SMARTDIRECTOR: KEYFRAME-CONDITIONED CINEMATIC VIDEO GENERATION WITH NARRATIVE PACING CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

The narrative quality of a video fundamentally determines its perceptual value. Although existing video generation methods can produce visually appealing content, they predominantly rely on sparse conditioning signals such as text prompts or first/last frames, which limits precise control over narrative structure and temporal pacing. In this paper, we propose SmartDirector, a framework that enhances the narrative capacity of video generation models through multiple keyframes. SmartDirector supports flexible generation scenarios including single-shot generation and multi-shot narrative synthesis. The framework operates in two stages: Director-Gen generates a low-resolution video conditioned on the provided keyframes, and Director-SR refines the output by exploiting high-resolution keyframes as semantic anchors to recover fine-grained details. To enable robust multi-keyframe training, we construct a data pipeline that curates single-shot and multi-shot sequences from movies. Director-Gen improves all six metrics over a same-backbone pairwise-generation baseline on the common 220-example KeyFrame-Compass subset. We further evaluate Director-SR independently on four standard video super-resolution benchmarks and assess the complete two-stage pipeline through a controlled Gen-only versus Gen+SR comparison.

1 INTRODUCTION

Recent advancements in video generation have propelled a paradigm shift from synthesizing short, single-shot clips (Team Wan et al., 2025; Kong et al., 2024; HaCohen et al., 2024) to creating long, multi-shot narratives (Wang et al., 2025c; Kling Team et al., 2025; Meng et al., 2025; Xiao et al., 2026). Although existing methods have demonstrated remarkable capabilities in generating visually stunning and high-fidelity videos, they predominantly rely on sparse conditioning signals, such as text prompts or first/last frames. Consequently, these approaches struggle to achieve precise control over fine-grained spatial-temporal content and narrative structure, significantly restricting their practical utility in real-world applications. In professional filmmaking, directors use storyboards to guide the production process and exercise fine-grained control over visual content. Storyboards serve as visual anchors that maintain coherence across multiple shots and regulate the temporal pacing (i.e., the rhythm and timing of visual content) within each individual shot. We therefore treat keyframes as the direct counterpart of storyboards and focus on multi-keyframe-conditioned video generation.

A naive approach treats each pair of adjacent keyframes as the endpoints of a short clip, generates the clips independently, and concatenates them. This neglects global context, causing discontinuities at keyframe boundaries and weakening narrative consistency. Recent work (Liu et al., 2025b;a) instead inserts keyframes into noisy latents at corresponding temporal positions before denoising. Yet this approach is limited by the causal temporal VAE (Team Wan et al., 2025; Kong et al., 2024; Yang et al., 2025). In a standard 3D VAE, the first frame is encoded independently, while later frames are encoded in groups with dependence on preceding frames. Replacing latents at arbitrary positions breaks this dependency and produces artifacts near the keyframes.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **Examples generated by SmartDirector.** SmartDirector enables high-fidelity video generation guided by multiple temporally positioned keyframes.

In this paper, we introduce SmartDirector, a flexible framework for video generation guided by multiple keyframes at user-specified temporal locations that supports both single-shot and multi-shot synthesis. To fully exploit the conditioning provided by multiple keyframes, the framework consists of two stages: a keyframe-conditioned generation stage and a keyframe-conditioned super-resolution stage, referred to as Director-Gen and Director-SR. In this work, we implement Director-Gen on the MiniMax H3 FL2VA backbone and evaluate it through controlled comparisons and ablations. We evaluate Director-SR separately on standard video super-resolution benchmarks and assess the complete pipeline through a controlled Gen-only versus Gen+SR comparison. In the Director-Gen stage, we propose a Multi-Chunk VAE strategy to address the causal limitation of the temporal VAE. During training, the video is partitioned into multiple chunks at the keyframe positions, with each keyframe serving as the first frame of its respective chunk and encoded independently by the VAE. The resulting multi-chunk latents are then processed by a Diffusion Transformer (DiT) (Peebles & Xie, 2023). To maintain global consistency, we apply full spatio-temporal attention within the DiT, enabling each chunk to attend to the global context across all chunks. Director-Gen produces low-resolution videos (e.g., 480p). Director-SR then upsamples them to high definition (e.g., 1080p), using high-resolution keyframes to recover fine details.

