VideoGen-of-Thought: Step-by-step generating multi-shot video with minimal manual intervention

Mingzhe Zheng[1,6] **Yongqi Xu**[2,6] **Haojian Huang**[3] **Xuran Ma**[1,6] **Yexin Liu**[1,6] **Wenjie Shu**[1,6]

Yatian Pang[4,6] **Feilong Tang**[1,6] **Qifeng Chen**[1,†] **Harry Yang**[1,6,†] **Ser-Nam Lim**[5,6,†]

[1] Hong Kong University of Science and Technology [2] Peking University
[3] University of Hong Kong [4] NUS [5] University of Central Florida [6] Everlyn AI

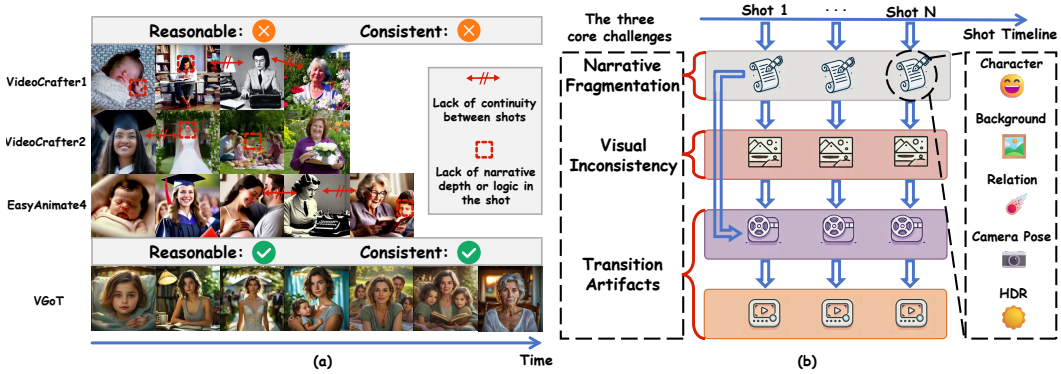

Figure 1: **Illustration of *VideoGen-of-Thought (VGoT)*. (a) Comparison of existing methods with *VGoT* in multi-shot video generation.** Existing methods struggle with maintaining reasonability and consistency across multiple shots, while *VGoT* effectively addresses these challenges through a multi-shot generation approach. **(b) Challenges solved by *VGoT***: addressing *narrative fragmentation* with dynamic storylines modeling across five domains *(characters/backgrounds/relations/camera/HDR)*, tackling *visual inconsistency* via identity-aware cross-shot propagation to create keyframes using IPP tokens derived from narrative elements, and solving *transition artifacts* during multi-shot video synthesizes through adjacent latent transition mechanisms.

Abstract

Current video generation models excel at short clips but fail to produce cohesive multi-shot narratives due to disjointed visual dynamics and fractured storylines. Existing solutions either rely on extensive manual scripting/editing or prioritize single-shot fidelity over cross-scene continuity, limiting their practicality for movie-like content. We introduce **VideoGen-of-Thought (VGoT)**, a step-by-step framework that automates multi-shot video synthesis **from a single sentence** by systematically addressing three core challenges: **(1) Narrative fragmentation**: Existing methods lack structured storytelling. We propose dynamic storyline modeling, which turns the user prompt into concise shot drafts and then expands them into detailed specifications across five domains (character dynamics, background continuity, relationship evolution, camera movements, and HDR lighting) with self-validation

† Corresponding author.
Project webpage: https://cheliosoops.github.io/VGoT/

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: What Makes a Good Video: Next Practices in Video Generation and Evaluation.

to ensure logical progress. **(2) Visual inconsistency**: previous approaches struggle to maintain consistent appearance across shots. Our identity-aware cross-shot propagation builds identity-preserving portrait (IPP) tokens that keep character identity while allowing controlled trait changes (expressions, aging) required by the story. **(3) Transition artifacts**: Abrupt shot changes disrupt immersion. Our adjacent latent transition mechanisms implement boundary-aware reset strategies that process adjacent shots' features at transition points, enabling seamless visual flow while preserving narrative continuity. Combined in a training-free pipeline, VGoT surpasses strong baselines by 20.4% in within-shot face consistency and 17.4% in style consistency, while requiring **10× fewer** manual adjustments. VGoT bridges the gap between raw visual synthesis and director-level storytelling for automated multi-shot video generation.

1 Introduction

Recent advancements in video generation techniques have yielded impressive results, particularly in creating short, visually appealing clips Blattmann et al. (2023a); Chen et al. (2023, 2024); Xu et al. (2024); Henschel et al. (2024). These advancements have been powered by increasingly sophisticated generative models, ranging from diffusion models Ho et al. (2020); Song et al. (2020b); Rombach et al. (2022); Blattmann et al. (2023a) to auto-regressive models Ge et al. (2022); Weng et al. (2023); Liu et al. (2024); Wang et al. (2024), supported by large-scale datasets Huang et al. (2020); Schuhmann et al. (2021, 2022). These methods have enabled the generation of high-quality and realistic short videos. However, generating multi-shot videos from a brief user input script remains a substantial challenge. Unlike single-shot video generation, which focuses on creating a coherent clip from a single prompt, multi-shot video generation requires the model to maintain both **reasonable storylines** and **visual consistency** across multiple shots. This task involves additional complexities, such as ensuring logical transitions between scenes and maintaining consistent appearance *(e.g., character identity, overall style, etc.)* throughout the video. Current video generation methods Chen et al. (2023, 2024); Xu et al. (2024); Hong et al. (2022); Yang et al. (2024) often fall short in these areas, resulting in fragmented narratives and inconsistent visual elements across shots. Furthermore, sometimes real movies require the same character to appear in different ways based on the storylines, which should be faithful to the same identity but not the same traits *(e.g., expression, appearance, relationships, etc.).* The requirement for high-level identity preservation Ye et al. (2023); Yuan et al. (2024); Zhou et al. (2025) across shots remains an open question, which is essential for cross-shot consistency.

Existing multi-shot video generation approaches suffer from several limitations. For instance, methods like MovieDreamer Zhao et al. (2024) require plenty of manual input, including mountains of script writing *(e.g., character appearance, scene elements, detailed plots, etc.)* and image selection. Other approaches, such as DreamFactory Xie et al. (2024), focus on multi-agent pipelines but require specific documents and repeated manual adjustment for each story, restricting the capability of easy usage. The need for heavy manual intervention not only increases the workload but also limits their practicality for automated movie-like content creation. In contrast, our work aims to decompose the complex task of multi-shot video generation into smaller, manageable problems and solve them in a step-by-step manner with minimal manual intervention, as shown in Figure 1. Our approach proposes to systematically and automatically address three core challenges: **(1) narrative fragmentation** through dynamic storyline modeling; **(2) visual inconsistency** via identity-aware cross-shot propagation; and **(3) transition artifacts** using adjacent latent transition mechanisms.

We propose **VideoGen-of-Thought (*VGoT*)**, an end-to-end framework that generates multi-shot video with *reasonable storylines* and *visual consistency* **from one sentence** with **minimal manual intervention**. Our framework addresses three fundamental challenges through a structured pipeline as shown in Fig 2: First, *VGoT* tackles *narrative fragmentation* through converting a brief user prompt into short descriptions for across shots to obtain a reasonable story draft: we introduce a dynamic storyline modeling that transforms user prompts into shot sequences through a two-stage process with self-validation mechanisms that enforce narrative coherence by rejecting candidates violating cinematic principles.

Additionally, to resolve *visual inconsistency*, we introduce identity-aware cross-shot propagation that extracts multi-aspect portrait schemata from narrative elements, ranging from different avatars in the same story to different traits of the same identity following the development of the story.

