HoloCine: Holistic Generation of Cinematic Multi-Shot Long Video Narratives

Yihao Meng[1,2]   Hao Ouyang[2]   Yue Yu[1,2]   Qiuyu Wang[2]   Wen Wang[2,3]

Ka Leong Cheng[2]   Hanlin Wang[1,2]   Yixuan Li[2,4]   Cheng Chen[2,5]   Yanhong Zeng[2]

Yujun Shen[2]   Huamin Qu[1]

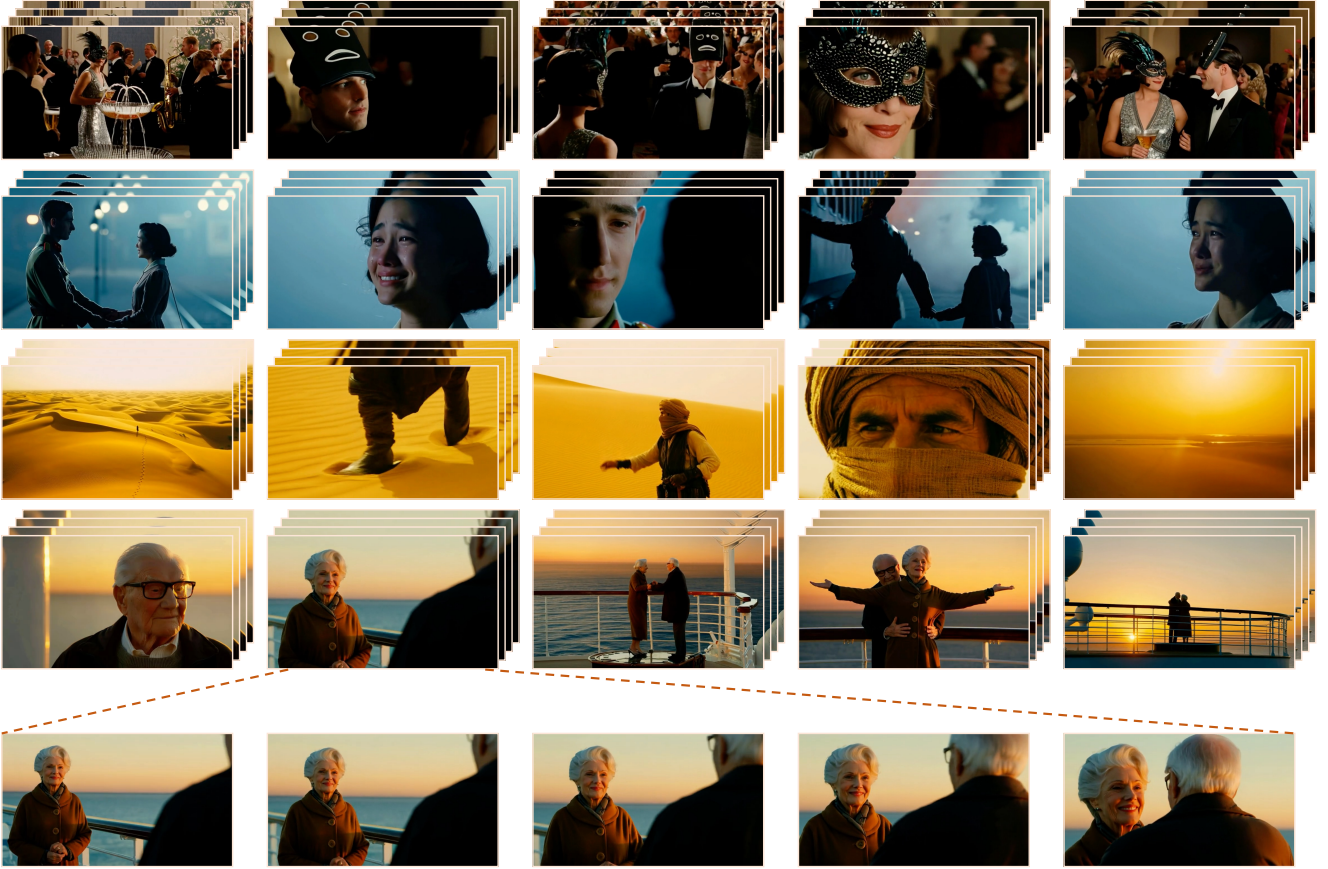[1] HKUST   [2] Ant Group   [3] ZJU   [4] CUHK   [5] NTU

Figure 1. From a text prompt alone, `HoloCine` generates coherent cinematic multi-shot video narratives in a holistic pass. The figure showcases our model's versatility, featuring diverse original scenes (top three rows) and a cinematic homage to Titanic (bottom rows). All scenes exhibit exceptional character consistency and narrative coherence. The expanded final row demonstrates smooth intra-shot motion and quality. Our code is available at: https://holo-cine.github.io/.

Abstract

State-of-the-art text-to-video models excel at generating isolated clips but fall short of creating the coherent, multi-shot narratives, which are the essence of storytelling. We bridge this "narrative gap" with `HoloCine`, a model that generates entire scenes holistically to ensure global consistency from the first shot to the last. Our architecture achieves precise directorial control through a Window Cross-Attention mechanism that localizes text prompts to specific shots, while a Sparse Inter-Shot Self-Attention pattern (dense within shots but sparse between them) ensures

the efficiency required for minute-scale generation. Beyond setting a new state-of-the-art in narrative coherence, `HoloCine` develops remarkable emergent abilities: a persistent memory for characters and scenes, and an intuitive grasp of cinematic techniques. Our work marks a pivotal shift from clip synthesis towards automated filmmaking, making end-to-end cinematic creation a tangible future.

1. Introduction

The field of generative AI has witnessed extraordinary progress, with text-to-video (T2V) synthesis emerging as a prominent frontier. Driven by the scaling of Diffusion Models [17] and Diffusion Transformers (DiTs) [32], state-of-the-art models [4, 27, 28, 41, 50] can now generate high-fidelity, single-shot video clips from textual prompts. Yet, this capability falls short of emulating the structure of most visual media. Films, television series, and documentaries are not single, unbroken takes. They are narratives constructed from sequences of distinct shots edited together to tell a cohesive story. This disconnect between current generative capabilities and the language of cinema represents the next major challenge: bridging the *"narrative gap"* by moving from single-clip generation to multi-shot, scene-level synthesis.

Current approaches to generating longer multi-shot videos often rely on a decoupled generation paradigm. Whether generating a video chunk-by-chunk [7, 9, 16, 34, 42, 52, 54], or first creating keyframes and then independently synthesizing the connecting shots [22, 47, 56–58], these methods model different parts of the video through separate processes. Even when conditioned on character or scene information to improve consistency [21, 23, 25, 30], the generation of individual shots remains largely independent. This fundamental decoupling inherently limits long-range coherence, leading to prevalent issues like error accumulation and consistency drift, where visual attributes such as character identity and background details degrade over time.

