STAGE: Storyboard-Anchored Generation for Cinematic Multi-shot Narrative

Peixuan Zhang[#1] Zijian Jia[#1] Kaiqi Liu[2,3] Shuchen Weng[*4,2] Si Li[*1] Boxin Shi[2,3]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications

[2]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University

[3]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

[4]Beijing Academy of Artificial Intelligence

{pxzhang,jiazijian,lisi}@bupt.edu.cn, liukq04@gmail.com, {shuchenweng,shiboxin}@pku.edu.cn

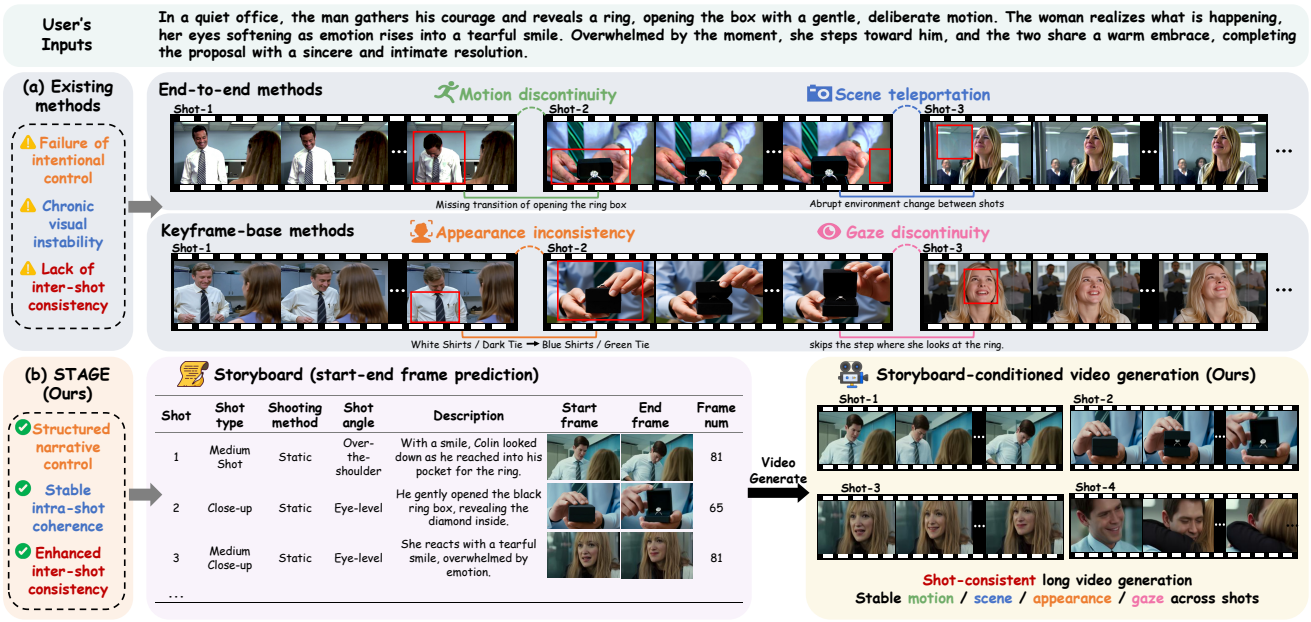

Figure 1. Illustration of our proposed STAGE workflow for multi-shot video generation. Given a user-provided story description, existing end-to-end [53] and keyframe-based [18] methods often suffer from incoherent inter-shot transitions (*e.g.*, motion discontinuities and appearance inconsistencies) that disrupt the narrative flow. We address this by explicitly predicting a structural storyboard composed of start-end frame pairs, which serve as visual anchors to ensure superior long-range shot consistency.

Abstract

*While recent advancements in generative models have achieved remarkable visual fidelity in video synthesis, creating coherent multi-shot narratives remains a significant challenge. To address this, keyframe-based approaches have emerged as a promising alternative to computationally intensive end-to-end methods, offering the advantages of fine-grained control and greater efficiency. However, these methods often fail to maintain cross-shot consistency and capture cinematic language. In this paper, we introduce **STAGE**, a **ST**oryboard-**A**nchored **GE**neration workflow to reformulate the keyframe-based multi-shot video generation task. Instead of using sparse keyframes, we propose STEP[2] to predict a structural storyboard composed of start-end frame pairs for each shot. We introduce the multi-shot memory pack to ensure long-range entity consistency, the dual-encoding strategy for intra-shot coherence, and the two-stage training scheme to learn cinematic inter-shot transition. We also contribute the large-scale ConStoryBoard dataset, including high-quality movie clips with fine-grained annotations for story progression, cinematic attributes, and human preferences. Extensive experiments demonstrate that STAGE achieves superior performance in structured narrative control and cross-shot coherence. Our code will be available at this url.*

# Equal contributions. * Corresponding author.

1. Introduction

Recent advancements in generative models have dramatically improved the visual quality of synthesized videos, fueling their adoption on short-clip content platforms [1, 2, 4]. To further unlock their potential for more sophisticated applications (*e.g.*, filmmaking), recent models are evolving to enhance storytelling ability [31, 57, 63]. However, this narrative ability typically requires composing a long video from multiple distinct shots depicting different stages or perspectives of a story, which still poses significant challenges for existing models.

While directly generating an entire multi-shot video in an end-to-end manner [14, 20, 45, 53] is an intuitive approach, it is computationally expensive and runs in an "all-or-nothing" paradigm, leading to an inefficient trial-and-error process with limited user control. To provide fine-grained user control, keyframe-based approaches [13, 52, 61, 62] first generate several keyframes to establish the video's narrative structure, and subsequently employ external I2V models [3, 8, 55] to synthesize each shot accordingly. However, these methods suffer from disrupted cross-shot coherence and struggle to capture cinematic language (Fig. 1 (a), the close-up reveals motion discontinuity when opening the ring box).

In this paper, we reformulate the keyframe-based approach, proposing a structural storyboard composed of Start-End frame pairs defining each shot (*i.e.*, $(F_1^{\mathrm{S}}, F_1^{\mathrm{E}}), (F_2^{\mathrm{S}}, F_2^{\mathrm{E}}), ...$), instead of a sparse keyframe sequence. This formulation offers three advantages: *(i) Structured narrative control:* These start-end frame pairs form a robust narrative scaffold, ensuring long-range consistency of entities (*e.g.*, character appearance) and settings (*e.g.*, scene background) throughout the entire video. *(ii) Intra-shot coherence:* The intra-shot pair $(F_i^{\mathrm{S}}, F_i^{\mathrm{E}})$ explicitly anchors visual content (*e.g.*, character) and the intended progression within the shot (*e.g.*, camera movements). *(iii) Inter-shot transitions:* The inter-shot pair $(F_i^{\mathrm{E}}, F_{i+1}^{\mathrm{S}})$ explicitly models the shot transition, effectively conveying complex cinematic language (*e.g.*, shot/reverse shot). However, realizing these intra-, inter-, and global advantages is non-trivial, demanding both a tailored model and a structurally annotated dataset.

To address this challenge, we introduce **STAGE**, a workflow for **ST**oryboard-**A**nchored **GE**neration for cinematic multi-shot narrative. The technical core of this workflow is our **ST**art-**E**nd frame-**P**air **P**rediction model (**STEP**^{2}), which iteratively visualizes the text-based storyboard into $(F_i^{\mathrm{S}}, F_i^{\mathrm{E}})$ pairs. To further enable STEP^2 to visualize the storyboard of each shot effectively, we propose: *(i)* a multi-shot memory pack to compress the history and ensure long-range entity consistency; *(ii)* a dual-encoding strategy to enforce intra-shot coherence and logical correlation; and *(iii)* a two-stage training scheme to understand the complex cinematic language of inter-shot transitions. In the complete STAGE workflow, a director agent first augments the user-provided story theme into a text-based storyboard. Our fully-trained STEP^2 model then iteratively generates the $(F_i^{\mathrm{S}}, F_i^{\mathrm{E}})$ pairs. These pairs are finally fed into an off-the-shelf video generation model (*e.g.*, WanX [44], Veo3.1 [5]) to generate video clips, which are then concatenated into the final multi-shot video.

To facilitate the training of STEP^2, we collect open-sourced movie clips [7] and construct the ConStoryBoard dataset[1]. It consists of 100K video clips with start-end frame pairs, along with structural storyboard annotations (*e.g.*, story progress of each shot) and diverse cinematic language attributes (*e.g.*, shot scale, shot length, camera angle, and camera movement). Building on this, we further curate the ConStoryBoard-HP by manually selecting high-quality frame pairs and constructing preference tuples for human-preference alignment post-training.

We conclude our contributions as follows:

- We reformulate the keyframe-based approach by introducing STEP^2, which predicts structural start-end frame pairs to enable controllable multi-shot video generation.
- We design a dual-encoding strategy to ensure logical correlation within each shot, and a multi-shot memory pack to provide long-range consistency of entities.
- We construct the ConStoryBoard dataset with fine-grained annotations, including a manually selected preference subset to optimize inter-shot cinematic language.