This system handles both inter-story avatar variations and intra-story identity evolution, generating identity-preserving portrait (IPP) tokens. These IPP tokens guide keyframe generation through hierarchical feature injection in pretrained diffusion models Song et al. (2020a); Rombach et al. (2022); Team (2024); Ye et al. (2023), ensuring style uniformity and identity fidelity across shots.

We encode each keyframe into latent space and utilize a video diffusion model to refine a noise map into video latent codes, representing k frames of shot video, conditioned on the keyframe latent and the corresponding textual latent. To address *transition artifacts*, we design a cross-shot transition mechanism with a FIFO-like Kim et al. (2024) latent reset strategy that processes adjacent shots' features at the boundary, ensuring seamless transitions and maintaining visual continuity across shots while preserving the logical coherence established in the storyline preparation.

Additionally, current evaluation protocols for multi-shot video generation remain inadequate due to the absence of dedicated datasets and task-specific metrics. To address this limitation, we propose four novel quantitative metrics for systematic assessment:

- **Within-Shot Face Consistency (WS-FC):** facial similarity between frames in a single shot.
- **Cross-Shot Face Consistency (CS-FC):** identity distance across shots.
- **Within-Shot Style Consistency (WS-SC):** style similarity between frames in a single shot.
- **Cross-Shot Style Consistency (CS-SC):** stylized bias across shots.

Experimental results demonstrate VGoT's superiority over state-of-the-art (SOTA) methods across all metrics. Quantitative comparisons reveal **20.4%** and **17.4%** improvements in WS-FC and WS-SC, respectively, compared to previous SOTA baselines. For cross-shot metrics, VGoT achieves **2.9**$\times$ higher in CS-FC and **106.6%** higher in CS-SC over baselines. Human evaluations (Table 2) confirm these findings, with VGoT receiving **66.7%** "Good" ratings for cross-shot consistency versus 27.2% for competitors.

The principal contributions of this work can be summarized as follows:

- **Automated Multi-Shot Generation Framework:** We present VGoT, the first end-to-end system that generates story-coherent multi-shot videos from single sentence inputs, obviously reducing manual intervention.
- **Three-Core Solution Architecture:** We propose a structured four-module pipeline addressing *narrative fragmentation* through LLM-powered story decomposition, *visual inconsistency* via identity-aware propagation, and *transition artifacts* using adjacent latent transition mechanisms.
- **Multi-Shot Evaluation Protocol:** We design a new multi-shot video assessment protocol featuring hierarchical consistency measurement with four quantitative metrics covering face/style consistency across both within-shot and cross-shot.

2 Related Work

Video generation has made significant strides following the great success of diffusion models, leading to two important research categories: long video synthesis and multi-shot story generation. These areas focus on generating high-quality, consistent videos either as extended single shots or as coherent sequences across multiple scenes.

Long Video Synthesis. Long video synthesis has advanced through diffusion-based methods and autoregressive approaches. Diffusion models based on Stable Diffusion Rombach et al. (2022) utilize iterative refinement to generate visually consistent frames and have been effective for short sequences He et al. (2022); Blattmann et al. (2023a); Chen et al. (2023, 2024); Xing et al. (2025); Wu et al. (2023); Blattmann et al. (2023b); Guo et al. (2023); Yang et al. (2024); Zhang et al. (2024); Geyer et al. (2023); Huang et al. (2024); Peng et al. (2024); Esser et al. (2023); Ho et al. (2022); Singer et al. (2022); Zeng et al. (2024); Zhou et al. (2022); Qiu et al. (2023); Pan et al. (2024). However, they struggle with maintaining coherence over extended video lengths. Autoregressive methods Ge et al. (2022); Li et al. (2024); Yin et al. (2023) predict frames sequentially but often face error accumulation, making long-term consistency challenging and computationally expensive. Additionally, training-free methods like FIFO-Diffusion Kim et al. (2024) generate long sequences

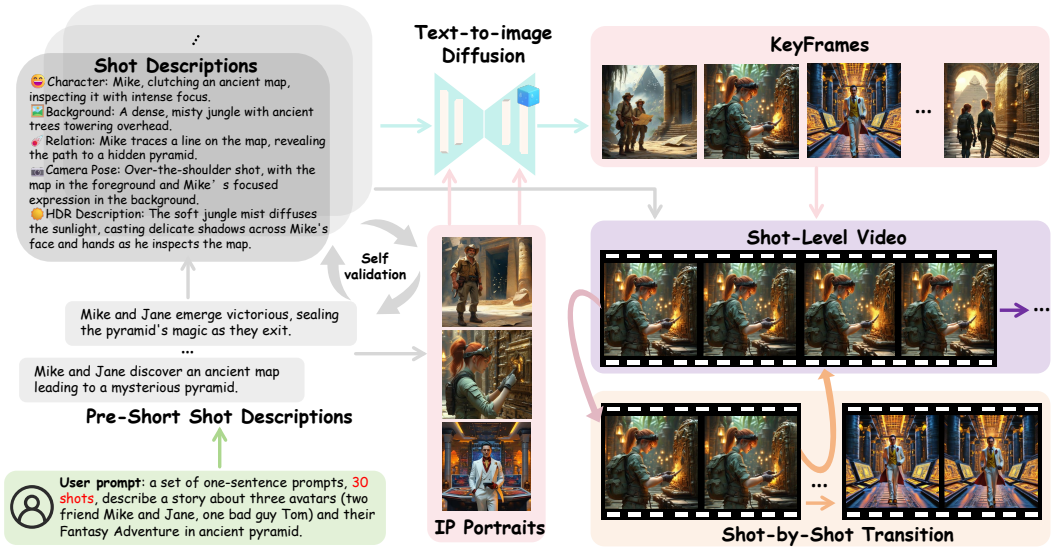

Figure 2: **The FlowChart of *VideoGen-of-Thought*. Left:** Shot descriptions are generated based on user prompts, describing various attributes such as character details, background, relations, and camera pose. Pre-shot descriptions provide a broader context for the upcoming scenes. **Middle Top:** Keyframes are generated using a text-to-image diffusion model conditioned with identity-preserving (IP) embeddings, which ensures consistent representation of characters throughout the shots. IP portraits help maintain visual identity consistency. **Right:** The shot-level video clips are generated from keyframes, followed by shot-by-shot transition inference to ensure temporal consistency across different shots. This collaborative framework ultimately produces a cohesive narrative-driven video.

without training but lack mechanisms to manage transitions across shots, limiting their effectiveness in narrative-driven content. Overall, while these approaches achieve visual fidelity, they fail to ensure logical coherence across extended sequences. In contrast, VideoGen-of-Thought (*VGoT*) leverages a modular approach that includes cross-shot smoothing mechanisms to ensure both visual consistency and narrative coherence, offering a more holistic solution for generating long-form videos.

Multi-Shot Video Generation. Multiple Shot Story Generation focuses on maintaining narrative coherency across distinct scenes, and existing approaches face critical limitations in automation scalability. Animate-a-Story He et al. (2023) uses retrieval-augmented generation to ensure visual consistency but struggles with maintaining logical narrative transitions. MovieDreamer Zhao et al. (2024) requires extensive manual scripting *(character profiles, scene details, etc.)* and image curation. DreamFactory Xie et al. (2024) demands repetitive adjustment and special documents for each story in its multi-agent system across stories. Flexifilm Ouyang et al. (2024) and StoryDiffusion Zhou et al. (2025) introduce conditional adaptability but still necessitate manual consistency fixes between shots. These methods share a common bottleneck: heavy reliance on human intervention for narrative and visual coherence. *VGoT* breaks this paradigm through systematic automation: 1) *Narrative Fragmentation* $\rightarrow$ LLM-driven story decomposition; 2) *Visual Inconsistency* $\rightarrow$ Story-derived identity propagation; 3) *Transition Artifacts* $\rightarrow$ Boundary-aware latent processing. This structured approach enables director-level storytelling from single prompts with 10× fewer manual interventions than prior works Zhao et al. (2024); Xie et al. (2024).