A more promising, emerging direction is the holistic pipeline, recently exemplified by LCT [15], where the entire multi-shot sequence is modeled jointly. This approach is powerful for maintaining global consistency but introduces two formidable challenges. First, achieving precise control is difficult, as per-shot instructions can be "diluted" within the context of the entire prompt. Second, the prohibitive computational cost of the self-attention mechanism, which scales quadratically with sequence length, makes generating longer, minute-scale videos practically intractable.

In this paper, we introduce `HoloCine`, a novel framework that unlocks the potential of holistic generation through two specialized architectural designs. For precise directorial control, our Window Cross-Attention mecha-nism localizes attention, directly aligning per-shot text prompts with their corresponding video segments to enable sharp, narrative-driven transitions. To overcome the computational bottleneck, our Sparse Inter-Shot Self-Attention leverages a hybrid pattern: it maintains dense attention within shots for motion continuity while using sparse connections based on compact summaries for efficient communication between shots. This design reduces computational complexity to a near-linear relationship with the number of shots, making minute-scale holistic generation feasible. Finally, to enable the training of our framework, we developed a robust data curation pipeline to build a large-scale, hierarchically annotated dataset of multi-shot scenes.

Extensive experiments validate the effectiveness of our proposed framework. `HoloCine` significantly outperforms strong baselines across major existing paradigms—including powerful pre-trained models [41], two-stage keyframe-to-video pipelines [22, 58], and other holistic approaches [46]. Our method establishes a new state-of-the-art in long-term consistency, narrative fidelity, and precise shot transition control. Ablation studies further confirm the critical roles of our novel components: Window Cross-Attention is essential for achieving fine-grained directorial control, while Sparse Inter-Shot Self-Attention is vital for scalability, delivering quality comparable to full attention at a fraction of the computational cost. Finally, our analysis reveals that `HoloCine` exhibits remarkable emergent capabilities. These include a persistent memory for characters and scene details across multiple shots and nuanced control over cinematic language, suggesting the model has developed a deeper, implicit understanding of visual storytelling. By enabling minute-scale holistic generation, our work shifts the paradigm from isolated clips to directing entire cinematic scenes, paving the way for automated, end-to-end filmmaking.

2. Related Work

2.1. Single-Shot Video Generation

The foundation of our work lies in the rapid progress of single-shot text-to-video (T2V) generation. Early models extended GAN architectures to the video domain [1, 10, 39, 44], but the advent of diffusion models marked a paradigm shift in quality and coherence [3, 8, 27, 41, 49, 50]. Foundational models like Imagen Video [18], Make-A-Video [38], and VDM [19] demonstrated the potential of cascaded diffusion and 3D U-Net architectures to generate short, high-fidelity clips. The introduction of the Diffusion Transformer (DiT) architecture [32] further improved scalability and generation quality, operating on latent patches and replacing the U-Net's inductive bias with the powerful self-attention

mechanism. Models like *Kling* [28] and other open-source efforts [27, 41, 50] have shown that scaling DiT-based architectures leads to remarkable capabilities in generating 5-second-long high-resolution single shot videos. However, these models are fundamentally designed for single shot videos and lack explicit mechanisms to construct a coherent narrative across multiple, distinct shots—the very essence of cinematic storytelling.

2.2. Multi-Shot and Scene-Level Video Generation

Bridging the gap from single shots to coherent scenes is an active area of research. One major approach is the hierarchical pipeline [2, 20, 30, 43, 48, 56], where an LLM first decomposes a story into sequential prompts, and a T2V model then generates each shot independently. To mitigate inconsistencies, works like VideoStudio [30] and MovieDreamer [56] add constraints using embeddings or visual tokens, but the isolated nature of shot generation remains a fundamental bottleneck. Another line of work relies on keyframe-based generation, such as Story Diffusion [35], IC-LoRA [22], and Captain Cinema [47]. These methods first create a sequence of consistent keyframes and then generate the video segments between them. Here, consistency is primarily enforced at the keyframe level, while the video infilling for each shot is still performed in isolation. Recognizing these limitations, a more promising paradigm has recently emerged: holistic generation [15, 26, 46]. LCT [15] proposes interleaved positional embeddings within MMDiT [14] architectures to jointly model all shots in a single, unified diffusion process. This approach inherently enforces global consistency, representing a significant conceptual advance. Aligning with this emerging paradigm, our framework, HoloCine, generates multi-shot videos holistically for inherent consistency. It further introduces two specialized mechanisms, Window Cross-Attention and Sparse Inter-Shot Self-Attention, to simultaneously provide precise directorial control and computational efficiency for practical scene-level synthesis.

2.3. Long Video Generation

Generating long videos, whether single-shot or multi-shot, inevitably confronts the quadratic complexity of the Transformer's self-attention mechanism. The predominant strategy to circumvent this is autoregressive generation, where a video is synthesized in sequential, overlapping chunks [7, 9, 16, 34, 42, 52, 54]. To mitigate error accumulation problem in this paradigm, one strategy enhances robustness by training models to denoise controlled noise injected into the historical context (Diffusion Forcing [7]). Some methods attempt to distill the entire past into a fixed-size latent vector. For example, TTTVideo [12] encodes context via an MLP at inference time, while FramePack [54] compresses multiple frames into a single vector to predict the next frame. While computationally manageable, this paradigm is prone to consistency drift and error accumulation, where visual fidelity degrades as the sequence lengthens.

To address the computational bottleneck more directly, another line of research explores efficient Transformer architectures. STA [55] utilizes localized 3D windows, processing video tile-by-tile in a manner compatible with FlashAttention [13]. SageAttention [53] combine selective token compression with a softmax-aware pass to prune key steps of the attention calculation. Radial Attention [29] employs a static $O(n \log n)$ mask derived from spatio-temporal energy decay, which enables longer video generation with quality that closely resembles dense attention. Most recent work such as Mixture of Contexts [5] has applied similar ideas to long-context video generation by partitioning tokens into chunks and using a trainable router to select relevant context. Our Sparse Inter-Shot Self-Attention is inspired by this line of work but is specifically tailored to the unique structure of multi-shot video. We hypothesize that the information required for consistency between shots is different from that required within a shot, allowing for a principled, structured sparsity pattern that is both efficient and effective for narrative synthesis.

3. Method

Our goal is to generate a coherent, multi-shot video sequence from a hierarchical text prompt in a single, holistic pass. To achieve this, we introduce HoloCine, a framework built upon the powerful DiT-based video diffusion model, Wan2.2 [41]. In the following sections, we detail our data curation and hierarchical annotation pipeline (Sec. 3.1), the Window Cross-Attention mechanism for explicit shot boundary control (Sec. 3.2.1), and the Sparse Inter-Shot Self-Attention mechanism that makes holistic generation computationally efficient (Sec. 3.2.2).

3.1. Data Curation and Annotation

One primary obstacle for multi-shot video generation is the lack of large-scale, high-quality datasets. Public video datasets are typically composed of isolated, short video clips. To address this, we developed a comprehensive data curation pipeline to process cinematic films and television series into a structured, multi-shot dataset.