2. Related work

2.1. Single-Shot video generation

The field of single-shot video generation has seen remarkable progress. Building on the success of image diffusion models [9, 35, 38, 49, 59, 60], early methods [15, 16, 40, 50] adapt pre-trained models by incorporating temporal layers. However, these approaches often result in limited temporal coherence. To address these limitations, research has shifted towards Diffusion Transformer architectures [32]. While state-of-the-art video foundation models [22, 44, 58] are able to generate high-fidelity videos with a single shot, they typically lack a native mechanism for multi-shot narratives. Consequently, naively applying them shot-by-shot results in jarring visual discontinuities that disrupt the narrative flow, marking the key challenge our work addresses.

2.2. Multi-Shot video generation

Generating multi-shot videos requires maintaining both content consistency and narrative coherence across shots. Existing methods extend single-shot models to generate full sequences, either via end-to-end visual conditioning [28, 61] or by modifying attention mechanisms [20, 29, 53].

[1] We will publish the dataset once the paper is accepted.

These approaches are computationally expensive and offer limited user control. To address these limitations, an alternative approach [34, 52, 57, 63] first generates sparse keyframes and then synthesizes the full video clips based on them. Despite their advances in text fidelity and story coherence, they largely neglect the cinematic language for inter-shot transitions, leading to abrupt cuts and logical disconnects that break the narrative flow. In contrast, our STAGE explicitly reformulates the keyframe-based approach as a start-end frame pair prediction problem, thereby directly modeling these inter-shot transitions.

2.3. Reinforcement learning for visual generation

Inspired by the success of Reinforcement Learning from Human Feedback (RLHF) in language models, aligning visual models with human preferences is a promising research area. Initial efforts involve directly fine-tuning models with scalar reward signals [10, 56] or employing reward weighted regression [24, 33]. Subsequently, policy gradient algorithms (*e.g.*, PPO [39]) are integrated into diffusion models, demonstrating notable efficacy in visual quality improvement [12, 30]. To avoid significant computational overhead and training instability, Direct Preference Optimization (DPO) [37] emerges and demonstrates its effectiveness in aligning visual models with human preferences for static image properties, such as text-image fidelity [43, 46] and aesthetic quality [25]. Building on this, we extend this technique to optimize for the complex temporal relationships (*e.g.*, cinematic language) that define coherent storytelling in multi-shot videos.

3. Dataset

Current keyframe-based video datasets [34, 53, 54] offer a single keyframe for each shot, primarily focusing on text-to-shot alignment while neglecting the start and end frames for modeling inter-shot transitions. This motivates us to construct the ConStoryBoard dataset, tailored to train our start-end frame pair prediction model.

The construction of ConStoryBoard begins with multi-shot videos collected from the Condensed Movies dataset [7], which we filter for high-resolution (over 1080p) and high aesthetic scores [23] (over 5.5). These filtered videos are then segmented into individual video clips for each shot using TransNetV2 [41]. For each resulting clip (i-th shot in the original video), we employ InternVL-3.5 [47] to generate a detailed text description D_i of the story progression, and structured cinematic attributes C_i (*e.g.*, shot scale and camera movement). We then extract the ground-truth start-end frame pair $(F_i^{\mathrm{S}}, F_i^{\mathrm{E}})$ from these annotated clips, and post-process them by cropping black bars and removing watermarks. Finally, the ConStoryBoard dataset comprises 100K training pairs and 1K testing pairs, where

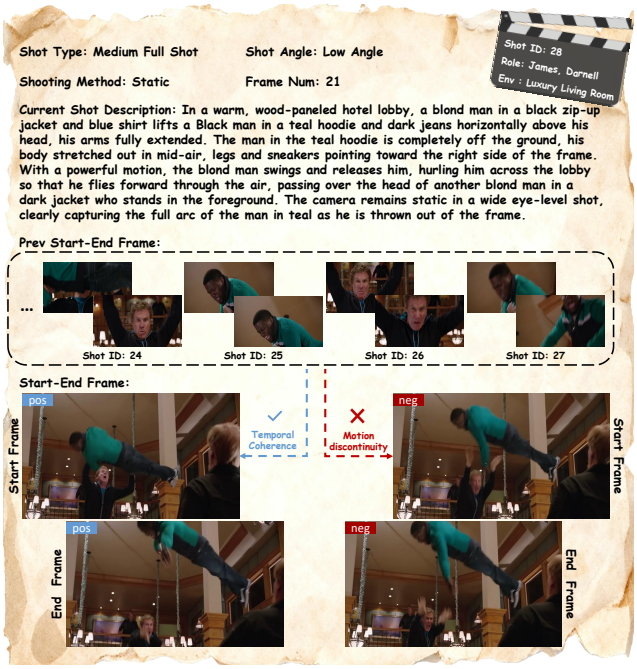

Figure 2. An example of the ConStoryBoard dataset.

each sample consists of a ground-truth pair $(F_i^{\mathrm{S}}, F_i^{\mathrm{E}})$ and its corresponding annotations (D_i, C_i).

To facilitate model alignment with human preferences, we further curate ConStoryBoard-HP (Human-Preferred) by manually selecting the most high-quality and cinematically significant pairs from our dataset. To construct the preference tuples, we define the ground-truth pair $(F_i^{\mathrm{S}}, F_i^{\mathrm{E}})$ as the positive sample y_w. Observing that internal frames from the same video clip typically represent a mismatched cinematic language (*e.g.*, an incomplete camera movement), we randomly sample two internal frames from the same clip as the corresponding negative sample y_l. These preference tuples (y_w, y_l) are used for the alignment post-training.

4. Method

This section begins with an introduction to the Start-End frame-pair prediction model (STEP^2) to iteratively generate start-end frame pairs (Sec. 4.1). Then, we design the two-stage training scheme to train the STEP^2 model for understanding cinematic language (Sec. 4.2). Finally, we present the STAGE workflow to demonstrate the approach of STEP^2 applied for multi-shot video generation (Sec. 4.3).

4.1. Start-end frame pair prediction model

Multi-shot memory pack. To enable iterative generation of start-end frame pairs with long-term temporal consistency of entities, the STEP^2 requires a memory mechanism for preceding shots. Since referring to all previous shots will bring a remarkable computational burden, we propose

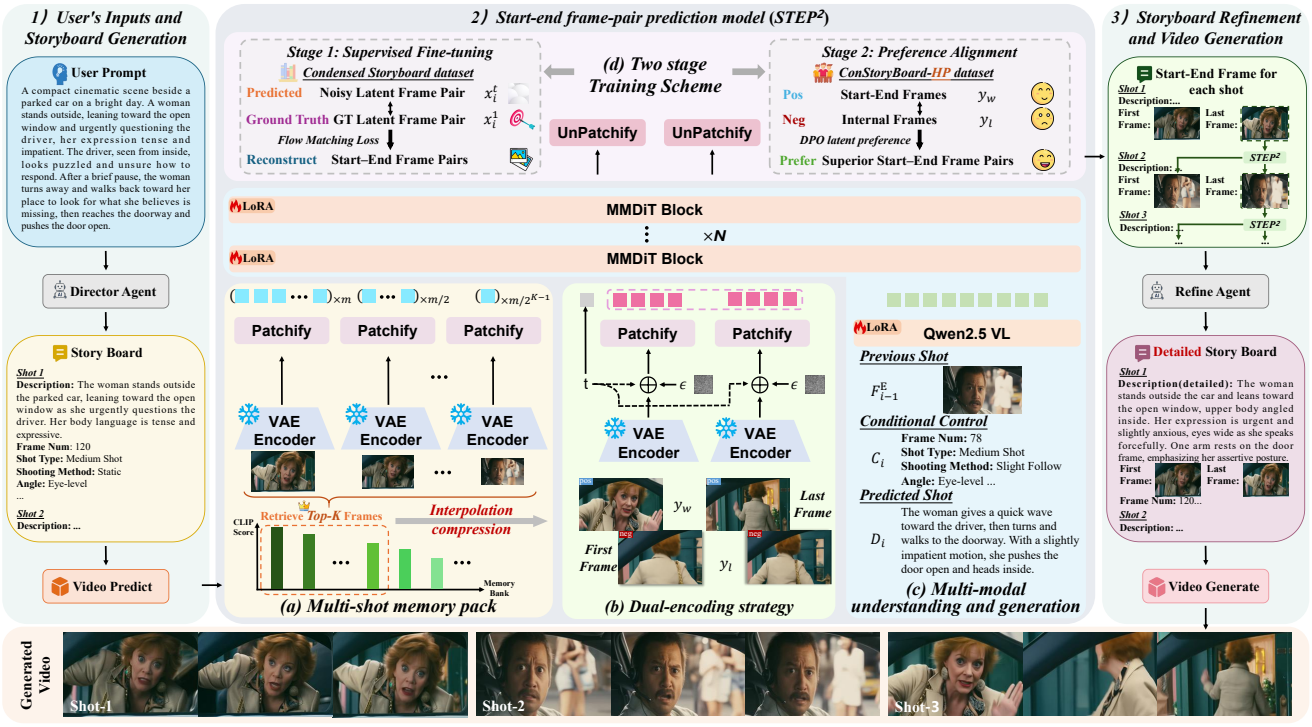

Figure 3. Overview of our proposed STAGE workflow. The core component of STAGE is the **Start-End frame-pair prediction model (STEP^2)**, which iteratively generates start-end frame pairs for each shot. To ensure long-term consistency, STEP^2 is equipped with the **(a) multi-shot memory pack** that compresses preceding visual context into a compact token. The **(b) dual-encoding strategy** is adopted to maintain intra-shot coherence by jointly encoding the start and end frames of the current shot. By integrating **(c) multi-modal understanding and generation** models into a unified architecture, STEP^2 performs robust reasoning from diverse contexts to ensure inter-shot coherence (Sec. 4.1). STEP^2 is optimized via a **(d) two-stage training scheme**: an initial supervised fine-tuning stage establishes a strong generative foundation, followed by a preference alignment stage to align outputs with human preferences (Sec. 4.2). During inference, STEP^2 is integrated into the STAGE workflow. A Director Agent first generates a structured storyboard from a user's theme. Then, STEP^2 produces the corresponding start-end frame pairs for each shot. Finally, a Refiner Agent uses these frames to create enhanced prompts that guide an off-the-shelf video model to synthesize the final multi-shot video (Sec. 4.3).

the multi-shot memory pack to compress these shots into a compact context.