3 Preliminaries and Problem Formulation

3.1 Preliminaries

Diffusion models Ho et al. (2020); Song et al. (2020a,b) are generative models trained to approximate data distributions $p(x)$ through iterative denoising of random noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The forward process gradually adds noise according to a variance schedule β_t:

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right) \tag{1}$$

over T timesteps, producing progressively noisier latents $\{x_t\}_{t=1}^{T}$. The reverse process learns a parameterized model μ_θ to reconstruct x_0 through transitions:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right) \tag{2}$$

where μ_θ and Σ_θ denote the predicted mean and variance. Training minimizes the noise prediction error via:

$$\mathcal{L}_{\text{uncond}} = \mathbb{E}_{x_0, \epsilon\sim\mathcal{N}(0,\mathbf{I}), t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right] \tag{3}$$

Latent diffusion models Rombach et al. (2022) map this process to a compressed space using a VAE Kingma and Welling (2013), enabling conditional generation through cross-attention:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{4}$$

where $Q = W_Q\mathcal{E}(x_t)$, $K = W_K\tau_\theta(y)$, and $V = W_V\tau_\theta(y)$ for text embeddings y. Video diffusion models He et al. (2022); Blattmann et al. (2023a) process frame sequences $\{z^f\}_{f=0}^{F-1} \in \mathbb{R}^{h\times w\times d}$ through temporal-aware denoising networks $\epsilon_\theta(z_t, t, c)$. FIFO-Diffusion Kim et al. (2024) extends this through queue-based latent processing:

$$Q_k = \{z^f\}_{f=f_k}^{f_k+n} \leftarrow \Phi(Q_k, \tau_k, c; \epsilon_\theta) \tag{5}$$

where Φ denotes the DDIM sampler Song et al. (2020a). While effective for frame continuity, this approach struggles with: (1) abrupt shot transitions that disrupt queue coherence, and (2) integration of image-based conditions. Our framework addresses these through reset mechanisms in adjacent latent-space processing.

3.2 Problem Definition

Given a one-sentence user input S specifying N shots (*e.g., "A story of Mary's life from birth to death"*), we aim to generate a multi-shot video V with *reasonable storylines* and *visual consistency* **from one sentence** through **minimal manual intervention**. The core challenges are:

- **Reasonability**: Maintaining logical narrative flow across evolving storylines
- **Consistency**: Preserving high-level identity Ye et al. (2023); Yuan et al. (2024); Zhou et al. (2025) while allowing trait variations (*e.g., aging expressions, contextual relationships*) across shots
- **Multi-Shot Generation**: Producing minute-level videos with diverse yet interconnected shots

VideoGen-of-Thought (VGoT) addresses these through script preparation $\{p_i\}_{i=0}^{N-1}$, identity-preserved keyframes $\{I_i\}_{i=0}^{N-1}$, and cross-shot video latents $\{z^f\}_{f_i=0}^{F-1}$. Our framework overcomes limitations in existing multi-shot methods through narrative-visual coherency mechanisms.

4 Method: VideoGen-of-Thought

In this section, we introduce a structured, step-by-step framework for generating multi-shot videos with *reasonable storylines* and *visual consistency* **from one sentence** with **minimal manual intervention**, addressing the core challenges of narrative fragmentation, visual inconsistency, and transition artifacts through four distinct yet collaborative modules (Fig 2):

4.1 Dynamic Storyline Modeling with Self-Vaildation

We formulate narrative generation as constrained multi-shot decomposition with auto-regressive validation. Given user prompt S and shot count N, our system first generates a story draft $S' = \{s_i\}_{i=1}^{N}$ of short shot descriptions, then produces structured scripts through:

$$\mathcal{P} = \{p_i\}_{i=1}^{N} = \bigcup_{i=1}^{N} \mathcal{M}_{\text{LLM}}(s_i|\mathcal{C}_{\text{film}}, \{p_j\}_{j=1}^{i-1}) \tag{6}$$

where $\mathcal{C}_{\text{film}}(p_{\text{cha}}, p_b, p_r, p_{\text{cam}}, p_h)$ encodes five cinematic dimensions: p_{cha} governs character appearance evolution and role relationships, p_b ensures background consistency across scene transitions, p_r maintains interaction patterns and event causality, p_{cam} specifies shot composition through camera movements, and p_h regulates HDR lighting continuity. The self-validation mechanism employs two criteria:

$$\mathcal{V}(p_i) = \mathbb{I}[C(p_i, p_{i-1}) > \tau_c] \cdot \mathbb{I}[K(p_i, \mathcal{C}_{\text{film}}) > \tau_k] \tag{7}$$

where $C : \mathcal{P} \times \mathcal{P} \to [0, 1]$ measures narrative coherence through pretrained textual feature extractor E_{GLM} GLM et al. (2024) to compute semantic similarity between consecutive shots, and $K : \mathcal{P} \times \mathcal{C} \to \{0, 1\}$ verifies constraint completeness via rule-based checks against $\mathcal{C}_{\text{film}}$. Thresholds $\tau_c = 0.85$ and $\tau_k = 1$ ensure strict adherence to cinematic principles.

Algorithm 1 Self-Validated Script Generation

Require: User prompt S, shot count N, constraints $\mathcal{C}_{\text{film}}$
Initialize $\mathcal{P} \leftarrow \emptyset$, $p_{\text{prev}} \leftarrow \emptyset$
Generate draft $S' \leftarrow \{s_1, ..., s_N\} = \mathcal{M}_{\text{LLM}}(S, N)$
for $i \leftarrow 1$ to N **do**
  repeat
    $p'_i \leftarrow \mathcal{M}_{\text{LLM}}(s_i, \mathcal{C}_{\text{film}}, p_{\text{prev}})$
    $C_i \leftarrow \mathbb{I}[C(p_i, p_{i-1}) > \tau_c]$
    $K_i \leftarrow \mathbb{I}[K(p_i, \mathcal{C}_{\text{film}}) > \tau_k]$
    $\mathcal{V}(p_i) \leftarrow C_i \cdot K_i$
  until $\mathcal{V}(p'_i) = 1$
  $\mathcal{P} \leftarrow \mathcal{P} \cup p'_i$, $p_{\text{prev}} \leftarrow p'_i$
end for

Our dynamic storyline modeling transforms user prompts into shot sequences through a two-stage process with self-validation mechanisms that enforce narrative coherence by rejecting candidates violating cinematic principles (Algorithm 1).