Shot Segmentation and Filtering. Our pipeline begins by collecting a large corpus of cinematic content from public sources. We then employ a shot boundary detection algorithm [40] to segment each video into individual shots, recording their start and end timestamps. These clips then undergo a rigorous filtering process, where we remove subtitles using [51] and discard clips that are too short, overly dark, or have low aesthetic scores.

Multi-Shot Sample Assembly: To construct coherent multi-shot samples, we sequentially group temporally con-
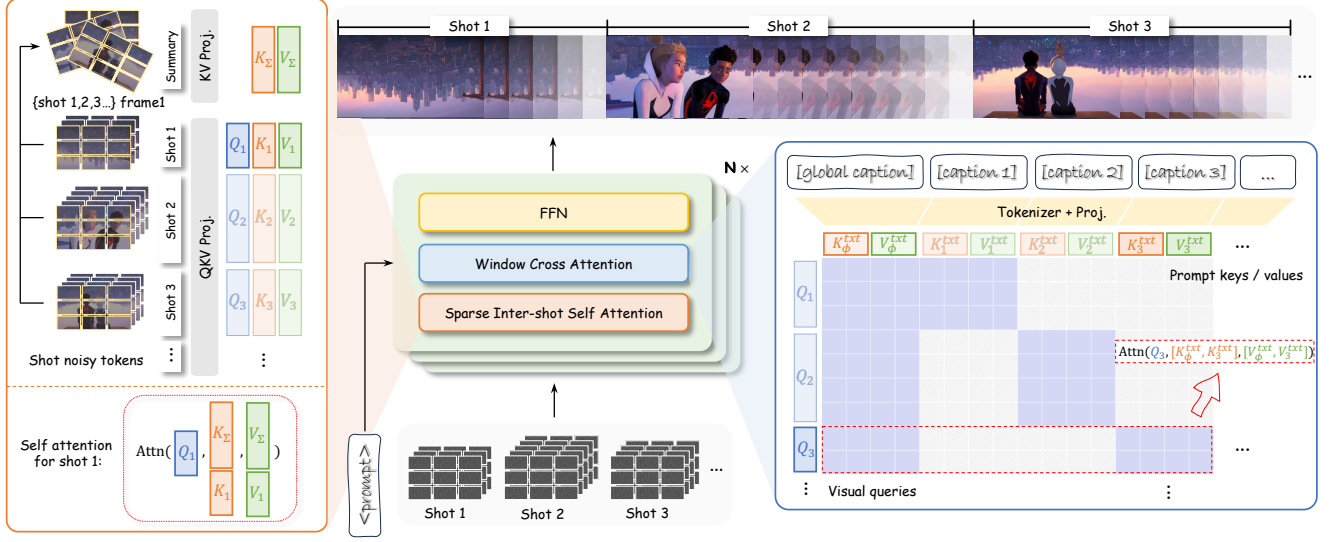
Figure 2. The architecture of our holistic generation pipeline, where all shot latents are processed jointly. The Window Cross-Attention provides precise directorial control by aligning each shot to its specific text prompt. The Sparse Inter-shot Self-Attention drastically reduces computational cost while preserving long-range consistency.

tiguous shots from the source video to form training samples. This grouping is guided by a target total duration (e.g., 5, 15, or 60 seconds), with shots being aggregated until the threshold is met within a certain tolerance. This process generates a diverse set of samples with varying numbers of shots, creating uniform batches for efficient training. The final dataset contains 400k samples with a controllable distribution of shots across these duration tiers.

Hierarchical Captioning: Each multi-shot sample is annotated with a hierarchical prompt structure using Gemini 2.5 Flash [11]. A global prompt describes the overarching scene, including the characters, environment, and plot. Following this, a series of per-shot prompts detail the specific actions, camera movements, and characters present in each individual shot [15]. A special [shot cut] tag is inserted between per-shot prompts to explicitly delineate shot boundaries. This two-tier structure provides the model with both global context and fine-grained, temporally localized guidance.

3.2. Holistic Multi-Shot Generation

The foundation of HoloCine is its holistic generation process, where the latent representations for all shots in a video are processed simultaneously within the diffusion model. This joint processing, primarily through a shared self-attention mechanism, allows the model to naturally maintain long-range consistency in aspects like character identity, background, and overall style, ensuring cohesion across all shot boundaries.

While this holistic design is powerful for maintaining consistency, its practical implementation requires careful

consideration of two key aspects. First, the model needs explicit guidance to align specific per-shot instructions with their corresponding visual segments. Without a mechanism to localize control, the textual guidance for any given shot would be "diluted" by the context of the entire prompt, making it difficult to execute precise control over the per-shot content and shot boundaries. Second, the computational cost of full self-attention, which scales quadratically $(O(L^2))$ with the sequence length L, becomes a prohibitive bottleneck for generating longer, minute-scale videos.

Our architecture directly integrates two specialized mechanisms to address these aspects: Window Cross-Attention for precise directorial control, and Sparse Inter-Shot Self-Attention for computational efficiency.

3.2.1. Window Cross-Attention

The Window Cross-Attention mechanism is designed to provide precise directorial control, addressing two fundamental requirements simultaneously: **what** to generate in each shot and **when** to transition between them. It achieves this by creating a localized link between segments of the video and segments of the text prompt.

Instead of allowing all video tokens to attend to the entire text prompt, our mechanism constrains the attention field to enforce a localized alignment. This is achieved by structuring the attention pattern based on the prompt's hierarchy. For the full sequence of video tokens, the attention each token can pay to the concatenated text prompt is not uniform; rather, it is selectively partitioned. Let Q_i be the query tokens corresponding to the i-th shot. We restrict Q_i to only attend to the key-value pairs derived from the global prompt ($KV_{\text{global}}^{\text{txt}}$) and its corresponding i-th per-shot

prompt (KV_i^{txt}). This operation is formally expressed as:

$$\text{Attention}(Q_i, KV^{\text{txt}}) = \text{Attention}\left(Q_i, \left[KV_{\text{global}}^{\text{txt}}, KV_i^{\text{txt}}\right]\right) \tag{1}$$

This localized attention provides an unambiguous signal for the model to execute sharp, temporally-aligned shot transitions, effectively allowing the text prompt to "direct" the shot cuts.

3.2.2. Sparse Inter-Shot Self-Attention

While the holistic design enables high-quality generation, applying full self-attention across the entire sequence of video tokens is computationally prohibitive for longer videos. To overcome this, we propose a Sparse Inter-Shot Self-Attention mechanism, which drastically reduces complexity while preserving necessary information flow.

Our key intuition is that the nature of consistency differs within a shot versus between shots. Specifically, **intra-shot consistency** demands dense, frame-to-frame temporal modeling to ensure smooth motion and action continuity. In contrast, **inter-shot consistency** is primarily concerning the persistence of characters, environment, and style—which does not require every frame of one shot to attend to every frame of another. Based on this, we structure our self-attention as follows.

Intra-Shot Attention: Within each shot i, we perform full, bidirectional self-attention. The query tokens Q_i from shot i attend to all key-value pairs KV_i from the same shot.