Specifically, as illustrated in Fig. 3 (a), given preceding shot start-end frame pairs $\{(F_j^{\mathrm{S}}, F_j^{\mathrm{E}})\}_{j=1}^{i-1}$ for i-th shot, we disorderly collect all frames in these pairs into a memory bank $\{F_j^{\mathrm{M}}\}_{j=1}^{2i-2}$, and encode all these frames into the latent space $\{m_j\}_{j=1}^{2i-2} = \{\mathcal{E}_{\mathrm{vae}}(F_j^{\mathrm{M}})\}_{j=1}^{2i-2}$ using a pre-trained VAE encoder $\mathcal{E}_{\mathrm{vae}}$. To guide the iterative generation with semantically relevant memory cues, we rank these preceding latent codes based on their CLIP similarity [36] as $\{m'_j\}_{j=1}^{2i-2} = \mathrm{Sort}\big(\{m_j\}_{j=1}^{2i-2}\big)$. After that, this ranked memory bank is further compressed into a packed memory token via a progressive spatial tiling mechanism:

$$M_i = \mathrm{SpatialTile}_{j\in\{1,\dots,2i-2\}}\big(\mathcal{P}(m'_j, A_j)\big), \qquad (1)$$

where $\mathcal{P}(m'_j, A_j)$ is the downsample function that compresses the latent code m'_j based on the compression rate $A_j = \frac{1}{2^j}$ to ensure the total area for M_i is mathematically converged (*i.e.*, $\sum_{j=1}^{\infty} A_j$=1). These packed memory tokens M_i are then fed into the generation model $\mathcal{E}_{gen}$ to present a potentially infinite generation history.

Dual-encoding strategy. To ensure the predicted start-end frame pair maintains intra-shot story coherence (*e.g.*, scenes remain visually consistent) and temporal dynamics (*e.g.*, a zoom-in cinematic language), we propose a dual-encoding strategy to enable the implicit visual context sharing between the start and end frames.

Specifically, as shown in Fig. 3 (b), we encode the ground truth start-end frame pair $(F_i^{\mathrm{S}}, F_i^{\mathrm{E}})$ separately using a pre-trained VAE encoder $\mathcal{E}_{vae}$, and then concatenate them along the sequence dimension to construct a single joint shot tensor $x_i = \big[\mathcal{E}_{\mathrm{vae}}(F_i^{\mathrm{S}}); \mathcal{E}_{\mathrm{vae}}(F_i^{\mathrm{E}})\big]$. As the common practice of flow matching [26], this shot tensor is then linearly interpolated with a Gaussian noise as:

$$x_i^t = t \cdot x_i^1 + (1-t) \cdot x_i^0, \qquad (2)$$

where $t \in [0, 1]$ is the continuous time step, $x_i^1 = x_i$ is the clean shot tensor, and $x_i^0 \sim \mathcal{N}(0, \mathbf{I})$ is Gaussian noise.

Multi-modal understanding and generation. To reason about shot semantics from diverse context, our STEP^2 first adopts a multi-modal large language model to understand the story progress and character performance of each shot. Specifically, as illustrated in Fig. 3 (c), the understanding model $\mathcal{E}_{mu}$ is built upon Qwen2.5-VL [6], which leverages the end frame of the previous shot F^{E}_{i-1}, the i-th shot corresponding text description D_i, and cinematic attributes C_i to generate unified context tokens $U_i = \mathcal{E}_{mu}(F^{\mathrm{E}}_{i-1}, D_i, C_i)$ for the i-th shot.

With the aforementioned memory tokens M_i and interpolated joint shot tensor x^t_i, these unified context tokens U_i are fed into the generation model $\mathcal{E}_{gen}$ to generate the corresponding start-end frame pair $(F^{\mathrm{S}}_i, F^{\mathrm{E}}_i)$. Specifically, the generation model $\mathcal{E}_{gen}$ is implemented as a diffusion Transformer architecture, comprising multiple MMDiT blocks [11] that includes the self-attention layers for global context interaction. This enables our STEP^2 to generate clean start-end frame pairs referring all integrated and compression semantics by solving the ODE:

$$\mathrm{d}x^t_i/\mathrm{d}t = \mathcal{E}_{\mathrm{gen}}(U_i, t, x^t_i, M_i), \tag{3}$$

from $t = 0$ to $t = 1$ using a numerical solver.

4.2. Two-stage training scheme

Supervised fine-tuning. We first conduct a supervised fine-tuning (SFT) on our ConStoryBoard dataset to pre-train the STEP^2 model, establishing a strong foundation for inter-shot transitions. Specifically, we employ the Low-Rank Adaptation (LoRA) technique [17] on both the understanding model $\mathcal{E}_{mu}$ (for high-level semantic understanding) and the generation model $\mathcal{E}_{gen}$ (for low-level frame generation). Following the flow matching formulation [27], our training target is the constant velocity vector field $v_t = x^1_i - x^0_i$, pointing from the noise $x^0_i \sim \mathcal{N}(0, \mathbf{I})$ to the ground-truth latent pair x^1_i:

$$\mathcal{L}_{\mathrm{SFT}} = \mathbb{E}_{x^1_i, x^0_i, \mathcal{C}_i, t} \| v_\theta(x^t_i, t, \mathcal{C}_i) - v_t \|^2, \tag{4}$$

where $\mathcal{C}_i = [D_i, C_i, \{(F^{\mathrm{S}}_j, F^{\mathrm{E}}_j)\}^{i-1}_{j=1}]$ represents the text description, the cinematic attribute, and all the preceding start-end frame pairs and v_θ is our entire STEP^2 model.

Preference alignment. To further align our STEP^2 model with human preferences, we perform the post-training using Direct Preference Optimization (DPO) [37]. Using the SFT-trained model as the reference model v_{ref} and extracting preference tuple (y_w, y_l) from the ConStoryBoard-HP subset, this process aims to maximize the likelihood of policy model v_θ prefers y_w over y_l:

$$\mathcal{L}_{\mathrm{DPO}} = -\mathbb{E}_{(y_w, y_l), \mathcal{C}_i, t} \left[\log \sigma \left(\beta (\mathrm{D}_\theta - \mathrm{D}_{\mathrm{ref}}) \right) \right], \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function and β is a scaling parameter. The preference differences for the policy and reference models are defined as:

$$\mathrm{D}_k = \| v_k(\hat{x}^t_i, t, \mathcal{C}_i) - \hat{v}^t \|^2 - \| v_k(\check{x}^t_i, t, \mathcal{C}_i) - \check{v}^t \|^2, \tag{6}$$

where $k \in \{\theta, \mathrm{ref}\}$, $\hat{x}^t_i$ and $\check{x}^t_i$ are noisy latent codes from the negative y_l and positive sample y_w, respectively. The $\hat{v}^t$ and $\check{v}^t$ are their corresponding ground-truth velocity vectors, defined as $\hat{v}^t = x^1_{i,l} - x^0_{i,l}$ and $\check{v}^t = x^1_{i,w} - x^0_{i,w}$. Here, $x^1_{i,l}$ and $x^1_{i,w}$ are the VAE-encoded latents for y_l and y_w.

4.3. STAGE workflow

We propose the STAGE workflow to generate coherent multi-shot videos $V = [V_1, \ldots, V_N]$ based on a story description T_{desc}, where our well-trained STEP^2 model is integrated to predict start-end frame pairs for each shot. This workflow is organized into three main stages:

1) User's Inputs and Storyboard Generation: Given a user-provided text-based story theme T_{desc}, we design a Director Agent G_{dire} with chain-of-thought [48] pre-defined prompts to generate a structured storyboard $\mathcal{S}$:

$$\mathcal{S} = \mathrm{G}_{\mathrm{dire}}(T_{\mathrm{desc}}), \tag{7}$$

where $\mathcal{S} = \{D_i, C_i\}^N_{i=1}$ is a sequence of N shot specifications $\mathcal{S}_i = (D_i, C_i)$, each one details the shot of story progress with a text description D_i and a set of cinematic attributes C_i.