Training our framework requires carefully curated data. We construct a data processing pipeline for curating long video sequences, as illustrated in Fig. 3. For the Director-Gen stage, we use a company-internal cinematic video corpus that cannot be redistributed, segment the videos into single shots, and employ a Vision-Language Model (VLM) to assess visual and semantic similarity among consecutive shots before aggregating them into coherent multi-shot sequences. We further annotate the sequences with structured descriptions using VLMs (Bai et al., 2025; Gemini Team et al., 2023). The resulting dataset contains both single-shot and multi-shot sequences, enabling robust training for both generation settings. For the super-resolution task, we use the open-source UltraVideo dataset (Xue et al., 2025). Our main contributions are summarized as follows:

- We propose SmartDirector, a unified framework that enables flexible keyframe-conditioned video generation in both single-shot and multi-shot settings.
- We identify the fundamental limitation imposed by the causal structure of the temporal VAE on keyframe insertion and propose a Multi-Chunk VAE strategy. This design respects the causal constraints while supporting user-specified keyframe locations up to the bounded temporal discretization imposed by the VAE.
- We design a keyframe-conditioned super-resolution module that exploits high-resolution keyframes as semantic anchors to recover fine-grained details.

2 Related Work

2.1 Video Generation

Video generation has evolved rapidly from synthesizing short single-shot clips (Team Wan et al., 2025; Kong et al., 2024; HaCohen et al., 2024; Bao et al., 2024) to broader long-range and cinematic generation (Zhou et al., 2024; Wu et al., 2026; Wang et al., 2025c; Kling Team et al., 2025; Meng et al., 2025; Xiao et al., 2026). However, these methods rely on sparse conditioning signals such as text prompts or the first/last frame, which limits their ability to control fine-grained spatial-temporal content and narrative structure. Recently, several approaches have attempted to incorporate multiple keyframes into the generation process to enable more precise control. For single-shot video generation, Pusa (Liu et al., 2025b) injects noise of different timesteps into distinct frames, while DreaMontage (Liu et al., 2025a) directly inserts keyframes into noisy latents at corresponding positions. Frame Guidance (Jang et al., 2026) introduces training-free guidance for frame-level control in video diffusion models, enabling sparse visual conditions at specified temporal locations. For multi-shot video generation, CaptainCinema (Xiao et al., 2026) generates video by conditioning on the first frames of each shot. However, these methods are limited to specific scenarios and lack the flexibility to support multiple temporally positioned keyframes for precise temporal and spatial control. We instead support flexible keyframe control for both single-shot and multi-shot generation.

2.2 Video Super Resolution

Video super-resolution has been studied extensively over the past decades. Early approaches were predominantly based on GANs (Chu et al., 2020; Chan et al., 2022), while recent methods have shifted toward diffusion models (Wang et al., 2025a; Chen et al., 2025b; Yu et al., 2026). SeedVR (Wang et al., 2025a) introduces a shifted window attention mechanism to enable effective restoration on long video sequences. DoVE (Chen et al., 2025b) proposes an efficient one-step diffusion model for real-world video super-resolution. However, existing VSR methods primarily focus on pixel-level enhancement, often treating each frame as a restoration target rather than a semantic object to be reconstructed. Consequently, they struggle to address common artifacts in low-resolution videos generated in the first stage, such as distorted small faces and incorrect text. A concurrent work, SparkVSR (Yu et al., 2026), also explores keyframe-conditioned video super-resolution. In contrast, our Director-SR is designed as the refinement stage of a unified keyframe-conditioned generation framework, using multiple high-resolution keyframes as semantic anchors to reconstruct fine-grained details and correct generative artifacts throughout the sequence.

3 Method

SmartDirector is a two-stage framework that comprises a keyframe-conditioned generation stage (Director-Gen) and a keyframe-conditioned super-resolution stage (Director-SR), as illustrated in Fig. 2. This section first provides a brief review of the Flow Matching framework, then details the proposed method, and finally describes the data curation pipeline.

3.1 Flow Matching

Let $z_0 \sim Z_0$ be a data sample and $z_1 \sim Z_1$ a noise sample. Recent image generation models adopt the Rectified Flow framework (e.g., Esser et al., 2024). The framework is also common in video generation (e.g., Team Wan et al., 2025; Kong et al., 2024; Chen et al., 2025a). Rectified Flow (Liu et al., 2023) defines the interpolated latent z_t as

$$z_t = (1-t)z_0 + tz_1, \tag{1}$$

for $t \in [0, 1]$. The model is trained to regress the velocity field $\boldsymbol{v}_\theta(z_t, t)$ by minimizing the Flow Matching objective (Lipman et al., 2023):

$$\mathcal{L}(\theta) = \mathbb{E}_{t,z_0,z_1}\left[\|\boldsymbol{v} - \boldsymbol{v}_\theta(z_t, t)\|^2\right], \tag{2}$$

where the target velocity field is $\boldsymbol{v} = z_1 - z_0$.