Inter-Shot Attention: To facilitate information exchange between shots, we create a global context summary. For each shot j, we select a small, representative subset of its key-value tokens, $KV_{\text{summary},j}$ (e.g., tokens from the first frame of this shot). These summary tokens from all shots are concatenated to form a global key-value cache, $KV_{\text{global}} = [KV_{\text{summary},1}, \ldots, KV_{\text{summary},N_{\text{shots}}}]$. The query tokens Q_i from shot i also attend to this global cache.

The complete self-attention for shot i is formulated as:

$$\text{Attention}(Q_i, KV) = \text{Attention}\left(Q_i, [KV_{\text{global}}, KV_i,]\right) \tag{2}$$

This design drastically reduces computational complexity. If a video has N_s shots of length L_{shot}, and each shot is summarized by S tokens, the total complexity of full attention would be $O((N_s L_{\text{shot}})^2)$. Our sparse attention, however, reduces this to approximately $O(N_s \times (L_{\text{shot}}^2 + L_{\text{shot}} \cdot N_s \cdot S))$. Since S (e.g., the number of tokens in one frame) is much smaller than L_{shot}, this complexity is significantly lower and scales much more favorably with the number of shots, making it feasible to holistically generate minute-level and longer multi-shot videos. We conduct ablation studies on the method for selecting summary tokens, such as using the first frame, first and last frames, or a learnable mechanism.

4. Experiments

In this section, we present a comprehensive experimental evaluation of our proposed framework, `HoloCine`. Sec. 4.1 describes the training and implementation details of `HoloCine`. Sec. 4.2 introduce the baselines and metrics for the cinematic multi-shot video generation task, and demonstrate our superior performance over these baselines. In Section Sec. 4.3, we analyze the effect of our key proposed modules, including the Window Cross-Attention and Sparse Inter-Shot Self-Attention mechanisms. In Sec. 4.4, we discuss some advanced capabilities of our model, including emergent memory capability. and controllability of cinematographic language.

4.1. Implementation Details

Training Setup. Our framework is built upon the 14B parameter version of `wan2.2`, a powerful DiT-based video diffusion model, which we adapt for the multi-shot task. We train our model on our curated dataset of 400k multi-shot video samples. The dataset includes videos at multiple duration levels (5s, 15s, and 60s) with a maximum of 13 shots per video, and all samples are processed at a resolution of 480×832. The model is trained for 10k steps with a learning rate of 1×10^{-5} and a linear warmup schedule. The entire training process is conducted on 128 NVIDIA H800 GPUs. To manage the significant memory requirements of training on such long video sequences, we employ a hybrid parallelism strategy, using Fully Sharded Data Parallelism (FSDP) to shard the model parameters and Context Parallelism (CP) to split the long token sequences across multiple devices.

Attention Implementation. The implementation of our proposed attention mechanisms is optimized for efficiency. For our Sparse Inter-Shot Self-Attention, where computational cost is a primary concern, we leverage the highly efficient `varlen` (variable-length) sequence functionality from FlashAttention-3[13]. For each query shot, we construct its corresponding Key and Value context by concatenating its own dense local tokens with the shared global summary tokens. These resulting variable-length sequences are then packed into single tensors, which allows the GPU to compute the complex, sparse attention pattern in a single, optimized kernel launch without any overhead from padding tokens. In contrast, for the Window Cross-Attention, since the text prompt sequences are short and this operation constitutes a small fraction of the total computation, we simply apply an attention mask to restrict the attention region. This approach is highly effective and incurs negligible performance overhead.

4.2. Comparison

Settings. We compare `HoloCine` against three categories of strong baselines representing the main paradigms for
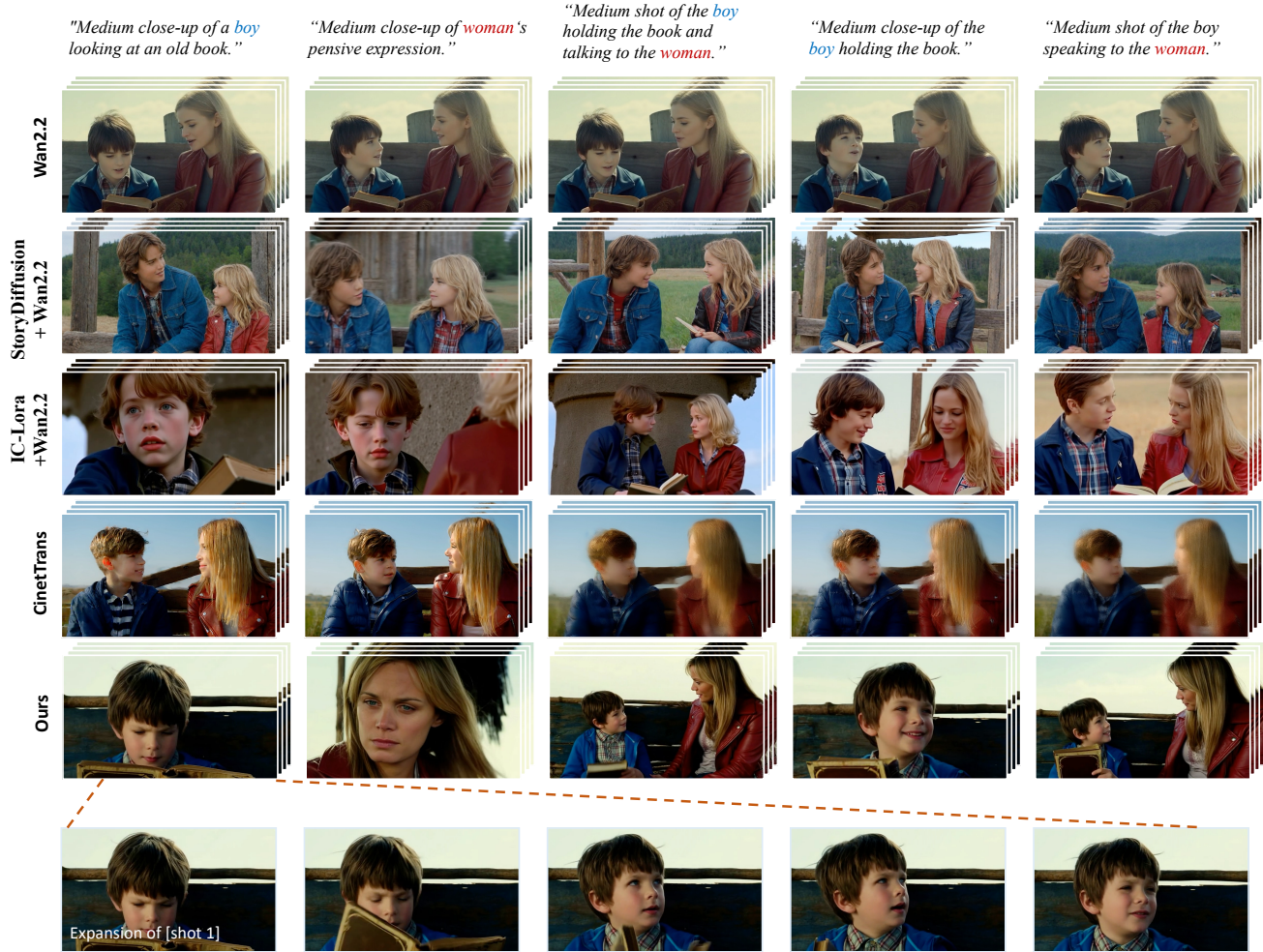
Figure 3. Qualitative comparison on a complex multi-shot prompt. Our method successfully generates a coherent sequence of distinct shots aligned with per-shot descriptions, while baseline methods fail in maintaining consistency, prompt fidelity, or handling shot transitions.

multi-shot long video generation:

- **Pre-trained Video Diffusion Model.** We test the capability of the powerful pre-trained video diffusion model, Wan2.2 14B [41], for the multi-shot task. We provide the model with our full hierarchical prompt (concatenated global and per-shot descriptions) and task it with generating the entire multi-shot sequence in one pass. This baseline evaluates whether a state-of-the-art model can understand and execute multi-shot instructions without our proposed architectural modifications.