2) Iterative Start-End Frame Generation: Leveraging the well-trained Start-End frame-Pair Prediction model (*i.e.*, STEP^2), the abstract storyboard for each shot is iteratively translated into concrete visual anchors (*i.e.*, a start-end frame pair $(F^{\mathrm{S}}_i, F^{\mathrm{E}}_i)$):

$$(F^{\mathrm{S}}_i, F^{\mathrm{E}}_i) = \mathrm{STEP}^2(D_i, C_i, \{(F^{\mathrm{S}}_j, F^{\mathrm{E}}_j)\}^{i-1}_{j=1}). \tag{8}$$

Since the first shot ($i = 1$) has no preceding context, the model is conditioned solely on its specification (D_i, C_i). This process iteratively generates a complete sequence of frame pairs to the last shot ($i = N$).

3) Storyboard Refinement and Video Generation: To bridge the gap between static keyframes and dynamic video, we employ a Shot Refiner Agent G_{refiner}. This agent integrates the original storyboard specification (D_i, C_i) with the generated visual anchors $(F^{\mathrm{S}}_i, F^{\mathrm{E}}_i)$ to produce a comprehensive text description R_i for each shot:

$$R_i = G_{\mathrm{refiner}}(D_i, C_i, F^{\mathrm{S}}_i, F^{\mathrm{E}}_i). \tag{9}$$

Subsequently, this refined description R_i and the start-end frame pair $(F^{\mathrm{S}}_i, F^{\mathrm{E}}_i)$ are fed into an off-the-shelf video generation model G_{video} (*e.g.*, WanX [44], Veo3.1 [5]) to guide the generation process of the video clip V_i of the i-th shot:

$$V_i = G_{\mathrm{video}}(R_i, F^{\mathrm{S}}_i, F^{\mathrm{E}}_i). \tag{10}$$

Finally, the multi-shot narrative video $V = [V_1, \ldots, V_N]$ is produced by concatenating all individual clips.

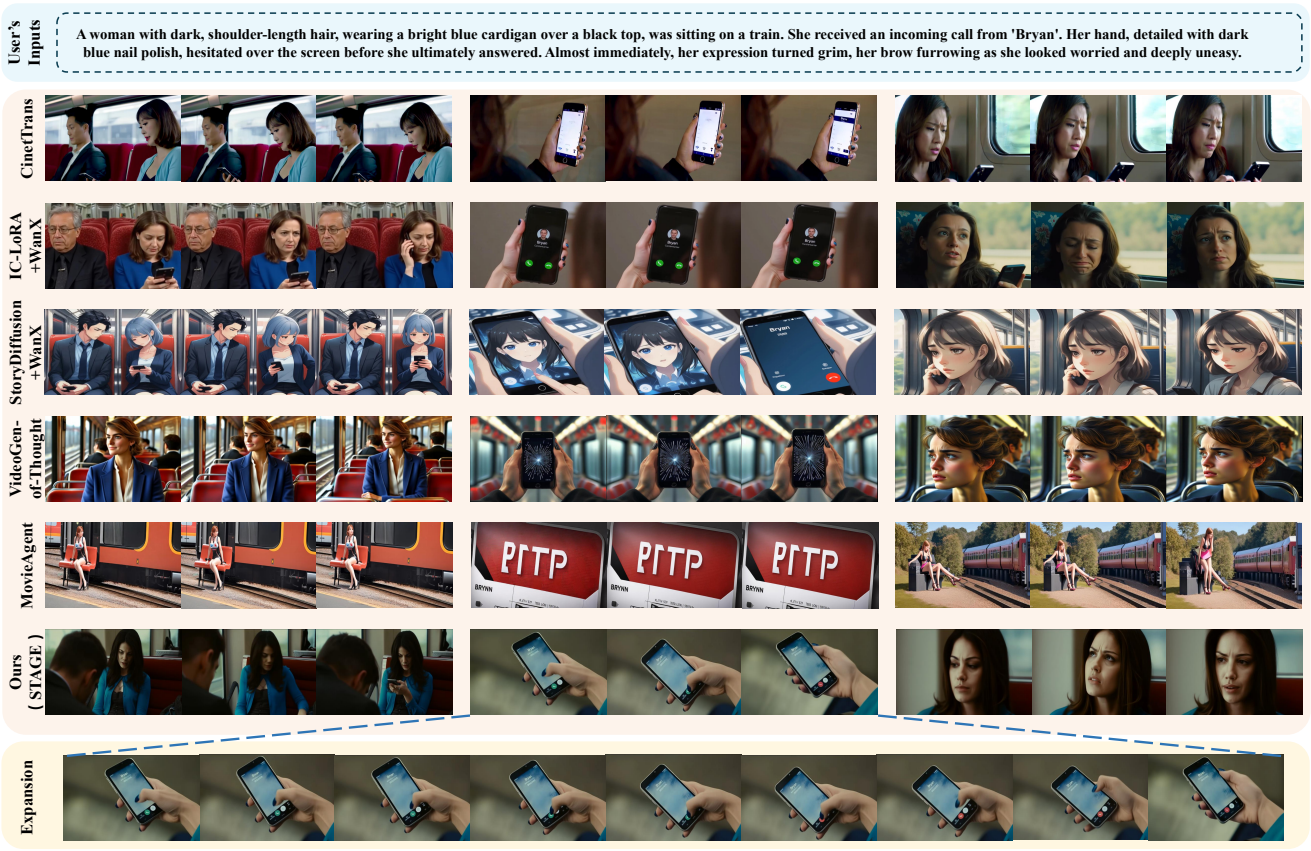

Figure 4. Visual quality comparisons with multi-shot video generation methods.

5. Experiment

5.1. Training details

Our implementation is based on Qwen-Image [51] where VAE is frozen. The rank of LoRA weight is set to 64. Training is performed on 8 A800 GPUs using the Adam optimizer [21] at a learning rate of 1×10^{-4}. Our two-stage training scheme consists of an initial 100K iterations of supervised fine-tuning, followed by an additional 20K iterations of preference alignment.

5.2. Quantitative evaluation metrics

We comprehensively evaluate STAGE across five aspects[2]: *(i)* Overall video quality. We assess the perceptual quality of generated videos using the **Aesthetic Quality (AQ)** and **Image Quality (IQ)** metrics from VBench [19]. *(ii)* Text video consistency. To measure the alignment between the generated video and the input text description, we employ the **Overall Consistency (OC)** metric from VBench [19]. *(iii)* Intra-shot coherence. We evaluate the temporal consistency within individual shots using **Subject Consistency (SC)** and **Background Consistency (BC)** from VBench [19]. *(iv)* Inter-shot coherence. We extend VBench's intra-shot metrics for inter-shot evaluation, defining **Subject Consistency-Extra (SC-E)** and **Background Consistency-Extra (BC-E)** metrics. *(v)* Inter-shot transition. We introduce **Transition Vector Similarity (TVS)** to measure cinematic transition quality. To further evaluate the quality of multi-shot video, we employ the VLM [42] to rate **Overall Video Quality (OVQ)**, **Video Text Consistency (VTC)**, **Inter-Shot Consistency (ISC)**, **Shot Transition Smoothness (STS)** on a scale of 0 to 1 score.

[2]Metric details are provided in the supplementary materials.

5.3. Comparison with State-of-the-art Methods

We compare our approach with state-of-the-art multi-shot video generation methods, including end-to-end approaches (*i.e.*, CineTrans [53]) and key-frame-based approaches (*i.e.*, IC-LoRA [18] + WanX [44], StoryDiffusion [63] + WanX [44], MovieAgent [52], and VideoGen-of-Thought [62]).

Qualitative comparisons. In Fig. 4, we present visual quality comparison results with aforementioned methods. Both CineTrans [53] and IC-LoRA [18] + WanX [44] fail to maintain consistency across multiple shots (*e.g.*, Fig. 4 first and second row, the appearance of the train car win-

Table 1. Quantitative experiment results of comparison and ablation. ↑ (↓) means higher (lower) is better. Throughout the paper, the best performances are highlighted in **bold**.