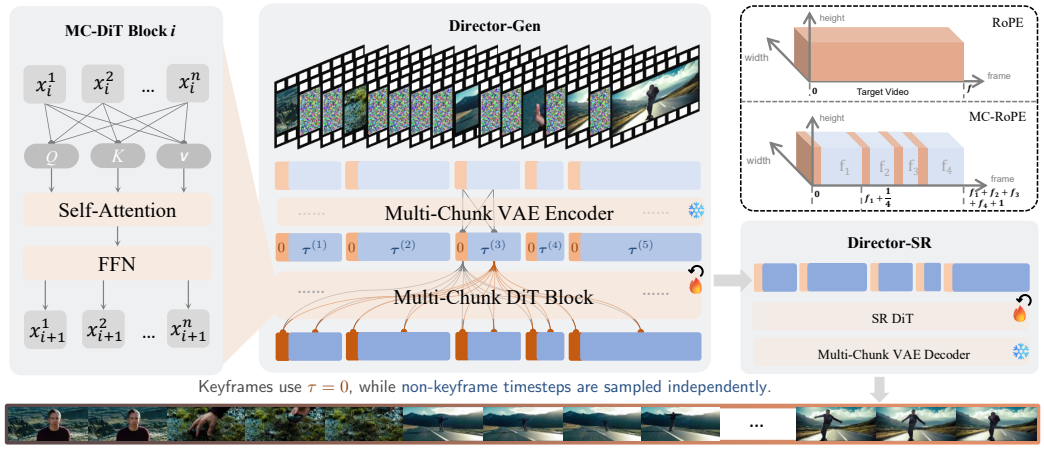

Figure 2: **Overview of the proposed SmartDirector framework.** SmartDirector is a two-stage framework for keyframe-conditioned video generation. In the first stage, Director-Gen synthesizes a low-resolution video from the provided keyframes via a Multi-Chunk VAE that encodes chunks independently at keyframe boundaries and a Multi-Chunk DiT with full spatio-temporal attention and MC-RoPE for coherent cross-chunk modeling. Orange boundary latents remain clean with timestep zero. Each $\boldsymbol{\tau}^{(j)}$ contains independently sampled timesteps for the non-keyframe latent frames in chunk j. In the second stage, Director-SR refines the low-resolution output into a high-resolution video under the guidance of high-resolution keyframes.

3.2 DIRECTOR-GEN

Training Director-Gen requires a set of videos, each paired with a structured caption c and a set of keyframes $\{I_k\}$ (k denotes the keyframe index). To address the causal limitation of the 3D VAE, we propose a Multi-Chunk VAE strategy.

We first split the video V into n video chunks $\{V_j\}$ (j denotes the chunk index) at the keyframe positions, ensuring that each keyframe serves as the first frame of its respective chunk. For simplicity, we assume the first frame is always provided as a keyframe (i.e., I_0), so the number of chunks equals the number of keyframes. Because the temporal VAE requires each chunk to contain $4m+1$ video frames for a non-negative integer m, a requested keyframe location is aligned to the nearest valid chunk boundary. The resulting discrepancy is bounded by two video frames. We encode each clean chunk with a 3D causal VAE $\mathcal{E}$ for spatiotemporal compression, yielding latents $z_j = \mathcal{E}(V_j)$. Through this design, each keyframe is encoded independently by the VAE. After encoding, we use a chunk-wise vectorized timestep: entries at keyframe positions are fixed to zero, while timestep values at non-keyframe latent positions are sampled independently. We then apply the flow-matching interpolation in Eq. 1 position-wise only to the non-keyframe latents, while keeping the keyframe latents clean. The training loss is likewise evaluated only at non-keyframe positions. At inference, the same masking scheme keeps the provided keyframe latents fixed throughout sampling. Following the native H3 formulation, the caption c is encoded by the backbone's text encoder, and the resulting hidden states are packed with the visual tokens and jointly processed through self-attention in the single-stream transformer. These chunk latents are patchified into visual tokens $\{x_j \in \mathbb{R}^{f_j \times d \times h \times w}\}$, where f_j, d, h, and w denote the latent frame count, channel number, height, and width, respectively. We concatenate all chunk tokens along the temporal dimension to form a unified latent sequence:

$$x = \mathrm{Concat}(\{x_j\}), \quad x \in \mathbb{R}^{(\sum f_j) \times d \times h \times w}. \tag{3}$$

The sequence x is then processed by the DiT. To maintain global consistency, we apply full spatio-temporal attention across all chunks.