- **Two-Stage Keyframe-to-Video Generation.** This paradigm first generates a set of consistent keyframes, one for each shot, and then uses a powerful I2V model to animate them into video clips. We evaluate two state-of-the-art methods for the keyframe generation stage: StoryDiffusion [58], which produces a complete multi-shot image sequence, and IC-LoRA [22], which generates keyframes using in-context learning. For a fair compari-

son, we utilize our base model, the wan2.2 14B, as the I2V component for both pipelines.

- **Holistic Multi-Shot Generation.** We compare against CineTrans [46], a most recent work that also performs holistic generation of multi-shot videos.

To facilitate a comprehensive evaluation of the multi-shot video generation task, we constructed a new benchmark dataset. We leveraged the capabilities of Gemini 2.5 Pro [11] to generate 100 diverse hierarchical text prompts, each containing explicit instructions for shot transitions. This test set spans a wide range of genres and narrative structures, enabling a robust assessment of a model's ability to maintain consistency and control across complex sequences. To ensure a fair comparison, we adapted the hierarchical prompts for the two-stage methods. We generated a distinct prompt for each shot by merging the global context with the shot-specific instructions. This process involved resolving abstract character ID tags (e.g.,

[character1]) into their full textual descriptions, ensuring all methods received equivalent semantic information.

We note that most related works LCT [15], Mixture of Concept [5], and Captain Cinema [47] are not open-sourced. Therefore, a direct quantitative comparison is not feasible. We will provide a qualitative comparison against their published results in the appendix.

Evaluation Metrics. We evaluate the models on five crucial aspects: overall video quality, semantic consistency(prompt adherence), intra-shot consistency, inter-shot consistency, and transition control. For overall quality, prompt adherence, and intra-shot consistency, we utilize the comprehensive VBench [24] benchmark. To specifically assess inter-shot consistency, we compute a ViCLIP-based similarity score between pairs of shots that are annotated to contain the same character. Furthermore, to better evaluate the model's ability to follow explicit shot-cut instructions, we propose the Shot Cut Accuracy (SCA) metrics. More details on these evaluation metrics are presented in Sec. A. This metric holistically assesses shot control by quantifying both the accuracy of the number of cuts and the temporal precision of their placement.

Quantitative Results. As shown in Tab. 1, our model HoloCine establishes a new state-of-the-art by achieving superior performance on the vast majority of metrics. It achieves the top scores across all categories central to the multi-shot task: Transition Control, Inter-shot Consistency, Intra-shot Consistency, and Semantic Consistency. While we note that StoryDiffusion+Wan2.2 performs slightly better on Aesthetic Quality, we argue that our holistic generation method, which produces all shots within a unified modeling process, is fundamentally better suited for this task. This architectural choice is precisely why HoloCine excels at maintaining consistency and control, validating its effectiveness in creating coherent narratives where prior paradigms have struggled.

Qualitative Results. In Fig. 3, we provide a qualitative comparison on a complex narrative prompt to illustrate the superiority of our method. The pre-trained base model, Wan2.2, fails to comprehend the multi-shot instructions, producing only a single, static shot without any transitions. The two-stage methods, while capable of generating different images, struggle with both prompt fidelity and long-range consistency. For example, the prompt for the second shot is "Medium close-up of woman's pensive expression," yet both StoryDiffusion + Wan2.2 and IC-LoRA + Wan2.2 generate a medium shot of the boy and woman together. Their struggle with long-range consistency is especially evident in shots 4 and 5, where the characters' features diverge significantly from the initial shots. The complexity of the prompt and the required length of the video also prove challenging for CineTrans, causing significant image degradation and preventing it from correctly performing

the specified shot transitions. In contrast, our method successfully parses the hierarchical prompt to generate a coherent sequence of five distinct shots. As shown, each shot precisely matches its corresponding text description while maintaining high character and stylistic consistency throughout the entire video, demonstrating the effectiveness of our holistic generation approach.

Comparison with Commercial Models. To further situate HoloCine's capabilities, we conducted a qualitative comparison with leading closed-source commercial models. As illustrated in Fig. 4, while models like Vidu [37] and Kling 2.5 Turbo [28] generate visually impressive clips, they struggle with the core task of multi-shot storytelling. Given a hierarchical prompt, they produce a single, continuous shot, failing to understand or execute the specified shot transitions. In contrast, HoloCine demonstrates narrative comprehension and control on par with the latest state-of-the-art model, Sora 2 [31]. Both models successfully parse the prompt to generate a coherent sequence of distinct shots—transitioning from a medium shot to a dramatic close-up—while maintaining high character and stylistic consistency. This result validates that our framework's ability to create complex, directed narratives is comparable to the leading proprietary solutions in the field.

4.3. Ablation Studies

We perform a series of ablation studies to validate our key architectural choices. The qualitative results are presented in Fig. 5. To facilitate rapid experimentation, all ablation studies were conducted on the wan2.2 5B model.

Window Cross-Attention. Without our window cross-attention, this model exhibited a severe degradation in shot control, as evidenced by a significantly lower Shot Cut Accuracy (SCA) and per-shot semantic consistency score. As illustrated in the top row of Fig. 5, the model fails to execute shot cuts, ignoring prompt instructions for new content (e.g., the close-up in Shot 3) and remaining locked into the initial scene. This confirms that our windowed attention is crucial for precise shot boundary and content control.

Sparse vs. Full Self-Attention. We then compare our sparse self-attention to the full, dense attention baseline. While both produce high-quality, consistent videos (second and fourth rows in Fig. 5), the full attention model is computationally prohibitive for generating long sequences. Our sparse attention mechanism, in contrast, provides a highly effective trade-off. It retains the vast majority of the generative quality while offering a fundamental improvement in efficiency and scalability, making complex, scene-level generation feasible.

Inter-Shot Summary Token. A critical aspect of our sparse attention design is the inter-shot communication facilitated by summary tokens, where each shot attends to

Table 1. **Quantitative results**. The best and runner-up are in **bold** and <u>underlined</u>.