4.2 Identity-Aware Cross-Shot Propagation

We resolve visual inconsistency through cross-shot propagation machanism, which maintains critical attributes (*e.g., hairstyle, facial structure*) while permitting narrative-driven variations (*e.g., expression, aging*). Using scripts $\mathcal{P}$, we generate keyframes with consistent visual identities through a two-stage process:

$$\mathbf{I} = \{\mathbf{I}_i\}_{i=1}^{N} = \mathcal{F}(\mathcal{P}, \Psi) \tag{8}$$

where $\mathcal{F}$ represents our identity-preserving generation pipeline and Ψ denotes the parameters of the character schema. For each shot script $p_i \in \mathcal{P}$, we extract:

$$e_i^T = E_{\text{GLM}}(p_i) \in \mathbb{R}^d, \quad \mathcal{C}_{\text{char}} = \mathcal{M}_{\text{LLM}}(\mathcal{P}) = \{c_j\}_{j=1}^{M} \tag{9}$$

where E_{GLM} is the text encoder and $\mathcal{C}_{\text{char}}$ contains M identity descriptors (e.g., *Young Mary, Elderly Mary* when describing given scripts *Mary's life*). Identity-Preserving Portrait (IPP) tokens are synthesized through:

$$\text{IPP}_j = \mathcal{M}_I(c_j) \in \mathbb{R}^{H \times W \times 3}, \quad e_j^I = E_{\text{CLIP}}(\text{IPP}_j) \in \mathbb{R}^{d_v} \tag{10}$$

where $\mathcal{M}_I$ is a pre-trained text-to-image model and E_{CLIP} denotes CLIP Radford et al. (2021)'s vision encoder. We inject identity features into diffusion via cross-attention:

$$Q = W_Q z_t \in \mathbb{R}^{n \times d_k} \tag{11}$$

$$K = \lambda[W_K e_i^T; W'_K e_j^I] \in \mathbb{R}^{2n \times d_k} \tag{12}$$

$$V = \lambda[W_V e_i^T; W'_V e_j^I] \in \mathbb{R}^{2n \times d_v} \tag{13}$$

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \tag{14}$$

Figure 3: **Visual showcases of *VGoT* generated multi-shot videos.**

where $z_t \in \mathbb{R}^{n \times d}$ is the latent in step t, λ balances the influence of text / identity, and $[;]$ denotes concatenation. Keyframe generation integrates both modalities:

$$I_i = \mathcal{D}\left(z_T \middle| e_i^T, e_{j(i)}^I\right), \quad j(i) = \mathcal{M}_{\text{LLM}}(p_i, \mathcal{C}_{\text{char}}) \tag{15}$$

By anchoring IPP tokens in narrative-derived character descriptors $\mathcal{C}_{\text{char}}$ and integrating them via attention mechanisms, we achieve robust identity fidelity across shots.

4.3 Adjacent Latent Transition Mechanisms

We address transition artifacts through latent-space noise management across shot boundaries. Given keyframes $\{I_i\}_{i=1}^N$ and script embeddings $\{e_i^T\}_{i=1}^N$, we generate shot-wise latents:

$$Z_i = \mathcal{M}_V(e_i^T, e_i^I, \epsilon_i) \in \mathbb{R}^{f \times c \times h \times w} \tag{16}$$

where $e_i^T = E_{\text{GLM}}(s_i)$ uses simplified shot description s_i rather than detailed script p_i, e_i^I is the keyframe embedding, and $\epsilon_i \sim \mathcal{N}(0, \mathbf{I})$ is the initial noise. Inspired by FIFO Kim et al. (2024), we implement boundary-aware noise reset for cross-shot transitions:

$$\epsilon_{\text{boundary}} \sim \mathcal{N}(0, \beta_i \mathbf{I}), \quad \beta_i = \gamma \cdot (1 - \frac{i}{N}) \tag{17}$$

$$Z_{\text{final}} = \mathcal{R}(Z_1, Z_2, \ldots, Z_N) \tag{18}$$

where β_i controls noise magnitude at shot boundaries with scaling factor γ, and $\mathcal{R}$ denotes our reset function:

$$\mathcal{R}(Z_1, \ldots, Z_N) = \left[Z_1^{1:f}, Z_2^{1:f}, \ldots, Z_N^{1:f}\right] \tag{19}$$

For each transition between shots i and $i+1$, we reset the diffusion process with:

$$z_{i+1}^0 = \epsilon_{\text{boundary}} + \alpha \cdot z_i^f \tag{20}$$

where $\alpha \in [0, 1]$ controls the temporal continuity and z_i^f is the final latent frame of the shot i. The complete video generation becomes:

$$V = \mathcal{D}\left(\bigcup_{i=1}^{N} Z_i\right) \in \mathbb{R}^{T \times 3 \times H \times W} \tag{21}$$

Figure 4: **Visual comparison of *VGoT* and baselines**

Table 1: **Quantitative comparison with state-of-the-art T2V baselines.** We compare average CLIP scores, and the average FC and SC scores within and across shots between VGoT and baseline models. We use **bold** to highlight the highest and *underline* for the second high.

Model	CLIP ↑	WS-FC ↑	CS-FC ↑	WS-SC ↑	CS-SC ↑
EasyAnimate Xu et al. (2024)	0.2402	0.4705	0.0268	0.7969	0.2037
CogVideo Yang et al. (2024)	0.2477	<u>0.6099</u>	0.0222	0.7424	<u>0.2069</u>
VideoCrafter1 Chen et al. (2023)	0.2478	0.3706	0.0350	0.7623	0.1867
VideoCrafter2 Chen et al. (2024)	<u>0.2529</u>	0.5569	<u>0.0686</u>	<u>0.7981</u>	0.1798
VGoT	**0.2557**	**0.8138**	**0.2688**	**0.9717**	**0.4276**

This mechanism decodes $\{Z_i\}_{i=1}^{N}$ into a coherent video $V \in \mathbb{R}^{T\times 3\times H\times W}$ through boundary reset operations, preserving both narrative flow and visual continuity across $T = N \times f$ frames without requiring additional training.

5 Experiments

5.1 Experiment Settings

Current video datasets lack sufficient multi-shot narratives with consistent characters across scenes. We therefore constructed a benchmark dataset using *VGoT* to create ten 30-shot stories (300 shots in total) for evaluation. Each story originates from a single user input S, which generates shot outlines S' and detailed scripts $\mathcal{P}$ across 30 shots and five domains defined in Eq. 6. For quantitative assessment, we employ four key metrics: Within-Shot Face Consistency ($\Omega_{\text{WS-FC}}$), Cross-Shot Face Consistency ($\Omega_{\text{CS-FC}}$), Within-Shot Style Consistency ($\Omega_{\text{WS-SC}}$), and Cross-Shot Style Consistency ($\Omega_{\text{CS-SC}}$) defined in Appendix C. We additionally report CLIP score Radford et al. (2021), PSNR Fardo et al. (2016), and IS Barratt and Sharma (2018). Our implementation uses multiple GPT-4o Achiam et al. (2023) for scripting, Kolor Team (2024) as base model for keyframes, and DynamiCrafter Xing et al. (2025) for video generation, compared against EasyAnimate Xu et al. (2024), CogVideo Hong et al. (2022), and VideoCrafter Chen et al. (2023, 2024) on NVIDIA H100 GPUs.

5.2 Comparison Evaluation

Our evaluation compares *VGoT* against state-of-the-art text-to-video models using narrative scenarios from Sec 5.1. Quantitative results in Table 1 demonstrate *VGoT*'s superiority as follows.

VGoT achieves **0.8138** $\Omega_{\text{WS-FC}}$ and **0.9717** $\Omega_{\text{WS-SC}}$, outperforming the best baselines (VideoCrafter2's 0.5569 $\Omega_{\text{WS-FC}}$ and 0.7981 $\Omega_{\text{WS-SC}}$) by **46.1%** and **21.7%** respectively. For cross-shot consistency,

Table 2: **Human Evaluation.** We compare *VGoT* with baseline models in terms of Within-Shot Consistency, Cross-Shot Consistency, and Visual Quality.