The DiT applies H3's native 3D Rotary Positional Embeddings (RoPE) to the packed visual tokens. In our multi-chunk representation, the first latent of each chunk is a boundary latent that represents a single keyframe, whereas each regular latent summarizes r consecutive video frames, where r is the VAE's temporal compression ratio. Using one increment for both types treats them as equally long,

while resetting coordinates at each chunk loses the global temporal order. We therefore propose **Multi-Chunk RoPE (MC-RoPE)**, which assigns boundary-aware temporal coordinates to the concatenated chunks. In normalized regular-latent-step units, a regular latent contributes an increment of 1, while a boundary latent contributes an increment of $1/r$. Specifically, let $\mathcal{K}$ denote the set of latent indices that begin a new chunk, excluding the initial index. For $i \geq 1$, we construct the temporal coordinate u_i as

$$u_i = \begin{cases} u_{i-1} + 1, & \text{if } i \notin \mathcal{K}, \\ u_{i-1} + \frac{1}{r}, & \text{if } i \in \mathcal{K}, \end{cases} \tag{4}$$

with $u_0 = 0$. The VAE used by the model has a temporal compression ratio of $r = 4$, so the normalized boundary increment is $1/r = 0.25$.

3.3 DIRECTOR-SR

Videos produced by the Director-Gen stage are low-resolution (e.g., 480p) due to the computational cost of diffusion models. Consequently, they struggle to preserve fine details such as facial features and text, limiting their practical applicability. Existing video super-resolution (VSR) methods focus on pixel-level restoration and lack precise frame-level control, making them insufficient for correcting generative artifacts introduced in the Director-Gen stage.

To exploit the high-resolution keyframes as semantic anchors, we design a keyframe-conditioned super-resolution module in the Director-SR stage. During training, each sample consists of paired low-resolution (LR) and high-resolution (HR) videos, denoted as V^{LR} and V^{HR}. A subset of HR frames is sampled from V^{HR} to serve as keyframes $\{I_k\}$. Following existing methods (Zhuang et al., 2025; Yu et al., 2026), V^{LR} is synthesized by applying degradation operations to V^{HR}. As in the Director-Gen stage, we adopt the Multi-Chunk VAE strategy to respect the causal structure of the VAE, obtaining the HR latents z^{HR} and LR latents z^{LR}. Let $\mathcal{U}(z^{LR})$ denote the LR latent after spatial upsampling to the dimensions of z^{HR}, and let $\mathcal{K}$ denote the keyframe latent indices. We construct the conditioned LR endpoint $\bar{z}^{LR}$ as

$$\bar{z}_i^{LR} = \begin{cases} z_i^{HR}, & i \in \mathcal{K}, \\ \mathcal{U}(z^{LR})_i, & i \notin \mathcal{K}. \end{cases} \tag{5}$$

We define $z_0 = z^{HR}$ as the clean target and $z_1 = \bar{z}^{LR}$ as the conditioned LR endpoint. Consistent with Eq. 1, the interpolation and target velocity are

$$z_t = (1-t)z^{HR} + t\bar{z}^{LR}, \qquad \boldsymbol{v} = \bar{z}^{LR} - z^{HR}. \tag{6}$$

Director-SR is trained with the standard flow-matching loss

$$\mathcal{L}_{\mathrm{SR}}(\theta) = \mathbb{E}_t \left[\|\boldsymbol{v} - \boldsymbol{v}_\theta(z_t, t)\|_2^2 \right]. \tag{7}$$

At inference, sampling starts from the conditioned LR endpoint at $t = 1$ and integrates the learned velocity field backward to $t = 0$ to recover the HR latent. The HR latent replacements at the keyframe indices provide the keyframe conditioning. Director-SR is designed to refine Director-Gen outputs but can also operate independently on low-resolution videos.

3.4 DATA PIPELINE

Training SmartDirector requires a large corpus of videos paired with structured captions that describe both the overall narrative and the content of each individual shot. To this end, we build a scalable data pipeline that proceeds in three steps, as illustrated in Fig. 3.

Video collection and shot segmentation. We use a non-redistributable company-internal cinematic video collection. Each video is segmented into single-shot clips with AutoShot (Zhu et al., 2023). To construct multi-shot samples that preserve narrative continuity, a vision-language model assesses the visual and semantic similarity of consecutive shots and aggregates shots that share the same scene and storyline, yielding multi-shot videos with coherent semantics.