Method	Quantitative metrics evaluation								LLM evaluation			
	AQ ↑	IQ ↑	OC ↑	SC ↑	BC ↑	SC-E ↑	BC-E ↑	TVS ↑	OVQ ↑	VTC ↑	ISC ↑	STS ↑
Comparison with state-of-the-art methods												
CineTrans [53]	0.5652	0.6120	0.2018	0.9437	0.9504	0.6197	0.7428	0.0455	0.7972	0.3551	0.5585	0.4931
IC-LoRA [18] + WanX [44]	0.6333	0.6951	0.2140	0.9567	0.9615	0.5319	0.7438	0.2090	0.7597	0.3897	0.4901	0.4696
StoryDiffusion [63] + WanX [44]	0.6941	0.7018	0.2087	0.9456	0.9640	0.5780	0.7988	0.1441	0.5343	0.2069	0.4813	0.4575
VideoGen-of-Thought [62]	0.7210	0.6630	0.1689	0.9599	0.9554	0.6278	0.7830	0.0966	0.8106	0.1120	0.5086	0.4507
MovieAgent [52]	0.5742	0.7069	0.0711	0.9664	0.9384	0.4993	0.6473	0.0079	0.4895	0.1931	0.4511	0.4182
Ours (STAGE)	**0.7689**	**0.7305**	**0.2713**	**0.9695**	**0.9685**	**0.6917**	**0.8207**	**0.2732**	**0.8929**	**0.6069**	**0.6985**	**0.6255**
Ablation study												
W/o MMP	0.7344	0.7220	0.2143	0.9631	0.9592	0.6088	0.7311	0.2370	0.8835	0.3642	0.5466	0.5228
W/o DES	0.7217	0.7145	0.2488	0.9542	0.9476	0.6803	0.8124	0.2680	0.7803	0.5063	0.6552	0.5167
W/o TTS	0.7476	0.7240	0.2635	0.9613	0.9633	0.6636	0.8037	0.2195	0.8733	0.5766	0.6614	0.5111

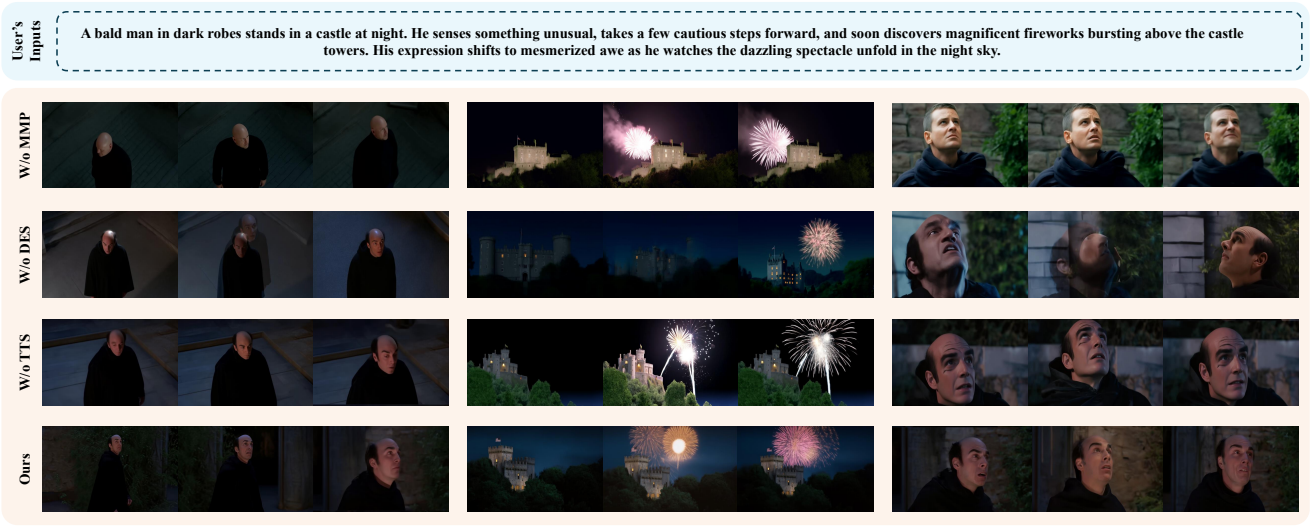

Figure 5. Ablation study results with different variants of our STAGE framework.

dow and the surrounding environment in the first shot are inconsistent with those in the third, leading to a disjointed narrative). StoryDiffusion [63] + WanX [44] produces an overly cartoonish style and lower overall quality (*e.g.*, Fig. 4 third row, the entire sequence exhibits a stylized aesthetic that is unsuitable for cinematic applications). VideoGen-of-Thought [62] suffers from poor temporal coherence in its action sequences (*e.g.*, Fig. 4 fourth row, the woman's hands are on her arm in the first shot but abruptly holds a phone with hands in the second, breaking the continuity of motion). MovieAgent [52] exhibits low text fidelity (*e.g.*, Fig. 4 fifth row, despite the prompt is sitting on a train, the generated video depicts a woman beside the train, failing to accurately reflect the user's intent). In contrast, our STAGE workflow successfully maintains high inter-shot consistency while ensuring shot smooth transitions, resulting in a coherent and visually superior cinematic result.

Quantitative comparisons. We present quantitative comparisons in Tab. 1, where our method significantly outperforms relevant multi-shot generation approaches across all eight quantitative metrics and four LLM-based evaluations. This demonstrates its ability to create multi-shot cinematic videos with high fidelity to user instructions (OC, VTC), strong consistency across intra-shot and inter-shot (SC, BC, SC-E, BC-E, ISC), seamless cinematic transitions (TVS, STS), and superior overall quality (AQ, IQ, OVQ).

Table 2. Percentage (%) of user ratings in the four experiments of human evaluation for the results.

Method	VQE	TAE	SCE	ITE
CineTrans [53]	13.2	18.4	6.4	16.4
IC-LoRA [18] + WanX [44]	24.4	12.8	13.2	7.2
StoryDiffusion [63] + WanX [44]	1.2	8.0	5.6	3.6
VideoGen-of-Thought [62]	2.8	4.4	0.4	2.0
MovieAgent [52]	0.8	3.2	1.6	1.2
Ours (STAGE)	**57.6**	**53.2**	**72.8**	**69.6**

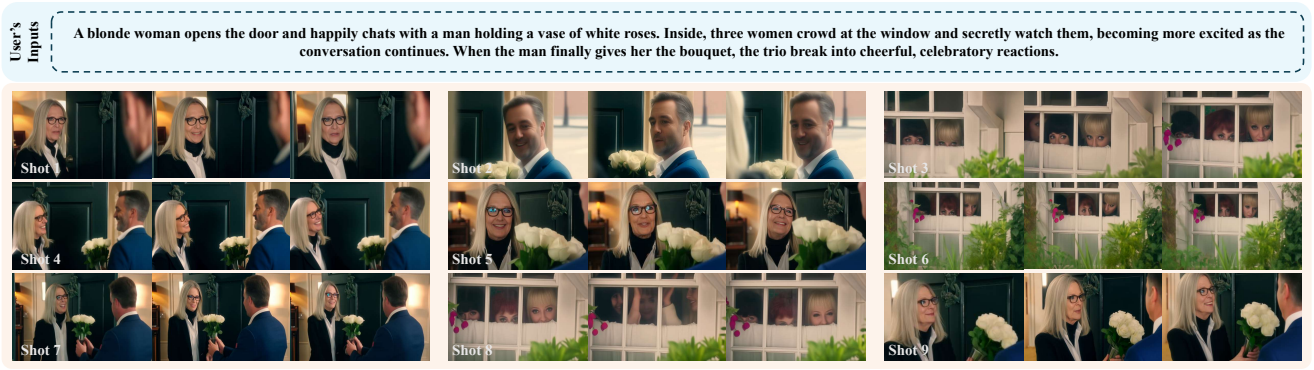

Figure 6. A long-range multi-shot video generation result of our STAGE framework, demonstrating high cross-shot consistency.

User study. In addition to qualitative and quantitative comparisons, we conduct four user studies to evaluate human preference: *(i)* **Visual Quality Evaluation (VQE):** Participants are shown the generated results produced by STAGE and relevant multi-shot video generating methods. They are asked to select the most visually pleasing video. *(ii)* **Text Alignment Evaluation (TAE):** Given a corresponding text description, participants are instructed to select the video from the same set of generating results that best matches the description. *(iii)* **Shot Consistency Evaluation (SCE):** This study measures intra- and inter-shot visual consistency in multi-shot videos. Participants evaluate which sequence maintains stable appearance, motion, and spatial coherence across frames and shots. *(iv)* **Inter-shot Transition Evaluation (ITE):** This evaluation measures perceptual continuity and narrative logic between shots, where participants select the video with smoother visual transitions and more natural scene progression. For each experiment, we randomly select 20 samples from the dataset, and recruit 25 volunteers from Amazon Mechanical Turk (AMT) to provide independent evaluations. As shown in Tab. 2, our model achieves the highest preference scores in all experiments.

5.4. Ablation study

We discard several modules and establish three baselines to study the impact of the corresponding modules. The evaluation scores and visual results of the ablation study are presented in Tab. 1 and Fig. 5, respectively.

W/o Multi-shot Memory Pack (MMP). We discard the multi-shot memory pack, which causes the model to be unable to obtain context from preceding shots. Therefore, the model struggles to maintain visual consistency across the generated sequence (lower SC-E and BC-E scores. As shown in the first row of Fig. 5, the character's appearance changes, and the scene illogically shifts from night to day across the first shot and the third shot).

W/o Dual-Encoding Strategy (DES). We remove the dual-encoding strategy and generate the first and last frames independently. Lacking mutual reference, the start and end frames fail to share context, which leads to degraded intra-shot consistency (lower SC and BC scores. As shown in the second row of Fig. 5, the castles generated in the start and end frames are inconsistent).

W/o Two-stage Training Scheme (TTS). We discard the two-stage training scheme, training the model solely through supervised fine-tuning. Lacking exposure to negative examples, the model struggles to capture valid inter-shot transitions, leading to motion discontinuities (lower TVS score. As shown in the third row of Fig. 5, an abrupt cut where the man is looking down in the first shot, but the camera is already shooting the sky in the second shot).

5.5. Long multi-shot video generation

As shown in Fig. 6, we demonstrate STAGE's capability to generate long multi-shot videos by composing individual shots into a coherent and extended narrative.