	Within-Shot Consistency			Cross-Shot Consistency			Visual Quality		
	Bad ↓	Normal ∼	Good ↑	Bad ↓	Normal ∼	Good ↑	Bad ↓	Normal ∼	Good ↑
EasyAnimate Xu et al. (2024)	0.3333	0.3232	0.3434	0.3535	0.3535	0.3131	0.4646	0.2727	0.2828
CogVideo Yang et al. (2024)	0.1341	0.4146	0.4512	0.2927	0.5976	0.2317	0.1463	0.4512	0.5244
VideoCrafter1 Chen et al. (2023)	0.5446	0.2574	0.1980	0.6436	0.1881	0.1683	0.6535	0.1782	0.1683
VideoCrafter2 Chen et al. (2024)	0.1262	0.4854	0.3883	0.3495	0.3786	0.2718	0.1748	0.4951	0.3981
VGoT	0.0889	0.2556	0.6556	0.0889	0.2444	0.6667	0.0889	0.2111	0.7000

Table 3: **Ablation Studies.** We evaluate the impact of removing key modules from our proposed framework. Metrics include CLIP Score, PSNR, IS, FC score, and SC score

	CLIP average ↑	PSNR ↑	IS ↑	WS-FC ↑	CS-FC ↑	WS-SC ↑	CS-SC ↑
w/o DSM w/o IPP	0.1146	24.3265	7.4624	0.7364	0.1129	0.9406	0.3650
w DSM w/o IPP	0.1146	24.3265	7.5783	0.7305	0.1174	0.9471	0.3663
w/o DSM w IPP	**0.1223**	23.9228	7.4521	**0.8745**	**0.3291**	0.9486	**0.4186**
Full Method	0.1111	**25.7857**	**7.5194**	0.8303	0.2738	**0.9487**	0.3859

*FC is denoted as Face Consistency, and SC is denoted as Style Consistency

VGoT's **0.2688** $\Omega_{\text{CS-FC}}$ and **0.4276** $\Omega_{\text{CS-SC}}$ surpass VideoCrafter1's second-best 0.0686 $\Omega_{\text{CS-FC}}$ and CogVideo's 0.2069 $\Omega_{\text{CS-SC}}$. While outperform text-visual alignment via best CLIP score (**0.2557**), *VGoT* maintains this performance while requiring **10**× **less** manual input, more qualitative analysis

Human evaluations (Table 2) confirm these findings: VGoT receives **66.7%** "Good" ratings for cross-shot consistency versus 27.2% for VideoCrafter2 and 23.2% for CogVideo. In visual quality, 70.0% of evaluators rate VGoT's outputs as "Good" compared to 52.4% for CogVideo. This preference is qualitatively validated in Fig 4, which demonstrates VGoT's superior maintenance of character consistency and visual coherence across extended narratives compared to baseline outputs.

5.3 Ablation Studies

We analyze two core components through systematic removal: (1) Dynamic Storyline Modeling (DSM) and (2) Identity-Preserving Portraits (IPP). Using the 30-shot cycling narrative, Table 3 reveals three critical patterns:

1. **CLIP-Logic Tradeoff**: The full model achieves lowest CLIP score (0.1111 vs 0.1223 baseline) but highest PSNR (25.79) and IS (7.52), confirming that DSM's narrative enrichment and IPP's consistency control prioritize cinematic quality over literal prompt matching.

2. **Consistency Costs**: Removing DSM boosts $\Omega_{\text{CS-FC}}$ by 20.2% (0.3291 vs 0.2738) and SC by 8.5% (0.4186 vs 0.3859), but as Fig 5 shows, this comes at severe narrative diversity loss—identical camera angles and repetitive scenes dominate DSM-ablated outputs.

3. **Component Synergy**: IPP alone achieves 0.8745 $\Omega_{\text{WS-FC}}$ (5.3% higher than full model), but combined DSM+IPP delivers optimal balance—8.3% better $\Omega_{\text{CS-SC}}$ than IPP-only versions while maintaining visual quality (25.79 PSNR vs 23.92).

These results prove both components are essential: DSM enables story progression while IPP ensures continuity, together resolving the consistency-diversity paradox in multi-shot generation.

6 Conclusion

We present VideoGen-of-Thought (*VGoT*), a structured framework that automates multi-shot video generation with *reasonable storylines* and *visual consistency* **from one sentence** through **minimal manual intervention**, through three core innovations: dynamic storyline modeling, identity-aware cross-shot propagation, and adjacent latent transition mechanisms. *VGoT* achieves 20.39% higher within-shot face consistency and 17.36% better style consistency than previous state-of-the-art methods, with 10× fewer manual adjustments than alternatives while maintaining director-level narrative flow. Our work redefines automated multi-shot generation, bridging raw visual synthesis with cinematic storytelling through systematic visual stories decomposition.

References

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

Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, and Ying Fu. Partial fc: Training 10 million identities on a single machine. In *ICCVW*, 2021.

Xiang An, Jiangkang Deng, Jia Guo, Ziyong Feng, Xuhan Zhu, Yang Jing, and Liu Tongliang. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *CVPR*, 2022.

Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv: 1801.01973*, 2018.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv: 2311.15127*, 2023a.

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

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *IJCV*, 2018.

Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the IEEE Conference on European Conference on Computer Vision*, 2020a.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020b.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023.

Fernando A. Fardo, Victor H. Conforto, Francisco C. de Oliveira, and Paulo S. Rodrigues. A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms. *arXiv preprint arXiv: 1605.07116*, 2016.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022.

Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Ostec: One-shot texture completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Jia Guo, Jiankang Deng, Niannan Xue, and Stefanos Zafeiriou. Stacked dense u-nets with dual transformers for robust face alignment. In *BMVC*, 2018.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022.

Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*, 2023.

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

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv: 2205.15868*, 2022.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. *European Conference on Computer Vision*, 2020.

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.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. In *NeurIPS*, 2024.

Diederik P. Kingma and M. Welling. Auto-encoding variational bayes. *International Conference On Learning Representations*, 2013.

Zongyi Li, Shujie Hu, Shujie Liu, Long Zhou, Jeongsoo Choi, Lingwei Meng, Xun Guo, Jinyu Li, Hefei Ling, and Furu Wei. Arlon: Boosting diffusion transformers with autoregressive models for long video generation. *arXiv preprint arXiv: 2410.20502*, 2024.

Haozhe Liu, Shikun Liu, Zijian Zhou, Mengmeng Xu, Yanping Xie, Xiao Han, Juan C. Pérez, Ding Liu, Kumara Kahatapitiya, Menglin Jia, Jui-Chieh Wu, Sen He, Tao Xiang, Jürgen Schmidhuber, and Juan-Manuel Pérez-Rúa. Mardini: Masked autoregressive diffusion for video generation at scale. *arXiv preprint arXiv: 2410.20280*, 2024.

Yichen Ouyang, Hao Zhao, Gaoang Wang, et al. Flexifilm: Long video generation with flexible conditions. *arXiv preprint arXiv:2404.18620*, 2024.

Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2920–2930, 2024.

Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.

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

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

Xingyu Ren, Alexandros Lattas, Baris Gecer, Jiankang Deng, Chao Ma, and Xiaokang Yang. Facial geometric detail recovery via implicit representation. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 2023.

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.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv: 2111.02114*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

K. Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference On Learning Representations*, 2014.

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.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2020a.

Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, S. Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference On Learning Representations*, 2020b.

Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024.

Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *arXiv preprint arXiv: 2406.09399*, 2024.

Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, Chong Luo, Yueyi Zhang, and Zhiwei Xiong. Art•v: Auto-regressive text-to-video generation with diffusion models. *arXiv preprint arXiv: 2311.18834*, 2023.

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.

Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawend F. Bissyand, and Saad Ezzini. Dreamfactory: Pioneering multi-scene long video generation with a multi-agent framework. *arXiv preprint arXiv: 2408.11788*, 2024.

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 *European Conference on Computer Vision*, pages 399–417. Springer, 2025.

Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv: 2405.18991*, 2024.

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.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.

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

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. *arXiv preprint arXiv:2411.17440*, 2024.

Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8850–8860, 2024.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, pages 1–15, 2024.

Canyu Zhao, Mingyu Liu, Wen Wang, Jianlong Yuan, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequence. *arXiv preprint arXiv:2407.16655*, 2024.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2025.