Structured video captioning. We then annotate each video with a systematic, multi-aspect caption. To describe camera behavior, we combine VGGT (Wang et al., 2025b) for geometric camera trajectory estimation with Qwen3-VL (Bai et al., 2025) for visual interpretation (e.g., pan, zoom, dolly). To

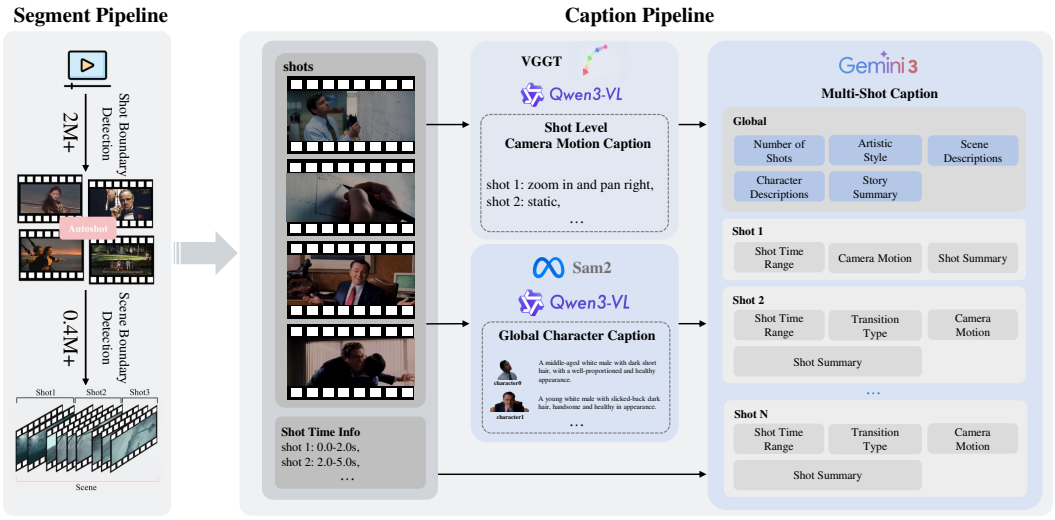

Figure 3: **Overview of the data pipeline.** Starting from large-scale cinematic videos, we perform shot segmentation with AutoShot and VLM-based multi-shot aggregation, and then generate structured captions that cover camera motion, per-character appearance, and a holistic description of the whole multi-shot sequence together with descriptions of each individual shot.

characterize on-screen subjects, we track each character with SAM2 (Ravi et al., 2024) across the entire video and generate an appearance description for every tracked identity via Qwen3-VL, ensuring consistent character grounding across shots.

Hierarchical caption aggregation. Finally, we feed the shot-level visual content, camera descriptions, and character descriptions into Gemini-3-Pro (Gemini Team et al., 2023) to produce a structured caption in a unified format. The caption contains a holistic description that summarizes the overall narrative of the multi-shot video, as well as per-shot descriptions that specify the visual content, camera motion, and active characters of each individual shot.

4 Experiments

We organize the experiments around five questions: (i) whether Director-Gen improves multi-keyframe execution under a common evaluation protocol, (ii) how it compares with results reported for a broader set of video generation systems, (iii) which components are responsible for the gain, (iv) how Director-SR performs independently on video super-resolution benchmarks, and (v) how super-resolution affects the complete two-stage pipeline.

4.1 Experimental Setup

Generation benchmark. KeyFrame-Compass contains 227 short-video cases (61 one-take and 166 multi-shot) (Tang et al., 2026). The public Seedance 2.0 interface rejects seven cases during platform content and access validation, including copyright-related restrictions, and returns no generated video. We therefore fix the remaining 220 accepted cases before evaluating any model outputs and use exactly these example IDs for every reproduced method. The supplementary artifact `keyframe_compass220_ids.txt` provides the retained ID list. We issue a single Seedance generation request per retained case and perform no retry-based or output-based filtering. Director-Gen, the same-backbone baseline, PuSA, and Frame Guidance receive the same prompts and ordered keyframes. Seedance 2.0 receives the corresponding prompts and visual references through the controls exposed by its interface, without imputing unsupported controls. Every method produces one 480p video per retained example. All evaluation settings follow the official open-source KeyFrame-Compass protocol. Output duration and frame rate are matched where each interface permits, and a fixed seed policy is used for methods that expose seed control.

Table 1: **Comparison on KeyFrame-Compass.** Results in (a) are evaluated on our fixed evaluation set using the method-specific input controls described in the text, whereas results in (b) are copied from the benchmark paper. The best and second-best values in (a) for each metric are shown in bold and underlined, respectively.