6. Conclusion

In this paper, we introduced STAGE, a novel storyboard-anchored workflow for generating coherent multi-shot videos. Our key insight is to reformulate the task as a start-end frame pair prediction problem. This allows our proposed model (STEP^2) to effectively ensure cross-shot consistency and narrative structure using a multi-shot memory pack and a dual-encoding strategy. To capture nuanced cinematic language, we employ a two-stage training scheme with preference alignment, supported by our newly constructed ConStoryBoard dataset. Extensive experiments show that STAGE significantly outperforms state-of-the-art methods in narrative controllability, consistency, and cinematic quality. We believe STAGE is a significant step towards practical and user-directable video generation for long-form storytelling and AI-assisted filmmaking.

Limitations. While multi-shot memory pack ensures long-range shot consistency with start-end frame pairs, the reliance on an off-the-shelf video generator for infilling can introduce temporal inconsistencies within the clips.

7. Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62136001), the Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012), and the Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance.

References

[1] Kling. Accessed December 9, 2024 [Online] https://klingai.kuaishou.com/, 2024. 2

[2] Sora. Accessed February 15, 2024 [Online] https://sora.chatgpt.com/, 2024. 2

[3] Runway gen-4. Accessed April 1, 2025 [Online] https://runwayml.com/, 2025. 2

[4] Sora 2. Accessed September 30, 2025 [Online] https://sora.chatgpt.com/, 2025. 2

[5] Veo3.1. Accessed October 16, 2025 [Online] https://aistudio.google.com/models/veo-3/, 2025. 2, 5

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5

[7] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2, 3

[8] 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. 2

[9] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *ICLR*, 2024. 2

[10] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023. 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. 5

[12] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *NeurIPS*, 2023. 3

[13] Jingwen He, Hongbo Liu, Jiajun Li, Ziqi Huang, Yu Qiao, Wanli Ouyang, and Ziwei Liu. Cut2next: Generating next shot via in-context tuning. *arXiv preprint arXiv:2508.08244*, 2025. 2

[14] Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *CVPR*, 2025. 2

[15] 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. 2

[16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022. 2

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 5

[18] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024. 1, 6, 7

[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 *CVPR*, 2024. 6

[20] Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M Rehg, and Tobias Hinz. Shotadapter: Text-to-multi-shot video generation with diffusion models. In *CVPR*, 2025. 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[22] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[23] LAION-AI. aesthetic-predictor. https://github.com/LAION-AI/aesthetic-predictor, 2022. 3

[24] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 3

[25] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *CVPR*, 2025. 3

[26] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 4

[27] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 5

[28] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videostudio: Generating consistent-content and multi-scene videos. In *ECCV*, 2024. 2

[29] Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Yixuan Li, Cheng Chen, Yanhong Zeng, et al. Holocine: Holistic generation of cinematic multi-shot long video narratives. *arXiv preprint arXiv:2510.20822*, 2025. 2

[30] Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In *CVPR*, 2024. 3

[31] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. In *WACV*, 2024. 2

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2

[33] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019. 3

[34] Quynh Phung, Long Mai, Fabian David Caba Heilbron, Feng Liu, Jia-Bin Huang, and Cusuh Ham. Cineverse: Consistent keyframe synthesis for cinematic scene composition. *arXiv preprint arXiv:2504.19894*, 2025. 3

[35] 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. *arXiv preprint arXiv:2307.01952*, 2023. 2

[36] 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 *ICML*, 2021. 4

[37] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 2023. 3, 5

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[39] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3

[40] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[41] Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *ACMMM*, 2024. 3

[42] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6

[43] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, 2024. 3

[44] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 5, 6, 7

[45] Bo Wang, Haoyang Huang, Zhiying Lu, Fengyuan Liu, Guoqing Ma, Jianlong Yuan, Yuan Zhang, Nan Duan, and Daxin Jiang. Storyanchors: Generating consistent multi-scene story frames for long-form narratives. *arXiv preprint arXiv:2505.08350*, 2025. 2

[46] Fu-Yun Wang, Yunhao Shui, Jingtan Piao, Keqiang Sun, and Hongsheng Li. Diffusion-npo: Negative preference optimization for better preference aligned generation of diffusion models. *arXiv preprint arXiv:2505.11245*, 2025. 3

[47] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 3

[48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022. 5

[49] Shuchen Weng, Peixuan Zhang, Zheng Chang, Xinlong Wang, Si Li, and Boxin Shi. Affective image filter: Reflecting emotions from text to images. In *ICCV*, 2023. 2

[50] Shuchen Weng, Haojie Zheng, Peixuan Zhang, Yuchen Hong, Han Jiang, Si Li, and Boxin Shi. Vires: Video instance repainting via sketch and text guided generation. In *CVPR*, 2025. 2

[51] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 6

[52] Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent cot planning. *arXiv preprint arXiv:2503.07314*, 2025. 2, 3, 6, 7

[53] Xiaoxue Wu, Bingjie Gao, Yu Qiao, Yaohui Wang, and Xinyuan Chen. Cinetrans: Learning to generate videos with cinematic transitions via masked diffusion models. *arXiv preprint arXiv:2508.11484*, 2025. 1, 2, 3, 6, 7

[54] Jinheng Xie, Jiajun Feng, Zhaoxu Tian, Kevin Qinghong Lin, Yawen Huang, Xi Xia, Nanxu Gong, Xu Zuo, Jiaqi Yang, Yefeng Zheng, et al. Learning long-form video prior via generative pre-training. *arXiv preprint arXiv:2404.15909*, 2024. 3

[55] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *ECCV*, 2024. 2

[56] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *NeurIPS*, 2023. 3

[57] Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Ying-Cong Chen. Seed-story: Multimodal long story generation with large language model. In *ICCV*, 2025. 2, 3

[58] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2

[59] Peixuan Zhang, Shuchen Weng, Jiajun Tang, Si Li, and Boxin Shi. Towards deeper emotional reflection: Crafting affective image filters with generative priors. *IEEE transactions on pattern analysis and machine intelligence*, 2025. 2

[60] Peixuan Zhang, Shuchen Weng, Chengxuan Zhu, Binghao Tang, Zijian Jia, Si Li, and Boxin Shi. Affective image editing: Shaping emotional factors via text descriptions. *International Journal of Computer Vision*, 2026. 2

[61] Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequence. *arXiv preprint arXiv:2407.16655*, 2024. 2

[62] Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, and Lim Sernam. Videogen-of-thought: A collaborative framework for multi-shot video generation. In *NeurIPS NextVid Workshop Oral*, 2025. 2, 6, 7

[63] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *NeurIPS*, 2024. 2, 3, 6, 7

Supplmentary Material:
STAGE: Storyboard-Anchored Generation for Cinematic Multi-shot Narrative

Peixuan Zhang[#1] Zijian Jia[#1] Kaiqi Liu[2,3] Shuchen Weng[*4,2] Si Li[*1] Boxin Shi[2,3]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications

[2]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University

[3]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

[4]Beijing Academy of Artificial Intelligence

{pxzhang,jiazijian,lisi}@bupt.edu.cn, liukq04@gmail.com, {shuchenweng,shiboxin}@pku.edu.cn

8. Appendix

8.1. Motivation of start-end frame pairs

While existing video generation models have demonstrated significant advances in high-fidelity text-to-video alignment, they struggle to meet user requirements due to the inherent ambiguity of text descriptions. This often leads to inconsistent intra-shot appearance (*e.g.*, identity drift) and jarring inter-shot transitions (*e.g.*, motion discontinuity).

Recognizing that the visual modality offers richer semantics to guide generation, we reformulate the keyframe-based approach to predict start-end frame pairs of each shot. These frame pairs act as powerful visual anchors to ground the generation process and explicitly define the scene boundaries of each shot. As a result, this formulation inherently enables structured narrative control by ensuring intra-shot coherence and facilitating cinematically plausible inter-shot transitions. We have illustrated the advantages of our formulation in Fig. 1 of the main paper.

8.2. STAGE workflow details

As illustrated in Sec. 4.3 of the main paper, the STAGE workflow begins with the director agent, which generates an initial storyboard from the user's input. This storyboard is then fed to the **STEP^2 model** for iterative start-end frame generation. Subsequently, we employ a shot refiner agent to bridge the gap between static frames and dynamic video by creating enhanced prompts. For better reproducibility, we provide the specific prompts used for the director and shot refiner agents in Fig. 7.