A Limitations, Licenses, and Future Work

Limitations. VGoT relies on pretrained components without additional finetuning, which bounds performance by the base models' capabilities. In particular, DynamiCrafter Xing et al. (2025) may limit motion diversity under highly complex camera trajectories, reduce temporal stability in out-of-distribution scenes with rapid appearance changes, and constrain very long-range dependencies beyond adjacent-shot transitions. These constraints are characteristic of training-free pipelines and motivate future model-level improvements.

Licenses Declaration. We acknowledge and comply with licenses of third-party assets used in our pipeline. DynamiCrafter is distributed under the Apache License 2.0. Kolor is distributed under the Apache-2.0 license. We use these tools within their permitted scopes and cite their sources. Other services used for scripting (e.g., commercial LLM APIs) are accessed under their respective terms of service. Our released code and evaluation scripts will clearly indicate all external dependencies and their licenses.

Future Work. We plan to: (1) integrate stronger video backbones and optional finetuning to enhance motion diversity and long-range temporal reasoning; (2) extend identity handling to multi-subject IPP with fine-grained attribute disentanglement; (3) broaden cultural and linguistic coverage in prompts and benchmarks; (4) include optional professional and sturctured movie screenplay writing in the script generation process.

B Detailed Example of Dynamic Storylines

Dynamic Storylines Modeling plays a fundamental role in converting high-level user input into a detailed series of prompts for each shot within the multi-shot video generation process. The specific process is to convert a single sentence user input S, into a more detailed and structured description S', which is then decomposed into a set of prompts $P = \{p_1, p_2, \ldots, p_N\}$, corresponding to each of the N shots required for the complete video. This process uniformly adopts a large language model (LLM) and uses prompt engineering to ensure the reasonableness of the generated video. This process enables the generation of a logical, stepwise narrative structure that serves as the backbone for subsequent keyframe and shot-level generation.

For example, consider the user input: *e.g., a set of one-sentence prompts, 30 shots, describe a story about three avatars (two friend Mike and Jane, one bad guy Tom) and their fantasy adventure in an ancient pyramid.*. The LLM would first transform this input into a detailed version S' consisting of 30 one-sentence scripts like s_1: *"Mike's Discovery: Mike examines an ancient map with intense focus, revealing the path to the pyramid."* Subsequently, each s_i in S' is decomposed into prompts P, such as p_1: ***Character**: Mike, holding an ancient map with Jane by his side.* ***Background**: A dense jungle filled with mist and towering trees.* ***Relation**: Mike studies the map closely, pointing to a pyramid.* ***Camera Pose**: Medium shot focusing on Mike and Jane.* ***HDR Description**: Soft light filters through the trees, creating dynamic shadows on the map and characters."*. This structured approach ensures narrative coherence across multiple shots, laying the foundation for the generation process.

Our self-validation mechanism ensures cinematic rigor through iterative refinement. Consider a draft shot description: *2. Inside the helicopter, a diverse team of scientists, their faces filled with anticipation and anxiety, pore over maps and equipment*

Initial Generation Attempt:

$$p_2^{(1)} = \begin{cases} p_{\text{cha}} & \text{Team of scientists with mixed ages} \\ p_b & \text{Helicopter interior} \\ p_r & \text{Studying maps} \\ p_{\text{cam}} & \text{Close-up} \\ p_h & \text{(Missing HDR specification)} \end{cases}$$

Validation fails with $K(p_2^{(1)}, \mathcal{C}_{\text{film}}) = 0$ due to incomplete HDR description. The LLM regenerates with lighting constraints:

Validated Output:

$$p_2^* = \begin{cases} p_{\text{cha}} & \text{Diverse team in expedition gear, anxious expressions} \\ p_b & \text{Helicopter over Amazon rainforest} \\ p_r & \text{Team engrossed in equipment} \\ p_{\text{cam}} & \text{Wide shot showing interior/exterior} \\ p_h & \text{Sunlight streaming through windows, warm glow} \end{cases}$$

This satisfies both criteria: $C(p_2^*, p_1) = 0.92 > \tau_c$, $K(p_2^*, \mathcal{C}_{\text{film}}) = 1$, ensuring director-level coherence from user input under minimal manual intervention.

C Multi-Shot Evaluation Protocol

Existing evaluation frameworks predominantly focus on single-shot quality, leaving multi-shot assessment underspecified. We introduce a hierarchical protocol that quantifies both intra-shot and inter-shot consistency:

$$\Omega_{\text{multi-shot}} = \{\Omega_{\text{WS-FC}}, \Omega_{\text{CS-FC}}, \Omega_{\text{WS-SC}}, \Omega_{\text{CS-SC}}\} \tag{22}$$

Within-Shot Face Consistency (WS-FC) measures identity preservation within temporal sequences:

$$\Omega_{\text{WS-FC}}(V_i) = \frac{1}{f-1} \sum_{j=1}^{f-1} \cos\langle F_j^i, F_{j+1}^i \rangle \tag{23}$$

where F_j^i represents facial features from frame j of shot i extracted using InsightFace Ren et al. (2023); Guo et al. (2021); Gecer et al. (2021); An et al. (2022, 2021); Deng et al. (2020a,b); Guo et al. (2018); Deng et al. (2018, 2019).

Cross-Shot Face Consistency (CS-FC) evaluates identity preservation across shots:

$$\Omega_{\text{CS-FC}}(V) = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{1}{n} \sum_{j=1}^{n} \cos\langle F_{f-j+1}^i, F_j^{i+1} \rangle \tag{24}$$

where $n = 8$ in our implementation, comparing the last n frames of shot i with the first n frames of shot $i + 1$.

Within-Shot Style Consistency (WS-SC) quantifies stylistic coherence through VGG-19 Simonyan and Zisserman (2014) features:

$$\Omega_{\text{WS-SC}}(V_i) = \frac{1}{f-1} \sum_{j=1}^{f-1} \cos\langle S_j^i, S_{j+1}^i \rangle \tag{25}$$

where S_j^i represents flattened VGG-19 Simonyan and Zisserman (2014) features from frame j of shot i.

Cross-Shot Style Consistency (CS-SC) assesses style preservation between adjacent shots:

$$\Omega_{\text{CS-SC}}(V) = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{1}{n} \sum_{j=1}^{n} \cos\langle S_{f-j+1}^i, S_j^{i+1} \rangle \tag{26}$$

These metrics establish a comprehensive framework capturing both local and global consistency aspects essential for multi-shot video assessment, addressing limitations in current evaluation approaches.

D Additional Results

We validate *VideoGen-of-Thought (VGoT)* through four narrative archetypes. **Type 1: Longitudinal Character Development** demonstrates decade-spanning consistency using prompts like *"30-shot*

Figure 5: Visual Demonstration of the ablation studies of **VGoT**

story of Marco discovering ancient secrets through culinary journeys", where our framework maintains consistency across aging sequences. **Type 2: Multi-Actor Scenes** handles complex group dynamics in scenarios like *"Abandoned factory transformed into community art center"*, preserving relationship continuity between complex stories and multiple characters across shots through identity-aware propagation. **Type 3: Non-Human Narratives** extends identity preservation to fantastical subjects, as shown in *"Mycelium networks and mechanical bees restoring ecosystems"*, we explore the creatity of *VGoT* in marvelous entities. We also evaluate the capability of *VGoT* to create diverse stories with the same input: *"an immigrant's story of moving to a new country, struggling, and eventually finding success as an entrepreneur."* in **Type 4** showcases.

