos, addressing scenarios that require either background

Table 1. ***Quantitative comparison.*** We evaluate our approach against baseline methods across **Identity Consistency**, **Background (BG) Consistency**, **Text Alignment** and a **User Study** on 2, 3, and 4-shot videos under conditions with background changes (*diff bg*), without background changes (*same bg*), and with both types (*all*). `ShotAdapter` w/o TT indicates the model fine-tuned without Transition Tokens (TT) and `ShotAdapter` w/ 2-shots fine-tuned only on a dataset of 2-shot videos. The user study assesses Identity Consistency (Q1), Background Consistency (Q2), and Text Alignment (Q3). ↑ and ↓ indicate the direction toward better performance for each metric.

	Identity Consistency ↑						**BG Consistency ↑**			**Text Alignment ↑**			**User Study**		
Shot Number	2		3		4		2	3	4	2	3	4	Q1 (IC)	Q2 (BC)	Q3 (TA)
Background Change	*diff bg*	*same bg*	*diff bg*	*same bg*	*diff bg*	*same bg*	*same bg*			*all*			Ours vs Baseline (Selection ratio in 1-to-1 comparison)		
Random Shots (RS)	71.03	80.47	54.76	63.72	48.08	55.87	84.46	65.77	59.18	26.84	**26.47**	**25.44**	**77.19**% / 22.81%	**73.27**% / 26.73%	**56.73**% / 43.27%
Similar Shots (SS)	73.94	82.55	55.15	66.17	49.25	58.67	88.85	67.02	60.20	26.40	26.13	25.16	**72.92**% / 27.08%	**69.20**% / 30.80%	**53.13**% / 46.87%
Shots by Reference (SR)	**81.74**	84.98	67.92	72.97	57.83	67.74	82.11	64.85	56.81	25.59	23.97	21.98	**73.43**% / 26.57%	**82.28**% / 17.72%	**73.03**% / 26.97%
`ShotAdapter` w/o TT	77.17	84.78	68.95	70.98	58.83	70.24	87.94	72.93	70.48	26.64	23.15	22.84	N/A	N/A	N/A
`ShotAdapter` w/ 2-shots	78.05	85.46	70.12	71.53	56.99	68.37	89.08	75.53	73.19	25.97	23.59	22.97	N/A	N/A	N/A
`ShotAdapter`	78.67	**86.33**	**70.30**	**76.44**	**61.86**	**74.89**	**89.48**	**77.66**	**76.55**	**27.12**	23.65	22.17	N/A	N/A	N/A

Table 2. ***Transition token generalizability.*** We compute the absolute difference in frames between the generated and ground truth shot duration as the **Mean Shot Duration Error (MSDE)**, and the error per-shot is reported with a range of 2 to 8 shots per video.

Shots	**2**	**3**	**4**	**5**	**6**	**7**	**8**
MSDE	2.00	0.83	1.00	1.70	1.33	0.92	1.21

transitions (*e.g.* a transition from a living room to a walk-in closet) or distinct activities within the same setting (*e.g.* a character lifting weights, drinking water, and doing push-ups). Our model effectively generates multi-shot videos of the same characters, with "cuts" even in diverse settings (full videos are available in the Supplementary). In Fig. 6, we provide a qualitative comparison of our approach with single-shot video generation using extended prompts (row 1) and baseline methods RS, SS and SR for 2, 3, and 4-shot videos. Single-shot generation with localized shot control can result in scenes where actions are intermingled (Fig. 6 (a), where the model fails to transition from walking to eating ramen). RS produces the least consistent outcomes, generating characters with random identities (Fig. 6, zoomed-in faces highlighted with red bounding boxes) and incoherent backgrounds due to the random concatenation of shots. SS shows a minor improvement in identity consistency by selecting shots based on foreground similarity, yet still generates visually distinct identities, despite similar clothing (Fig. 6 (b)), and struggles to maintain coherent backgrounds. SR, achieves better identity consistency than previous baselines as it generates consistent characters using an off-the-shelf method [62] but suffers from quality degradation as the number of shots increases and lacks temporal coherence between keyframes, resulting in notable inconsistencies, such as complete environment changes (Fig. 6 (c)). In contrast, our approach effectively addresses these limitations, generating multi-shot videos with consistent character identity across different background requirements as directed by text prompts.

Transition token generalizability To further assess the generalizability of the transition token, we test our model on videos with 2 to 8 shots, using the **Mean Shot Duration Error** (MSDE) metric, which is calculated by averaging the absolute difference between ground-truth and generated shot durations in terms of frames, where SceneCut [3] is used to detect the cuts. Quantitative results (Table 2) show that despite the temporal compression applied during encoding and patchification, the model achieves an average offset of only 1 to 2 frames per shot, even in 8-shot examples. These findings confirm that the transition token serves effectively as an "End of Shot" marker and can be extended to accommodate multiple shots.