Method	HR↑	KFS↑	KOC↑	KPA↑	SVQ↑	DVQ↑
(a) Evaluation on our fixed evaluation set						
MiniMax H3	0.910	0.634	<u>0.994</u>	0.590	0.629	0.412
PuSA	<u>0.998</u>	0.639	<u>0.994</u>	0.707	0.467	0.426
Frame Guidance	0.402	0.243	0.786	<u>0.771</u>	0.322	0.484
Seedance 2.0 480p	0.992	<u>0.662</u>	0.986	0.695	**0.873**	**0.796**
Director-Gen (ours)	**1.000**	**0.783**	**0.999**	**0.905**	<u>0.856</u>	<u>0.774</u>
(b) KeyFrame-Compass results reported by Tang et al. (2026)						
Seedance 2.0†	0.961	0.652	0.984	0.717	0.862	0.844
Kling-3.0-Omni†	0.778	0.587	0.959	0.659	0.827	0.800
Gemini-Omni-Flash†	0.623	0.371	0.901	0.587	0.864	0.861
LTX-2.3†	0.967	0.735	0.985	0.873	0.665	0.518
Wan2.7-I2V†	0.614	0.368	0.748	0.376	0.778	0.688

Implementation details. We fine-tune all DiT parameters of the 33B MiniMax H3 FL2VA backbone for 30,000 steps on 64 NVIDIA H20 GPUs with a learning rate of 5×10^{-6}. For Director-SR, we fine-tune all DiT parameters of Wan-2.2-5B for 2,000 steps on 8 NVIDIA GPUs with a learning rate of 2×10^{-5}. The two stages use separate VAE paths: Director-Gen outputs are decoded to video and then re-encoded as inputs to Director-SR. The Multi-Chunk VAE, full spatio-temporal attention, and MC-RoPE follow Section 3. Locally run methods use matched resolution, video length, preprocessing, keyframes, and prompts; hosted systems are matched wherever their interfaces expose the corresponding controls.

Baselines. We additionally construct a same-backbone baseline using MiniMax H3 FL2VA. Given an ordered sequence of keyframes, it independently generates each interval conditioned on two adjacent keyframes and concatenates the generated segments. Its data, prompts, keyframe conditions, inference settings, and evaluation protocol are identical to Director-Gen. We also compare with PuSA (Liu et al., 2025b), Frame Guidance (Jang et al., 2026), and Seedance 2.0 at 480p (Team Seedance et al., 2026). PuSA and Frame Guidance are run with the same ordered keyframes and prompts as Director-Gen. Seedance 2.0 is evaluated with the same prompts and ordered visual references through its publicly available 480p generation interface. We separately report results copied from the KeyFrame-Compass paper to avoid conflating paper-reported and reproduced numbers.

4.2 Comparison with State-of-the-Art Methods

KeyFrame-Compass metrics. To keep the main comparison focused, we report six representative visual metrics: Hit Rate (HR), Keyframe Similarity (KFS), Keyframe Order Consistency (KOC), Keyframe Position Accuracy (KPA), Static Visual Quality (SVQ), and Dynamic Visual Quality (DVQ). These cover keyframe reproduction, keyframe ordering and localization, static appearance quality, and motion quality. Higher is better for all metrics in the tables below.

The upper panel is designed to isolate the effect of explicit multi-keyframe conditioning. The MiniMax H3 FL2VA baseline uses exactly the same backbone and evaluation setup as Director-Gen, but generates each adjacent-keyframe interval independently before concatenation. Director-Gen improves all six metrics over this same-backbone baseline and obtains the best HR, KFS, KOC, and KPA in the controlled comparison. In particular, KPA measures whether keyframe events occur at their requested temporal positions and thus directly evaluates the temporal control underlying narrative pacing. Seedance 2.0 achieves higher SVQ and DVQ, while Director-Gen ranks second on both visual-quality metrics, indicating strong keyframe execution and temporal localization without sacrificing overall visual quality.

The lower panel expands the context of the comparison without claiming a direct reproduction. The dagger marks values copied from the segment-specific short-video setting of the benchmark paper

Table 2: **Ablation study.** Keyframe-execution and visual-quality metrics are reported on the same evaluation set as the main comparison.

Variant	HR↑	KFS↑	KOC↑	KPA↑	SVQ↑	DVQ↑
Direct insertion	0.661	0.359	0.858	0.582	0.537	0.518
Standard RoPE	0.767	0.431	0.875	0.525	0.556	0.560
Chunk-local attention	0.803	0.450	0.932	0.550	0.571	0.598
Full Director-Gen	1.000	0.783	0.999	0.905	0.856	0.774

and rounded to three decimal places. In particular, the proprietary systems are reported for context only. Their source-paper sample intersection and prompt protocol differ from our evaluation setup. All claims under a controlled setting refer to the upper panel of Table 1.

4.3 Ablation Study

We run all ablations on the same evaluation set with fixed training budget, prompts, keyframes, and sampling parameters. **Direct insertion** removes the Multi-Chunk VAE. **Standard RoPE** uses the backbone's native scaled temporal coordinates without the boundary-specific rescaling of MC-RoPE. **Chunk-local attention** prevents attention across chunks. Each variant changes only the indicated component.