8.3. Evaluation metrics details

As illustrated in Sec. 5.2 of the main paper, we utilize eight quantitative metrics and four VLM-based evaluation scores to evaluate STAGE's performance across five key aspects. We list them as follows:

- **Overall video quality.** We assess the perceptual quality of generated videos using two metrics from VBench [3]:
  - **Aesthetic Quality (AQ)** quantifies the artistic appeal and visual composition of video frames using the LAION aesthetic predictor [5].
  - **Imaging Quality (IQ)** measures technical fidelity of video frames by detecting artifacts (*e.g.*, blur, noise, and distortion) with the MUSIQ model [4].
- **Text video consistency.** We measure the alignment between the generated video and the input text using a metric from VBench [3]:
  - **Overall Consistency (OC)** uses ViCLIP [11] to compute a video-text similarity score, reflecting semantic adherence to the prompt.
- **Intra-shot coherence.** We evaluate the temporal consistency within individual shots using VBench metrics [3]:
  - **Subject Consistency (SC)** measures the stability of a subject's appearance and identity throughout a single shot via DINO feature similarity [1].
  - **Background Consistency (BC)** assesses scene stability using CLIP features [7] to penalize background flickering or illogical changes.
- **Inter-shot coherence.** Since standard benchmarks lack metrics for evaluating long-range coherence in multi-shot videos, we introduce two metrics by extending VBench's [3] intra-shot consistency principles:
  - **Subject Consistency-Extra (SC-E).** We manually annotate shot pairs that are expected to share the same subject. For each annotated pair, we compute the DINO-based similarity for both their start frames and their end frames. The final score is the average of similarity across all annotated pairs.
  - **Background Consistency-Extra (BC-E).** We manually annotate shot pairs that are expected to share the same background. For each annotated pair, we compute the CLIP-based similarity for both their start

Prompts for Director Agents

Step1: Narrative & Visual Intent Analysis

You are a professional film director. First, carefully analyze the plot or scene provided by the user.
Please think through and output following these steps:
1. Core Narrative Identification
  - What is the core conflict or emotion of this scene?
  - Who are the key characters? What are their states and relationships?
  - What is the time, location, and environmental atmosphere of the scene?
2. Narrative Rhythm Analysis
  - How many narrative segments can this scene be divided into?
  - How does the emotional intensity change in each segment?
  - What are the key turning points or emotional climaxes?
3. Visual Storytelling Strategy
  - What overall visual style suits this scene? (Realistic/Dreamlike/Oppressive/Bright, etc.)
  - What visual elements need emphasis? (Character expressions/Environmental details/Actions/Spatial relationships)
  - How should the audience's perspective be guided? (Objective observer/Subjective immersion/God's eye view)

Step2: Shot-Level Narrative Decomposition

Based on the previous plot analysis, now divide the scene into specific shots.
Please think according to the following principles:
1. Shot Count Control
  - Plan 3-10 shots based on scene complexity
2. Visual Continuity Planning
  - Ensure compliance with the 180-degree rule
  - Plan match cuts between shots (action/eyeline/direction)
  - Design progressive shot size changes (e.g., Wide→Medium→Close or Close→Wide)

Step3: Technical Cinematography Decision Making

Now determine the specific technical parameters for each shot.
For each shot, reason through the following logic:
A. Shot Size Selection
Thinking logic:
- Need to show spatial relationships/environment → Extreme Wide Shot/Wide Shot
- Show character full-body actions → Full Shot
- Show character interaction and dialogue → Medium Shot
- Capture facial expressions and emotions → Close-Up/Medium Close-Up
- Emphasize details or emotional peaks → Big Close-Up/Extreme Close-Up
B. Camera Position & Angle
Thinking logic:
- Objective narrative/equal relationship → Eye-level
- Show character vulnerability/oppression → High angle
- Show character dominance/threat → Low angle
- Subjective perspective/dialogue → Over-the-shoulder/Shot-reverse-shot
- Observer perspective → Side angle
C. Camera Movement
Thinking logic:
- Stable, objective narrative → Static shot (Tripod)
- Gradually focus/intensify emotion → Push in
- Reveal information/emotional release → Pull out
- Follow action/maintain continuity → Tracking
- Display space/establish scene → Pan/Aerial
- Tension/chaos/subjective feeling → Handheld
- Smooth spatial transition → Dolly/Crane

Step4: Detailed Shot Execution Specification

Based on all previous analysis and decisions, now generate detailed shooting descriptions for each shot.
Each shot description must include:
1. Shot Content Core
  - What happens in this shot? (Specific events, actions, dialogue)
  - Who/what is the main subject of the frame?
  - What is the key narrative information or emotional change?
2. Visual Elements Checklist
  - Foreground: What specific objects/elements are present?
  - Midground: Main visual focus (characters/objects and their states)
  - Background: Environmental details, spatial depth, atmospheric creation
  - Lighting: Light source type (natural/artificial), direction, contrast, color temperature, atmospheric feel
3. Character Performance (if characters present)
  - What is the character doing? (Specific actions, precise body language)
  - Facial expressions and emotional state (eyes, micro-expressions, muscle tension)
  - Costume, hair, makeup details
  - Character's position in frame and spatial relationships
  - Character's movement trajectory and speed changes
4. Camera Execution Plan
  - Starting position: Location, height, angle, composition
  - Movement type: Static/push-pull/pan-tilt/follow (specific path)
  - Movement speed: Slow/steady/accelerating/rapid
  - Ending position: Final composition and visual focus
  - Focus control: Where is the focus? Is there focus shift?
5. Narrative & Emotional Function
  - What new information does this shot reveal?
  - What part of the plot does it advance?
  - What emotion does it create or reinforce? (Tense/warm/oppressive/relaxed, etc.)
  - How does it connect with the previous shot? (Action match/eyeline match/spatial continuity)
  - What groundwork does it lay for the next shot?
6. Technical Details Supplement
  - Recommended focal length (16mm/35mm/50mm/85mm/200mm, etc.)
  - Depth of field treatment (shallow DOF blur background/deep DOF all sharp)
  - Aperture, shutter speed recommendations (if special needs)
  - Special effects, filters, color grading (if needed)
Description writing principles:
- First clarify "what happens" — this is the shot's core
- Specific and executable (cinematographers can directly understand and shoot)
- Avoid vague terms ("some," "certain," "approximately," "maybe," etc.)
- Highlight this shot's unique visual characteristics and narrative function

Step5: Final Integrated Shot List Construction

Now integrate all previous stages of analysis and decisions into a complete shot list.
Please ensure:
1. All shots comply with the 180-degree rule
2. Shot size changes are reasonable (avoid jump cuts)
3. Clear match cut points between shots (action/eyeline/direction)
4. Rhythm aligns with narrative needs

Prompts for Shot Refiner Agent

You are a professional video storyboard artist. The user will provide the first and last frame images of a shot along with a preliminary description. You need to expand this into a complete prompt suitable for video generation models.
Task:
1. Based on the first and last frames, supplement complete visual and action details
2. Clearly describe the motion trajectory and transformation process from start to finish
3. Retain all key elements from the user's original description
Format Requirements:
- Single coherent text block, organized in chronological order
- Use present tense to describe actions
- Output the final prompt directly, without any prefix or explanation

Figure 7. Specific prompts for the director and shot refiner agents.

Prompts for Multi-Shot Video Quality Evaluation

Overall Video Quality

Assesses the overall stability, coherence, and consistency of the video across both individual shots and the full sequence. A score of 1 indicates that all major consistency dimensions—such as intra-shot stability of subject and background, inter-shot continuity of appearance and spatial logic, and uniformity of visual style—are maintained without disruptions throughout the video. A score of 0 indicates the presence of unstable shot-internal elements, mismatched subjects or environments between shots, significant visual style inconsistencies, etc., resulting in a viewing experience that lacks any reliable or coherent sense of continuity.

Video Text Consistency

Assesses how faithfully the video reflects the narrative content and storyline described in the text. A score of 1 indicates that key narrative elements—such as the intended plot progression, character actions and interactions, event sequencing, and overall story logic—are accurately represented in the video without contradictions or omissions. A score of 0 indicates the presence of incorrect or missing plot events, misaligned character behavior, altered story progression, or other substantial deviations from the described narrative intent, resulting in a video that does not reliably correspond to the textual storyline.

Inter-Shot Consistency

Assesses how smoothly the transition between two shots preserves subject consistency, background logic, and visual style coherence. A score of 1 indicates that all key continuity aspects—such as subject appearance and motion, spatial relationships in the environment, and overall visual styling—remain coherent and uninterrupted across the cut. A score of 0 indicates clear discontinuities in these elements, such as mismatched subject attributes, background inconsistencies, or noticeable shifts in color or lighting that disrupt the sense of continuity.

Shot Transition Smoothness

Assesses the smoothness, temporal stability, and motion continuity of the video across both individual shots and the full sequence. A score of 1 indicates that frame timing, subject motion, camera movement, and shot-to-shot temporal alignment remain consistently smooth without noticeable interruptions. A score of 0 indicates the presence of issues such as stuttering, frame drops, mismatched motion between shots, and significant temporal inconsistencies, resulting in a viewing experience that lacks any reliable sense of fluidity.

Figure 8. Evaluation prompts for the VLM-based evaluation metrics (OVQ, VTC, ISC, and STS).

frames and their end frames. The final score is the average of similarity across all annotated pairs.

- **Inter-shot transition.** We evaluate the quality of inter-shot transitions about logical narrative and visual coherence in multi-shot video generation.
  - **Transition Vector Similarity (TVS)** assesses the consistency between the semantic trajectory of a generated transition and that of its ground truth counterpart. Specifically, inspired by prior research [2, 6, 13] that the distance in CLIP space corresponds to semantic shift, we define a transition vector at each shot boundary as the difference between the CLIP embeddings of the new shot's first and the previous shot's last frame. This TVS score is then calculated as the cosine similarity between the generated transition vector and its corresponding vector from the ground truth video.

As presented in Tab. 1, we employ the Gemini [8] to complement our quantitative metrics with LLM-based evaluations. For better reproducibility, we present the evaluation prompts for Overall Video Quality (OVQ), Video Text Consistency (VTC), Inter-Shot Consistency (ISC), and Shot Transition Smoothness (STS) in Fig. 8.