Method	Transition Control↑	Inter-shot Consistency↑	Intra-shot Consistency		Aesthetic Quality↑	Semantic Consistency	
			Subject↑	Background↑		Global↑	Shot↑
Wan2.2	0.4843	0.6772	0.9054	0.9014	0.5568	0.1652	0.1364
CineTrans	<u>0.5370</u>	0.6152	0.8990	0.8998	0.4789	0.1568	0.1159
IC-LoRA+Wan2.2	-	0.7096	<u>0.9421</u>	<u>0.9303</u>	0.5246	<u>0.1808</u>	<u>0.1692</u>
StoryDiffusion+Wan2.2	-	<u>0.7364</u>	0.8487	0.8927	**0.5773**	0.1453	0.1644
HoloCine(Ours)	**0.9837**	**0.7509**	**0.9448**	**0.9352**	<u>0.5598</u>	**0.1856**	**0.1837**

Figure 4. Qualitative comparison with state-of-the-art commercial models. While Vidu and Kling 2.5 Turbo fail to interpret multi-shot instructions and generate only a single, continuous clip, `HoloCine` successfully executes complex shot transitions. Our method demonstrates narrative control and consistency comparable to the leading closed-source model, Sora 2, accurately rendering the sequence from medium shots to close-ups as directed by the prompt.

Table 2. **Ablations**. The best is in **bold**.

Method	Transition Control↑	Inter-shot Consistency↑	Aesthetic Quality↑	Semantic Consistency↑
WO WINDOW	0.6266	0.7009	**0.5755**	0.1562
FULL ATT WINDOW	0.8923	**0.7231**	0.5700	0.1738
SPARSE ZERO	0.9675	0.6761	0.5669	0.1642
SPARSE	**0.9736**	0.7225	0.5693	**0.1739**

the first-frame token of all other shots. To ablate this, we trained a variant where self-attention is confined strictly within each shot, disabling this information exchange. This results in a catastrophic loss of consistency (third row in Fig. 5), where the old man's identity and appearance change drastically between shots. This demonstrates that our inter-shot summary token mechanism is the critical component for maintaining narrative continuity and character consistency across the entire scene.

4.4. Advanced Capabilities

4.4.1. Emergent Memory Capability

Beyond generating high-quality and coherent shots, our model exhibits surprising emergent memory. This capability suggests the model is not merely learning superficial visual transitions but is building an implicit and persistent representation of scenes and objects. We demonstrate this memory in three key aspects.

Object/Character Permanence Across Viewpoints. Our model maintains consistent character identity across varying shots and angles. In Fig. 6(a), for instance, the artist's key features — her blonde hair, grey t-shirt, and apron—remain identical across a medium shot [Shot 2], a profile view [Shot 3], and a subsequent smiling shot [Shot 6], demonstrating a stable character representation.

Long-range Consistency and Re-appearance. The model demonstrates robust long-range consistency, recalling subjects after being interrupted by completely different shots.
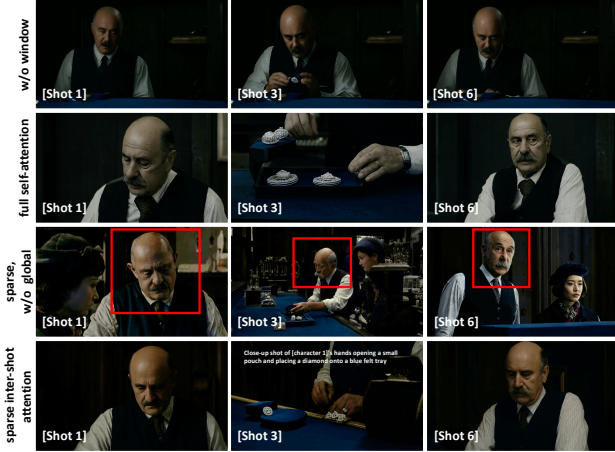
Figure 5. Qualitative results for our ablation study. We compare our full model (bottom row) against three variants. From top to bottom: removing window cross-attention prevents shot transitions; a full self-attention baseline works well but is computationally expensive; and removing inter-shot summary tokens leads to a complete loss of character consistency.

Fig. 6(b) shows an A-B-A sequence where a professor, introduced in [Shot 1], is accurately regenerated in [Shot 5] after a distractor shot of the library environment [Shot 2]. His distinct appearance is perfectly preserved, proving a memory that extends beyond adjacent shots.

Fine-grained Detail Persistence. Crucially, the model's memory extends to fine-grained, non-salient details, indicating a holistic scene understanding. As illustrated in Fig. 6(c), a specific blue magnet (highlighted) appears in the background of [Shot 1]. After an intervening shot, the model correctly recalls and renders the exact same magnet in its original position in [Shot 5], despite it not being a central element of the prompt.
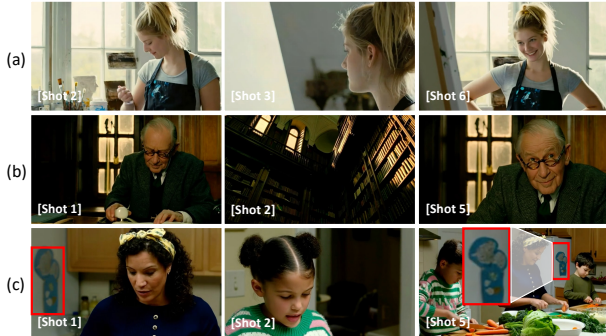

Figure 6. Qualitative results of our model's emergent memory capability. (a) Character Permanence: The subject's identity and appearance are consistently maintained across different camera angles and expressions. (b) Long-range Consistency: The subject is accurately recalled after an unrelated distractor shot. (c) Fine-grained Persistence: A non-salient background detail (the magnet, highlighted) is correctly preserved across an intervening shot.

4.4.2. Controllability of Cinematographic Language

By training on a vast corpus of cinematic data and high-level descriptive prompts, our model has developed a nuanced understanding of filmmaking techniques. Consequently, it exhibits high fidelity in interpreting and executing standard cinematographic commands, enabling precise narrative and stylistic control.

Shot Scale Control. The model accurately renders standard shot scales. As shown in Fig. 7(a), given prompts for a [Long shot], [Medium shot], and [Close-up shot] of the same statue, the model generates outputs that correctly correspond to established cinematographic definitions.

Camera Angle Control. Our model precisely follows the camera angle commands specified in the text prompt. As shown in Fig. 7(b), when prompted with [Low-angle], [Eye-level], and [High-angle] descriptions for the same subject, the model generates the corresponding views accurately. This demonstrates its ability to translate textual cinematographic instructions into correct geometric camera placements within the scene.

Camera Movement Control. Our model is capable of producing a wide range of dynamic and fluid camera movements specified in the prompt. As demonstrated in Fig. 7(c), the model accurately executes these commands to create compelling visual narratives. For instance, a [Tilt up] command generates a smooth vertical camera motion, gracefully revealing the full height of the tree. A [Dolly out] command results in the camera physically moving backward, progressively unveiling the broader context of the artist's studio. Furthermore, a [Tracking] shot correctly follows the motion of a subject, in this case, keeping the soaring eagle centered in the frame. This mastery over camera movement is crucial for creating professional and engaging cinema sequences.