Full Director-Gen outperforms all variants on all six metrics. Chunk-local attention is the strongest ablation on five metrics but remains clearly behind the full model, supporting the importance of cross-chunk context exchange. Standard RoPE reduces KPA from 0.905 to 0.525 because the native temporal indexing does not distinguish a single-frame chunk boundary from a regular temporally compressed latent transition. Since HR only counts valid matches within the expected temporal window and unmatched keyframes contribute zero to KFS, this temporal mismatch can also reduce HR and KFS. These results suggest that the boundary-aware coordinates of MC-RoPE improve temporal placement and maintain keyframe correspondence across chunks. Direct insertion has the highest KPA among the ablations but substantially lower HR, KFS, and KOC, indicating that temporal placement alone does not ensure faithful keyframe realization and ordering.

4.4 Director-SR Results

Director-SR is evaluated independently on UDM10, SPMCS, YouHQ40, and RealVSR. For context, we quote the PiSA-SR-reference results from the SparkVSR paper (Yu et al., 2026). SparkVSR uses training-matched degradations on the three synthetic datasets, native pairs on RealVSR, and the initial frame as reference. Because these are paper-reported results rather than a common rerun, the comparison is contextual rather than strictly controlled. We report the three shared metrics: PSNR, SSIM, and LPIPS.

Quantitative results. The four-benchmark evaluation reports reconstruction fidelity and perceptual similarity. PSNR and SSIM measure pixel-level recovery, while LPIPS captures perceptual similarity in faces, text, and small objects.

Among the reported values, Director-SR has lower LPIPS on all four datasets and higher PSNR and SSIM on SPMCS and YouHQ40, whereas SparkVSR has higher PSNR and SSIM on UDM10 and RealVSR. This pattern suggests a perceptual-fidelity trade-off rather than uniform dominance. The qualitative comparison further illustrates facial, textual, and object details not fully captured by aggregate scores.

4.5 End-to-End Gen+SR Evaluation

Table 4 evaluates the complete pipeline by applying Director-SR to 480p Director-Gen outputs on the same 220 cases. HR and KFS test whether the supplied keyframes remain recognizable after refinement. KOC and KPA measure whether their order and temporal locations are preserved. SVQ and DVQ measure appearance and motion quality at 1080p. Together, these metrics show whether refinement preserves the control established by Director-Gen.

Table 3: Reference comparison on four video super-resolution benchmarks. SparkVSR† values come from its PiSA-SR-reference variant. Best values are bold, and all results use three decimal places.

Method	UDM10			SPMCS			YouHQ40			RealVSR		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SparkVSR†	**23.430**	**0.671**	0.355	20.120	0.491	0.339	21.750	0.579	0.350	**19.720**	**0.618**	0.217
Director-SR	22.780	0.636	**0.202**	**21.010**	**0.521**	**0.224**	**22.120**	**0.582**	**0.137**	18.440	0.603	**0.146**

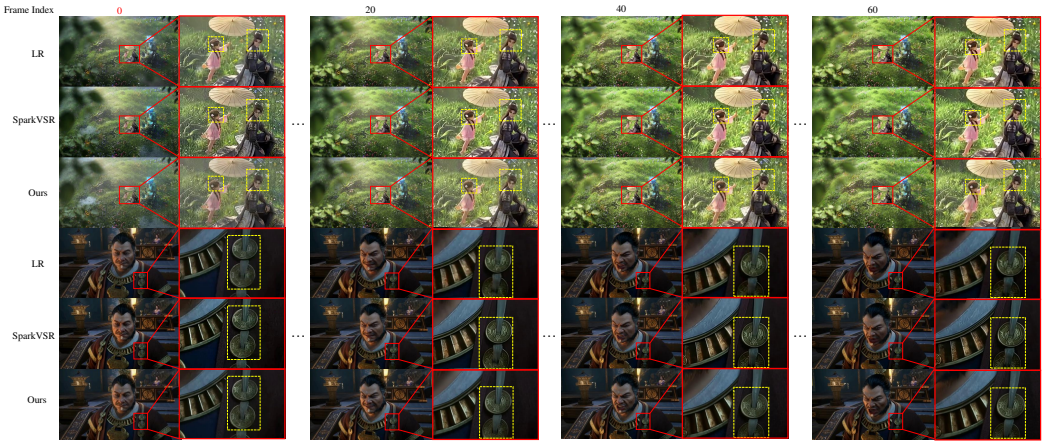

Figure 4: **Qualitative comparison with SparkVSR.** In the displayed examples, Director-SR produces clearer facial structures and textual or symbolic details.