During quantitative comparisons, we adopt Wan2.2-14B [10] as the video generation backbone to ensure a fair evaluation against state-of-the-art methods.

8.4. Ranking strategy of multi-shot memory pack

As illustrated in Sec. 4.1 of the main paper, the multi-shot memory pack is designed to ensure long-term temporal consistency by compressing preceding shots into a compact context. Instead of prioritize chronological order to highlight more recent scenes [9, 12], we rank preceding frames

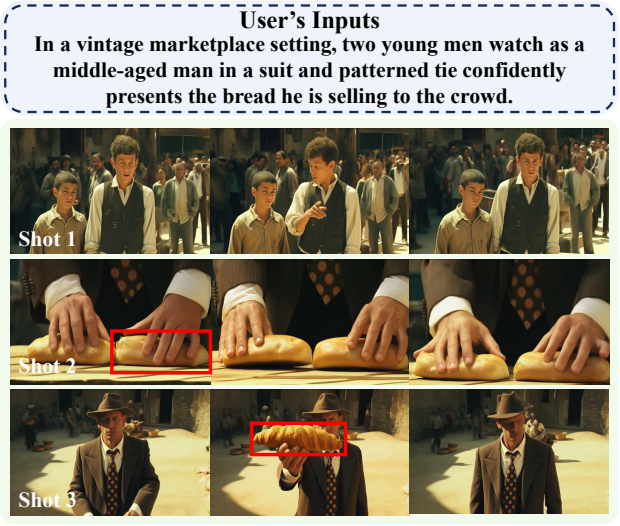

Figure 9. Visualization of failure cases.

based on their CLIP similarity, which prioritizes semantic relevance. To validate the efficacy of this similarity-based ranking, we conduct an additional ablation study. As results reported in Tab. 3, our ranking strategy demonstrates superior performance for the multi-shot cinematic video generation task.

Table 3. Ablation study for the multi-shot memory pack.

Methods	SC ↑	BC ↑	SC-E ↑	BC-E ↑	ISC ↑
W/o Ranking	0.9649	0.9557	0.6378	0.7792	0.6589
Ours (STAGE)	**0.9695**	**0.9685**	**0.6917**	**0.8207**	**0.6985**

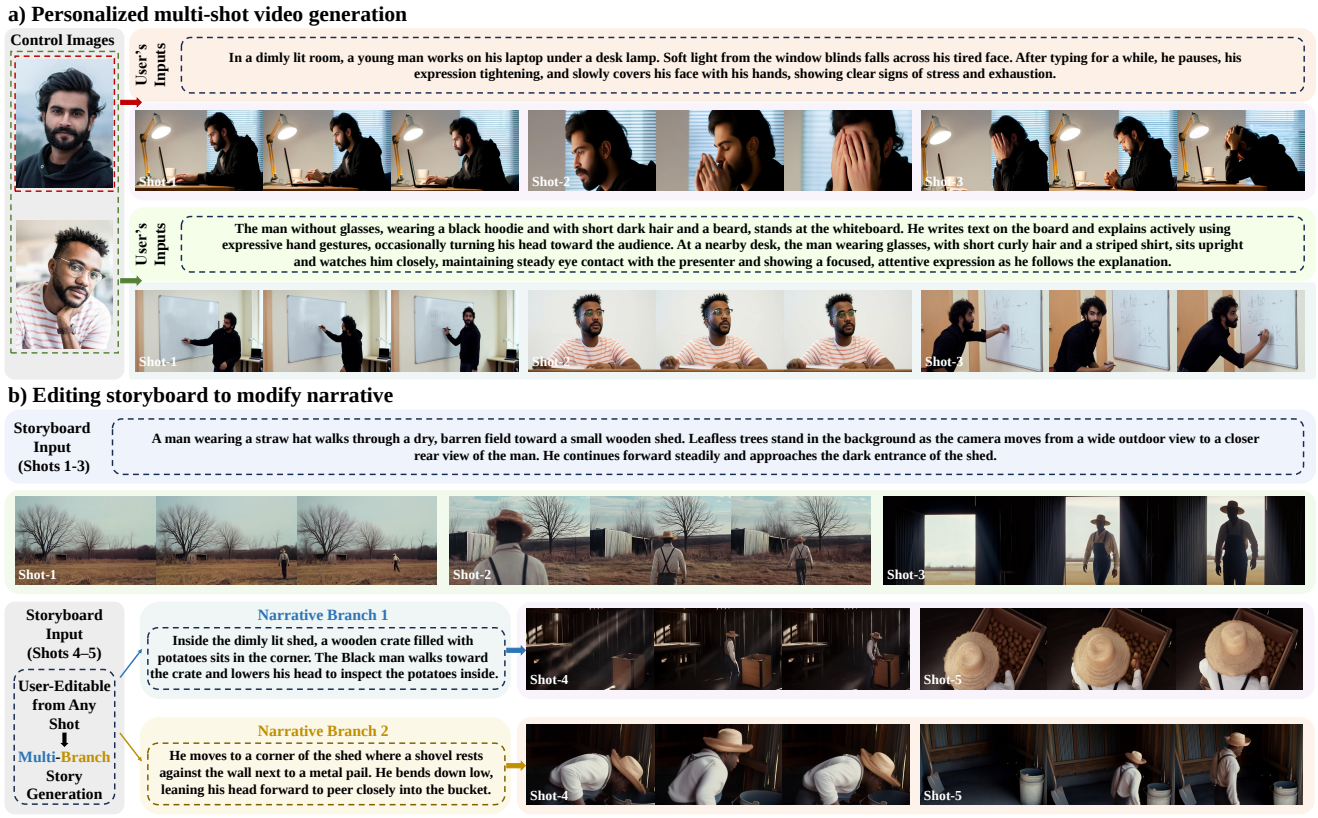

Figure 10. Additional application scenarios of our STAGE.

8.5. Failure case

As illustrated in Sec. 6 of the main paper, STAGE is designed to maintain long-term narrative and entity consistency across multiple shots according to start-end frame pairs. However, the internal video clip generation is handled by a pre-trained video generation model, which may introduce temporal inconsistencies within the generated clips in fine-grained visual details. Specifically, we present a failure case in Fig. 9. While our framework provides start and end frames to anchor the main visual appearance, the bread only appears in intermediate frames. This results in its appearance being under-specified and prone to changes due to text ambiguity.

8.6. Application

We show two representative application scenarios with our STAGE framework to highlight its advanced capability for controllable and flexible video creation.

Personalized multi-shot video generation. As illustrated in Fig. 10 (a), our STAGE framework can generate videos with customized person and scene. Given user-provided one or more reference images for a specific subject, STAGE leverages the multi-shot memory pack to ensure the subject is faithfully rendered with high consistency across varying contexts and actions. This enables generating personalized stories featuring user-specified characters.

Editing storyboard to modify narrative. As illustrated in Fig. 10 (b), our STAGE framework supports an interactive editing workflow. Given a storyboard for each shot, the user can freely modify text descriptions to edit the story progress and character performance. Our framework can re-synthesize the start-end frame pairs and generate the corresponding video clip, enabling seamless narrative branching while preserving the temporal consistency of entities. This capability enables fine-grained video design.

8.7. Additional dataset samples

As presented in Sec. 3 of the main paper, we introduce the ConStoryBoard dataset for training and evaluation. As showcase the diversity and scope of our data, we visualize several randomly selected samples in Fig. 11.

8.8. Organization of supplementary video

We provide a supplementary video to dynamically showcase our cinematic multi-shot video results. The video is structured as follows:

- **Multi-shot video generation:** We begin by presenting

qualitative results for our multi-shot video generation task (corresponding to Fig. 3 in the main paper).

- **Long multi-shot video generation:** We then demonstrate our method's capability in generating a 9-shot long video with 1080 frames (corresponding to Fig. 6 in the main paper).
- **Qualitative Comparisons:** We provide comparisons against state-of-the-art methods to highlight the advantages of our approach (corresponding to Fig. 1 and Fig. 4 in the main paper).
- **Applications:** We showcase additional application scenarios of our method (corresponding to Fig. 10).
- **Ablation Studies:** We present our ablation study results to demonstrate the contribution of each proposed component (corresponding to Fig. 5 in the main paper).
- **Failure Cases:** Finally, we include a failure case to demonstrate the future improvement (corresponding to Fig. 9).

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1

[2] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[3] 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 *CVPR*, 2024. 1

[4] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021. 1

[5] LAION-AI. aesthetic-predictor. https://github.com/LAION-AI/aesthetic-predictor, 2022. 1

[6] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3

[7] 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 *ICML*, 2021. 1

[8] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[9] 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. 3

[10] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3

[11] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 1

[12] Wilson Yan, Danijar Hafner, Stephen James, and Pieter Abbeel. Temporally consistent transformers for video generation. In *ICML*, 2023. 3

[13] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jiahui Zhang, Shijian Lu, Miaomiao Cui, Xuansong Xie, Xian-Sheng Hua, and Chunyan Miao. Towards counterfactual image manipulation via clip. In *ACMMM*, 2022. 3

Figure 11. Additional samples presentation of the ConStoryBoard dataset.