Moreover, we also provide additional comparison results to illustrate *VGoT*'s advantages over existing state-of-the-art methods. These comparisons include visual comparisons with four baselines: *EasyAnimate*, *CogVideo*, *VideoCrafter1*, and *VideoCrafter2*. Each comparative example is analyzed in terms of visual consistency, narrative coherence, and overall quality. As shown in Figure 7, *VGoT* consistently outperforms the baselines in terms of character continuity, background stability, and logical flow across shots. These results highlight *VGoT*'s ability to maintain coherent storytelling while also achieving high-quality visuals. We also prepared the original experiment data record for quantitative evaluation and ablation studies in our provided materials.

E User Study

To evaluate the user-perceived quality of videos generated by our *VGoT* framework, we conducted an extensive user study involving 10 participants. The participants were given 50 accelerated multi-shot videos, each generated either by *VGoT* or one of four baseline methods. The 10 input stories, consisting of 30 shots each, were randomly assigned to ensure diverse feedback and minimize bias. Each user was presented with 10 videos from different sources and asked to evaluate them on a scale of *good*, *normal*, or *bad*, based on three specific criteria: within-shot consistency, cross-shot consistency, and visual quality.

The results of the user study are summarized in Figure 8. The data indicates that users significantly preferred the videos generated by *VGoT*, especially regarding cross-shot consistency. Users found that *VGoT*'s videos maintained logical transitions between shots and preserved character appearances

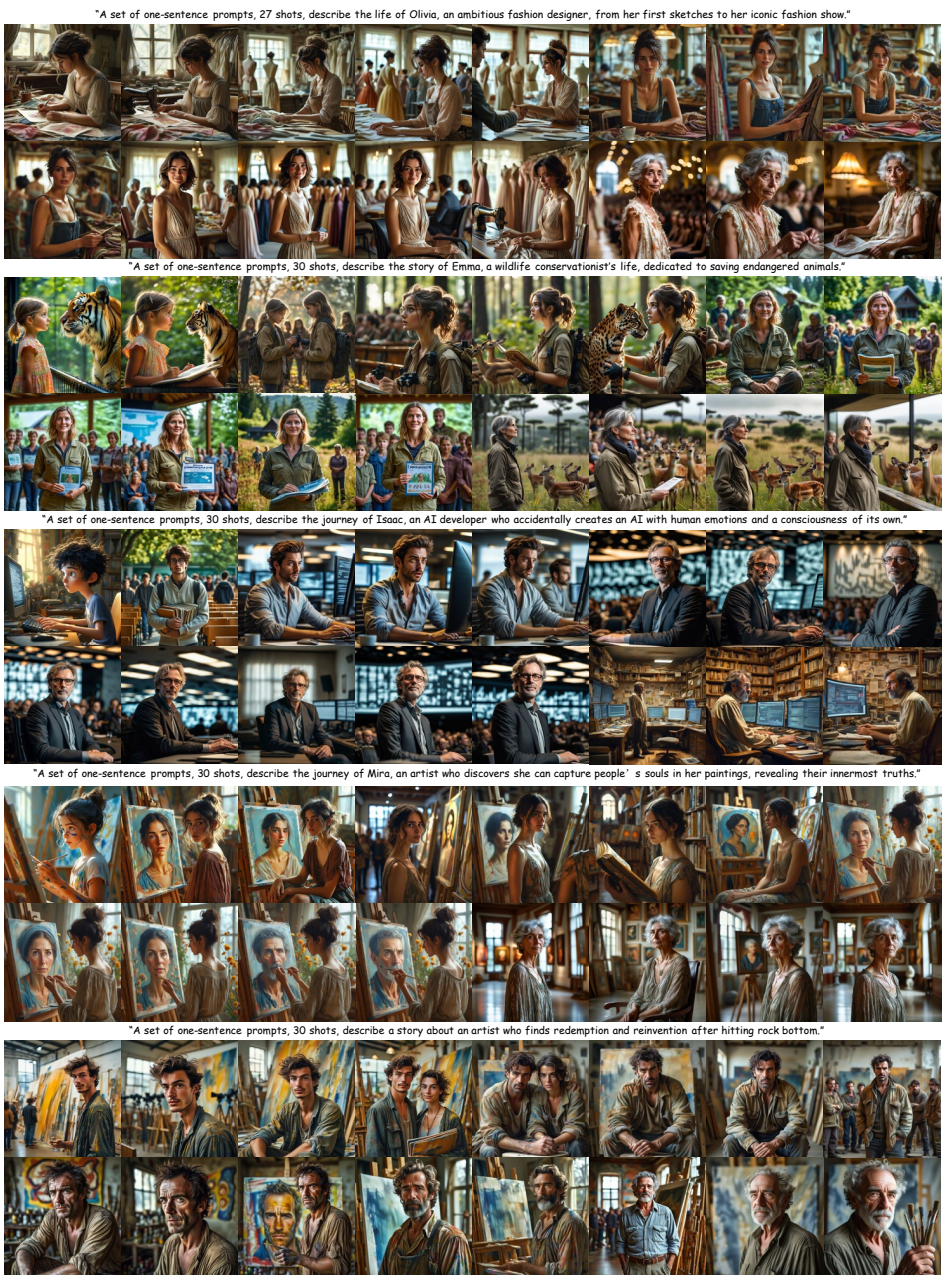

Figure 6: *VGoT* Visual complement of the multi-camera video generated.

across different scenes, reflecting the robustness of our approach. Compared to the baselines, *VGoT*'s results were rated highly for narrative coherence and overall quality, demonstrating the effectiveness of our collaborative multi-shot framework in meeting user preferences.

F Ethics Statement

Potential Harms Caused by the Research Process. Our study uses publicly available pretrained systems for scripting and generation (e.g., GPT-4o for script preparation, Kolor for keyframes, and DynamiCrafter for video synthesis) in accordance with their licenses and terms of service (see Section 5.1). Computation for metric evaluation and visual synthesis was performed on NVIDIA H100 GPUs. The 10-story benchmark used for evaluation is created by the authors specifically for