Limitations While our experiments demonstrate the effectiveness of our approach in T2MSV generation this study is limited to human foreground objects, as experiments with non-human subjects (*e.g.* animals) were not conducted. This limitation is primarily due to dataset filtering choices. Additionally, the maximum duration the model can generate is restricted by the underlying model used for fine-tuning, which is limited to 128 frames in this study. For future work, we aim to extend the duration by employing an autoregressive approach to generate additional shots conditioned on previously generated ones. It is worth noting that our method experiences a slight quality reduction in the user study compared to baselines, though it remains highly competitive. We hypothesize that this minor drop is primarily due to fine-tuning with a 90% smaller batch size compared to the baseline as well as our model better adhering to multiple text captions while the baselines often ignore larger parts of the text (see Supplementary for quality analysis).

5. Conclusion

In this paper, we present `ShotAdapter`, a lightweight framework that transforms single-shot T2V models into multi-shot T2MSV generators with minimal fine-tuning. Our approach incorporates a *transition token* and *localized attention masking*, applied to a multi-shot video dataset collected through a novel data collection pipeline. Extensive evaluations demonstrate that our method outperforms baseline models in identity and background consistency without compromising text alignment scores, as further validated by a user study. Additionally, our findings highlight the framework's generalizability to videos with an increasing number of shots, affirming the effectiveness of the "transition token" concept.

References

[1] Kling AI. Kling ai. https://klingai.com/, 2024. Accessed: 2024-11-04. 3

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023. 1

[3] Breakthrough. Pyscenedetect: A cross-platform tool for video scene detection. https://github.com/Breakthrough/PySceneDetect, 2024. Accessed: 2024-11-04. 8

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators, 2024. 2, 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[6] Vasileios T Chasanis, Aristidis C Likas, and Nikolaos P Galatsanos. Scene detection in videos using shot clustering and sequence alignment. *IEEE transactions on multimedia*, 11(1):89–100, 2008. 3

[7] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Diffusion transformers for image and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6441–6451, 2024. 3

[8] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2023. 7, 2

[9] Aram Davtyan, Sepehr Sameni, and Paolo Favaro. Efficient video prediction via sparsely conditioned flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23263–23274, 2023. 3

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 3

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 4

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*. 1, 3

[13] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 3

[14] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[15] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023. 3

[16] Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion. *arXiv preprint arXiv:2407.12899*, 2024. 3

[17] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 1

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

[19] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 6

[20] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 2, 3

[21] Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6507–6516, 2024. 2

[22] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023. 2, 3

[23] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. 3

[24] Luma Labs. Dream machine. https://lumalabs.ai/dream-machine, 2024. Accessed: 2024-11-04. 3

[25] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing re-

alistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. 3

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

[27] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 3

[28] 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. 3

[29] MiniMax. Minimax. https://hailuoai.com/, 2024. Accessed: 2024-11-04. 3

[30] 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 *European Conference on Computer Vision*, pages 401–418. Springer, 2024. 7, 2

[31] OpenAI. Chatgpt: A large language model. https://openai.com/chatgpt, 2024. Accessed: 2024-11-04. 6

[32] 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. 4

[33] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*. 1

[34] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 2, 3, 4

[35] Prolific. Prolific: Online participant recruitment for surveys and research. https://prolific.com/, 2024. Accessed: 2024-11-01. 7

[36] 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 *The Twelfth International Conference on Learning Representations*, 2024. 7, 2

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv e-prints*, pages arXiv–2204, 2022. 1

[38] 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. 1, 3

[39] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1, 3

[40] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6527–6536, 2024. 1

[41] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[42] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[43] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 5, 1

[44] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024. 3

[45] Kristin Thompson, David Bordwell, and Jeff Smith. *Film history: An introduction*. McGraw-Hill Boston, 2003. 3

[46] Ultralytics. YOLOv5: A state-of-the-art real-time object detection system. https://docs.ultralytics.com, 2021. Accessed: 2024-11-04. 5, 6

[47] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation, 2019. 1

[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, 2023. 7, 2

[49] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 3

[50] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2024. 6

[51] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 1, 3

[52] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2, 3

[53] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, pages 1–20, 2024. 3

[54] Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. Facestudio: Put your face everywhere in seconds. *arXiv preprint arXiv:2312.02663*, 2023. 3

[55] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1

[56] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023. 3

[57] HongJiang Zhang, Atreyi Kankanhalli, and Stephen W Smoliar. Automatic partitioning of full-motion video. *Multimedia systems*, 1:10–28, 1993. 3

[58] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/, 2024. 5

[59] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 1

[60] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. https://github.com/hpcaitech/Open-Sora, 2024. 3, 4

[61] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024. 3

[62] Zhengguang Zhou, Jing Li, Huaxia Li, Nemo Chen, and Xu Tang. Storymaker: Towards holistic consistent characters in text-to-image generation. *arXiv preprint arXiv:2409.12576*, 2024. 3, 6, 8, 1