4.5. Limitations

While our model excels at maintaining visual consistency, it exhibits limitations in causal reasoning. It can fail to comprehend how an action should alter an object's physical state. Fig. 8 illustrates this clearly. Given an empty glass [Shot 1] and the action of water being poured into it [Shot 2], the model fails to render the logical outcome. Instead, it regenerates the glass as empty in [Shot 3], prioritizing visual consistency with the initial shot over the physical consequence of the action. This highlights a key challenge for future work: advancing from perceptual consistency to logical, cause-and-effect reasoning.

5. Conclusion

In this work, we bridge the *"narrative gap"* in text-to-video generation with `HoloCine`, a holistic framework that synthesizes entire multi-shot scenes to ensure global
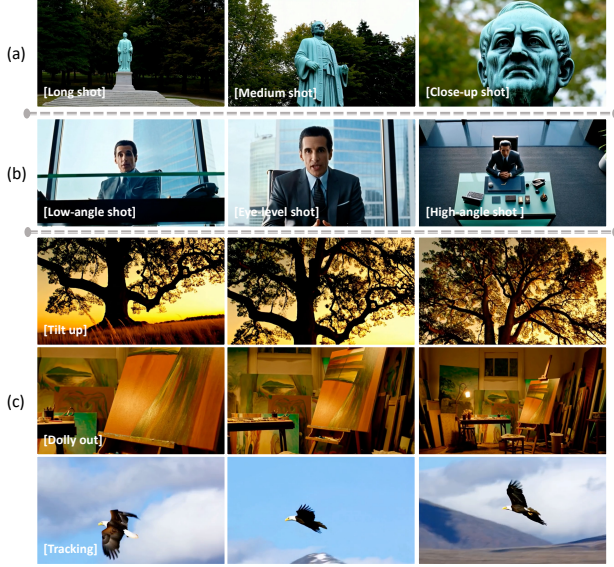
Figure 7. Controllability of Cinematographic Language. Our model demonstrates high fidelity in executing specific cinematic commands. (a) Shot Scale: The model accurately generates long, medium, and close-up shots. (b) Camera Angle: It correctly interprets low-angle, eye-level, and high-angle commands to set the camera's viewpoint. (c) Camera Movement: The model produces fluid and precise camera movements as prompted, including tilt up, dolly out, and tracking shots.

Figure 8. A failure case in causal reasoning. After an action (pouring water, [Shot 2]) is applied to an object (empty glass, [Shot 1]), the model fails to render its logical consequence. It incorrectly reverts to the initial empty state in [Shot 3], prioritizing visual consistency over the action's outcome.

narrative coherence. Our architecture achieves precise directorial control through a Window Cross-Attention mechanism while overcoming prohibitive computational costs with a Sparse Inter-Shot Self-Attention, making minute-scale generation feasible. `HoloCine` not only establishes a new state-of-the-art in consistency and shot control but also develops remarkable emergent capabilities, such as persistent memory for characters and a nuanced understanding of cinematic language. While our work identifies causal reasoning as a key challenge for future research, `HoloCine` represents a critical step towards the automated creation of complex visual narratives. By enabling minute-scale holistic generation, it shifts the paradigm from isolated clips to directing entire scenes, making end-to-end film generation a tangible and exciting future.

References

[1] Md. Tahmeed Abdullah, Sejuti Rahman, Shafin Rahman, and Md. Fokhrul Islam. VAE-GAN3D: leveraging image-based semantics for 3d zero-shot recognition. *Image Vis. Comput.*, 147:105049, 2024. 2

[2] Hritik Bansal, Yonatan Bitton, Michal Yarom, Idan Szpektor, Aditya Grover, and Kai-Wei Chang. TALC: time-aligned captions for multi-scene text-to-video generation. *CoRR*, abs/2405.04682, 2024. 3

[3] 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. *CoRR*, abs/2311.15127, 2023. 2

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. 2

[5] Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan L. Yuille, Leonidas J. Guibas, Maneesh Agrawala, Lu Jiang, and Gordon Wetzstein. Mixture of contexts for long video generation. *CoRR*, abs/2508.21058, 2025. 3, 7

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Int. Conf. Comput. Vis.*, pages 9630–9640. IEEE, 2021. 1

[7] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Adv. Neural Inform. Process. Syst.*, 2024. 2, 3

[8] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7310–7320. IEEE, 2024. 2

[9] Nan Chen, Mengqi Huang, Yihao Meng, and Zhendong Mao. Longanimation: Long animation generation with dynamic global-local memory. *CoRR*, abs/2507.01945, 2025. 2, 3

[10] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Trans. Graph.*, 39(4):75, 2020. 2

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé,

Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith She-shan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reit-ter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leich-ner, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis, Jieru Mei, and Mu Cai. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025. 4, 6

[12] Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 17702–17711. Computer Vision Foundation / IEEE, 2025. 3

[13] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christo-pher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Adv. Neural Inform. Process. Syst.*, 35:16344–16359, 2022. 3, 5

[14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rec-tified flow transformers for high-resolution image synthesis. In *Int. Conf. Mach. Learn.* OpenReview.net, 2024. 3

[15] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. *CoRR*, abs/2503.10589, 2025. 2, 3, 4, 7

[16] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2568–2577. Computer Vision Foundation / IEEE, 2025. 2, 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst.*, 2020. 2

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Sali-mans. Imagen video: High definition video generation with diffusion models. *CoRR*, abs/2210.02303, 2022. 2

[19] Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Adv. Neural Inform. Process. Syst.*, 2022. 2

[20] Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. Storyagent:

[21] Kaiyi Huang, Yukun Huang, Xintao Wang, Zinan Lin, Xuefei Ning, Pengfei Wan, Di Zhang, Yu Wang, and Xihui Liu. Filmaster: Bridging cinematic principles and generative AI for automated film generation. *CoRR*, abs/2506.18899, 2025. 2

[22] 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. *CoRR*, abs/2410.23775, 2024. 2, 3, 6

[23] Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video cus-tomization on diffusion transformer models without test-time tuning. *CoRR*, abs/2501.04698, 2025. 2

[24] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Com-prehensive benchmark suite for video generative models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21807–21818. IEEE, 2024. 7, 1

[25] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6689–6700. IEEE, 2024. 2

[26] Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Cey-lan, James M. Rehg, and Tobias Hinz. Shotadapter: Text-to-multi-shot video generation with diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 28405–28415. Computer Vision Foundation / IEEE, 2025. 3

[27] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *CoRR*, abs/2412.03603, 2024. 2, 3

[28] Kuaishou. Kling video model. https://kling.kuaishou.com/en, 2024. 2, 3, 7

[29] Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, Maneesh Agrawala, Ion Stoica, Kurt Keutzer, and Song Han. Radial attention: O(n log n) sparse atten-tion with energy decay for long video generation. *CoRR*, abs/2506.19852, 2025. 3