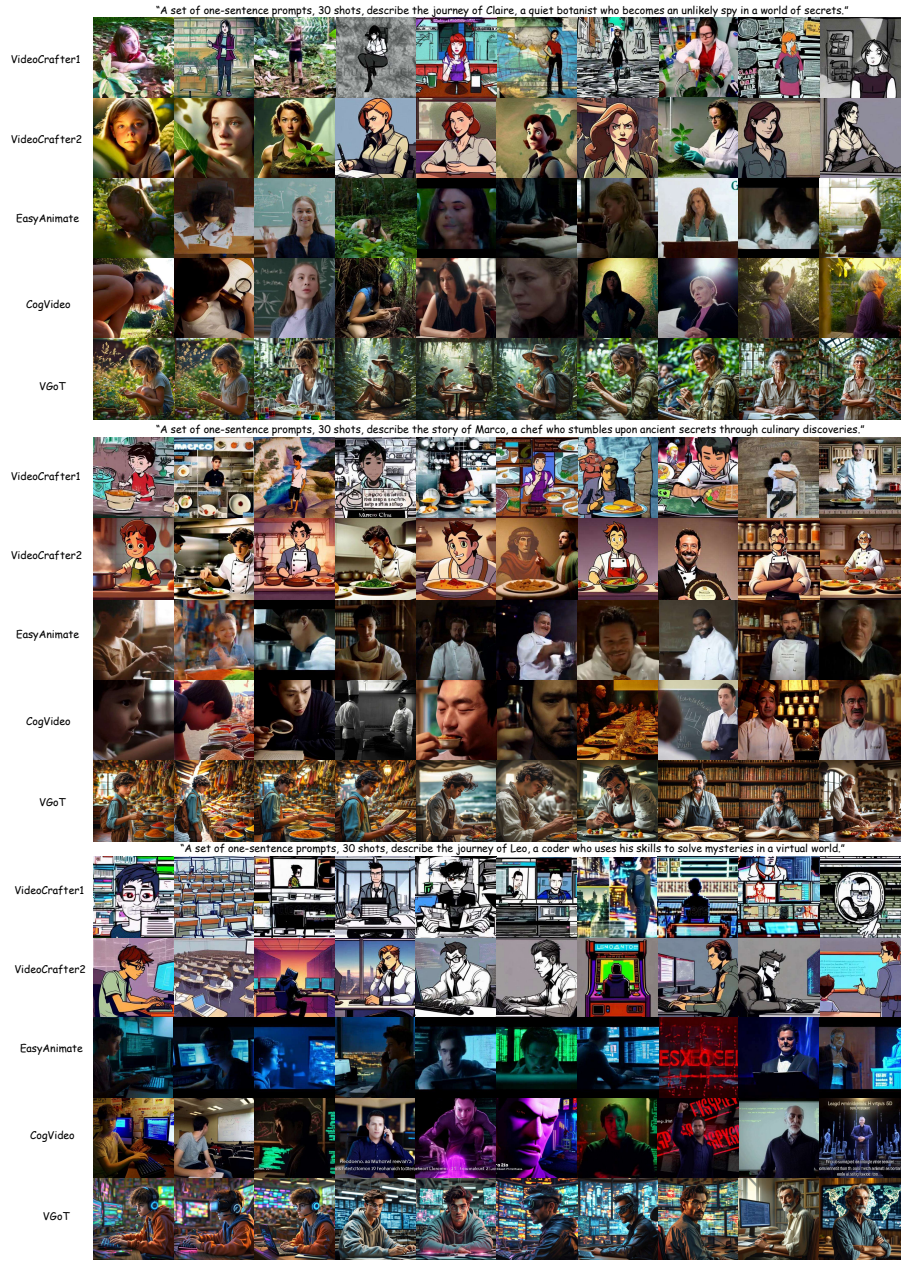

Figure 7: Visual comparison of *VGoT* with baselines Supplement.

this work; it does not contain personal data or copyrighted media beyond model outputs generated under the respective tools' usage policies. A small human evaluation with 10 participants was conducted to assess perceived quality and consistency (Figure 8); participants were informed of the study purpose, their privacy was protected, and compensation followed local norms. No sensitive personal information was collected, and we identified no additional risks to participants.

Societal Impact and Potential Harmful Consequences. VGoT is a training-free pipeline that automates multi-shot video generation from a single sentence. While this can benefit creative workflows and prototyping, risks remain. First, the environmental footprint of generative pipelines is

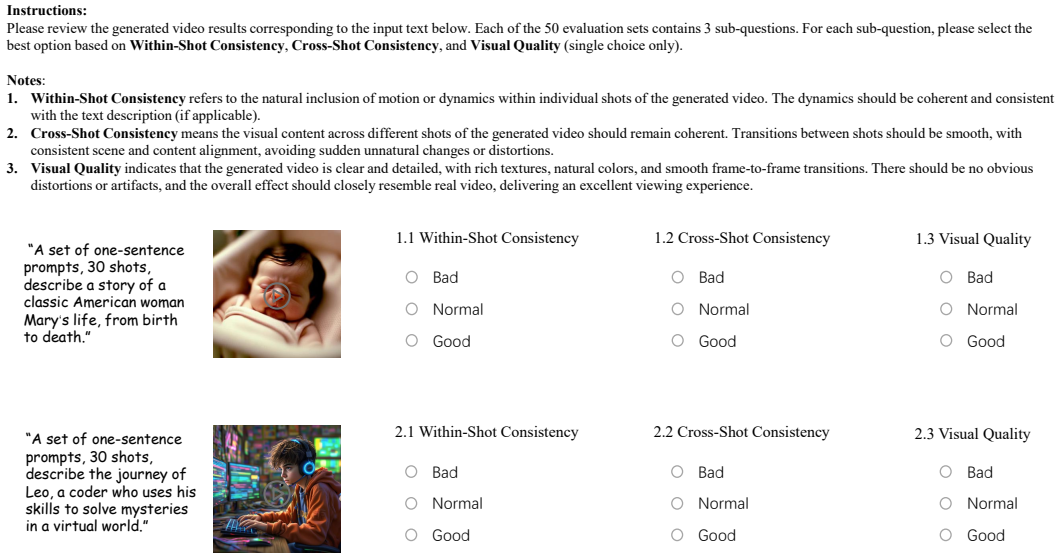

Figure 8: Designed user study interface. Each participant was required to rate 50 videos by answering three sub-questions for each video. Due to page limitations, only two videos are shown here.

non-negligible; H100-class accelerators consume substantial energy during inference and evaluation. Second, synthetic videos may be misused for disinformation or to fabricate misleading content if deployed irresponsibly. Third, bias can arise from scriptwriting and scene conventions (e.g., English-centric prompts or specific cultural settings), potentially reducing representativeness across regions or languages. Future work should prioritize energy-aware configurations, content provenance indicators, and broader cultural coverage in story prompts and evaluation scenarios.

Impact Mitigation Measures. We intend to release the VGoT code and evaluation scripts under an open-source license for academic research use, documented to clarify intended use and discourage misuse. We recommend that downstream deployments include visible AI-generation disclosures and optional watermarking, follow model and dataset licenses, and avoid use cases that could harm individuals or communities. We will maintain the released materials and welcome community feedback to improve responsible usage and coverage.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction state the three core contributions and reported gains, which match the methods in Section 4 and our experimental evidence in Sections 5.1 and C.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix A details our limitations tied to base model capacity and discusses future directions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We present a systematic problem–solution framework for multi-shot video generation in Section 4 and empirical evaluation protocol in Appendix C; we do not include formal theorems or proofs that require explicit assumption sets.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Core settings (10 stories × 30 shots), baselines, models, hardware, and metrics are specified in Sections 5.1 and C; we also state intent to release code and evaluation scripts to facilitate exact reproduction (Appendix F).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We intend to release the VGoT code and evaluation scripts for academic research with documentation and licensing guidance (Appendix F).

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify story construction (10 stories, 30 shots each), baselines, models (GPT-4o, Kolor, DynamiCrafter), hardware (H100), and metrics; our pipeline is training-free, so training hyperparameters are not applicable (Section 5.1).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Despite we conduct quantitative comparison in Table 1 and qualitative comparison in Figure 4, human evaluation in Table 2, and ablation study in Table 3 and Figure 5, we don't report point estimates without error bars, confidence intervals, or statistical tests.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We disclose compute resources (H100 GPUs) and a training-free inference pipeline; experiments are reproducible under the specified story set and metrics, with compute bounded by inference-time usage (Section 5.1).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We use synthetic prompts and public pretrained models and do not process personal data; we adhere to standard research ethics and anonymization for submission.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Our Ethics Statement (Appendix F) discusses potential positive uses and negative societal impacts (energy footprint, misuse risks, and cultural bias) and outlines mitigation measures.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release high-risk models or scraped datasets; our paper does not introduce assets that require special safeguards.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Appendix A declares licenses for DynamiCrafter (Apache License 2.0) and Kolor (Apache-2.0) and notes compliance with external services' terms.

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will introduce new pipeline code and evaluation scripts for VGoT and will provide documentation and license notes upon release (See Appendix F).

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Our Ethics Statement summarizes participant privacy protection and compensation norms; Figure 8 illustrates the interface. We provide the instruction in the supplemental material.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our human-subject study involved rating multi-shot video quality and consistency, which is generally considered minimal risk. Participants were informed about the evaluation task; as detailed in Appendix F, we protected privacy and identified no additional risks. We do not provide formal IRB approval details, which can be common for minimal-risk studies depending on institutional policies, but we followed ethical considerations regarding compensation and privacy.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLM usage is a core component of our method and is described in Section 4 (dynamic storyline modeling) and Section 5.1.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.