[30] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videostudio: Generating consistent-content and multi-scene

videos. In *Eur. Conf. Comput. Vis.*, pages 468–485. Springer, 2024. 2, 3

[31] OpenAI. Sora 2 technical report. `https://openai.com/research/sora-2`, 2025. Accessed: 2025-10-15. 7

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Int. Conf. Comput. Vis.*, pages 4172–4182. IEEE, 2023. 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021. 1

[34] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *Trans. Mach. Learn. Res.*, 2024, 2024. 2, 3

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10674–10685. IEEE, 2022. 3

[36] Christoph Schuhmann. Improved aesthetic predictor. `https://github.com/christophschuhmann/improved-aesthetic-predictor`, 2022. 1

[37] Shengshu Technology and Tsinghua University. Vidu: A sora-level text-to-video model. `https://www.vidu.com`, 2024. Accessed: 2025-10-15. 7

[38] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *Int. Conf. Learn. Represent.* OpenReview.net, 2023. 2

[39] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3616–3626. IEEE, 2022. 2

[40] Tomás Soucek and Jakub Lokoc. Transnet V2: an effective deep network architecture for fast shot transition detection. In *ACM Int. Conf. Multimedia*, pages 11218–11221. ACM, 2024. 3, 1

[41] 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, Xiaofeng Meng, 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. *CoRR*, abs/2503.20314, 2025. 2, 3, 6

[42] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *CoRR*, abs/2305.18264, 2023. 2, 3

[43] Wen Wang, Canyu Zhao, Hao Chen, Zhekai Chen, Kecheng Zheng, and Chunhua Shen. Autostory: Generating diverse storytelling images with minimal human efforts. *Int. Conf. Comput. Vis.*, 133(6):3083–3104, 2025. 3

[44] Yaohui Wang, Piotr Bilinski, François Brémond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal GAN for video generation. In *IEEE Winter Conf. Appl. Comput. Vis.*, pages 1149–1158. IEEE, 2020. 2

[45] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *Int. Conf. Learn. Represent.* OpenReview.net, 2024. 1

[46] 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. 2, 3, 6

[47] Junfei Xiao, Ceyuan Yang, Lvmin Zhang, Shengqu Cai, Yang Zhao, Yuwei Guo, Gordon Wetzstein, Maneesh Agrawala, Alan L. Yuille, and Lu Jiang. Captain cinema: Towards short movie generation. *CoRR*, abs/2507.18634, 2025. 2, 3, 7

[48] 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. *CoRR*, abs/2408.11788, 2024. 3

[49] Jinbo Xing, Menghan Xia, Yong Zhang, Hao 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 *Eur. Conf. Comput. Vis.*, pages 399–417. Springer, 2024. 2

[50] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *Int. Conf. Learn. Represent.* OpenReview.net, 2025. 2, 3

[51] YaoFANGUK. video-subtitle-extractor: A gui tool for extracting hard-coded subtitles from videos. `https://github.com/YaoFANGUK/video-subtitle-extractor`, 2021. 3

[52] Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22963–22974. Computer Vision Foundation / IEEE, 2025. 2, 3

[53] Jintao Zhang, Jia Wei, Haofeng Huang, Pengle Zhang, Jun Zhu, and Jianfei Chen. Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. *arXiv preprint arXiv:2410.02367*, 2024. 3

[54] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *CoRR*, abs/2504.12626, 2025. 2, 3

[55] Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhenghong Liu, and Hao Zhang. Fast video generation with sliding tile attention. *CoRR*, abs/2502.04507, 2025. 3

[56] Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequences. In *Int. Conf. Learn. Represent.* OpenReview.net, 2025. 2, 3

[57] Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. Videogen-of-thought: A collaborative framework for multi-shot video generation. *CoRR*, abs/2412.02259, 2024.

[58] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. In *Adv. Neural Inform. Process. Syst.*, 2024. 2, 6

HoloCine: Holistic Generation of Cinematic Multi-Shot Long Video Narratives

Supplementary Material

Appendix

A. Details on Evaluation metrics

We evaluate the models across five key dimensions: **aesthetic quality**, **semantic consistency**, **intra-shot consistency** (capturing subject and background stability), **inter-shot consistency**, and **transition control**.

A.1. Transition Control Evaluation Metrics

Furthermore, to comprehensively evaluate the model's ability to to follow explicit shot-cut instructions, we propose the Shot Cut Accuracy (SCA) metrics,specified in appendix A. This metric holistically assesses shot control by quantifying both the accuracy of the number of cuts and the temporal precision of their placement. To compute the SCA, we first apply a pre-trained shot boundary detector, TransNet V2 [40], to the generated video to obtain the set of cut locations in the generated videos. These are then compared against the ground truth cut locations specified in the input. SCA is defined as:

$$\text{SCA} = \exp(-\text{NSD}) \tag{S1}$$

where NSD is the Normalized Shot Discrepancy, representing the total error relative to the video's length in frames, F_{total}:

$$\text{NSD} = \frac{E_{\text{matched}} + E_{\text{penalty}}}{F_{\text{total}}} \tag{S2}$$

E_{matched} quantifies the frame-wise deviation between predicted and ground-truth cuts after a one-to-one matching process. E_{penalty} is a penalty term for any missed or extraneous cuts. This ensures that models are penalized not only for imprecise timing but also for failing to produce the correct number of shots. The SCA score ranges in $(0, 1]$, where 1 indicates a perfect match. The exponential formulation makes the metric particularly sensitive to large errors, heavily penalizing significant temporal deviations.

A.2. Aesthetic Quality

We assess the aesthetic and artistic value of each video frame using the LAION aesthetic predictor [36]. This metric reflects human-perceived qualities such as composition, color harmony, realism, naturalness, and overall artistic appeal of the generated frames.

A.3. Semantic Consistency.

We evaluate the alignment between the text prompt and the generated video by measuring two types of semantic consistency: global and shot-level. For global consistency, we extract the representations of the entire prompt and the full video using ViCLIP [45] and compute their cosine similarity. For shot-level consistency, the video is divided into segments based on the input shot prompts, and the cosine similarity between each shot clip and its corresponding shot-level prompt features is calculated using ViCLIP.

A.4. Intra-shot Consistency

To compute intra-shot consistency, we first employ the pre-trained shot boundary detector TransNet V2 [40] to identify cut locations within the generated videos. We then compute subject consistency and background consistency, following the design of VBench [24].

Subject Consistency. For the main subject in the video, we measure the stability of its visual appearance across frames. Specifically, we extract DINO [6] features for each frame and compute the average cosine similarity between consecutive frames and between each frame and the first frame.

Background Consistency. To evaluate the temporal stability of the scene background, we compute CLIP [33] feature similarities across frames. A higher similarity indicates a smoother and more coherent background transition over time.

A.5. Inter-shot Consistency

To assess consistency across different shots, a naive approach would be to extract ViCLIP features for each shot and compute the cosine similarity between them. However, since different shots may depict distinct characters or scenes, this simple comparison ignores diversity and may lead to biased results. To address this, we identify the characters described in the prompt and group the corresponding shots by character identity. We then compute the ViCLIP feature similarity among shots belonging to the same character group to obtain the inter-shot consistency score.