Table 4: **End-to-end effect of Director-SR on KeyFrame-Compass.** Both rows use the same Director-Gen outputs and evaluation cases. Higher is better for all metrics.

Pipeline	HR↑	KFS↑	KOC↑	KPA↑	SVQ↑	DVQ↑
Director-Gen (480p)	1.000	0.783	0.999	0.905	0.856	0.774
Director-Gen + Director-SR (1080p)	1.000	0.749	1.000	0.900	0.824	0.762

Director-SR increases the output resolution from 480p to 1080p while largely retaining the keyframe control and visual quality of Director-Gen. All six metrics remain close to their 480p values, with HR unchanged at 1.000 and KOC reaching 1.000. Because KeyFrame-Compass resizes videos before computing SVQ and DVQ, these scores assess resolution-normalized quality rather than directly rewarding native resolution. Together with the standalone SR results, this demonstrates substantially higher-resolution video without materially compromising quality or control.

5 Conclusion

We present SmartDirector, a two-stage framework for keyframe-conditioned cinematic video generation with fine-grained control over narrative structure and temporal pacing. Director-Gen synthesizes a low-resolution video from multiple keyframes, while Director-SR uses their high-resolution counterparts as semantic anchors for refinement. To respect the causal 3D VAE, our Multi-Chunk VAE partitions the latent sequence at keyframe boundaries. Full spatio-temporal attention and MC-RoPE then support cross-chunk context exchange and temporal continuity. On the MiniMax H3 FL2VA backbone, Director-Gen is effective in keyframe execution and temporal coherence across single-shot and multi-shot scenarios. SR benchmarks and the Gen-only versus Gen+SR comparison show that Director-SR raises resolution from 480p to 1080p without sacrificing control or visual quality.

AI USE STATEMENT

In the research workflow, vision-language models were used to aggregate and annotate real-video data and to perform model-based evaluation, as described in the corresponding method and evaluation sections. During manuscript preparation, generative AI tools were used to edit prose for clarity and grammar. All AI-assisted outputs and manuscript edits were reviewed by the authors, who take responsibility for the final content of this submission.

REPRODUCIBILITY AND DATA AVAILABILITY STATEMENT

Section 3 describes the model and data pipeline, and Section 4 specifies the evaluation metrics, subset-selection rule, and comparison protocols. The MiniMax H3 FL2VA generation evaluations use the public KeyFrame-Compass benchmark, and the fixed 220-example list is provided in `keyframe_compass220_ids.txt`. All reproduced methods use this preselected subset and the official open-source evaluation settings. The appendix separately reports portability evidence from a proprietary production implementation evaluated on an internal cinematic benchmark. The raw training videos and proprietary model cannot be redistributed because of corporate confidentiality and source-specific usage restrictions.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: A highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *CVPR*, pp. 5962–5971, 2022.

Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. In *CVPR*, pp. 23516–23527, 2025a.

Zheng Chen, Zichen Zou, Kewei Zhang, Xiongfei Su, Xin Yuan, Yong Guo, and Yulun Zhang. DOVE: Efficient one-step diffusion model for real-world video super-resolution. In *NeurIPS*, 2025b.

Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM TOG*, 39(4):75–1, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

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

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Sangwon Jang, Taekyung Ki, Jaehyeong Jo, Jaehong Yoon, Soo Ye Kim, Zhe Lin, and Sung Ju Hwang. Frame Guidance: Training-free guidance for frame-level control in video diffusion models. In *ICLR*, 2026.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-Omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

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

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Jiawei Liu, Junqiao Li, Jiangfan Deng, Gen Li, Siyu Zhou, Zetao Fang, Shanshan Lao, Zengde Deng, Jianing Zhu, Tingting Ma, et al. DreaMontage: Arbitrary frame-guided one-shot video generation. *arXiv preprint arXiv:2512.21252*, 2025a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Yaofang Liu, Yumeng Ren, Aitor Artola, Yuxuan Hu, Xiaodong Cun, Xiaotong Zhao, Alan Zhao, Raymond H Chan, Suiyun Zhang, Rui Liu, et al. Pusa V1.0: Unlocking temporal control in pretrained video diffusion models via vectorized timestep adaptation. *arXiv preprint arXiv:2507.16116*, 2025b.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4195–4205, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Yuqi Tang, Tengfei Liu, Yizheng Lai, Yuran Wang, Yang Shi, Wanshun Su, Zhuoran Zhang, Qixun Wang, Xiaohan Zhang, Xinlei Yu, et al. KeyFrame-Compass: Towards comprehensive evaluation of keyframe-conditioned video generation. *arXiv preprint arXiv:2607.14202*, 2026.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In *CVPR*, pp. 2161–2172, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pp. 5294–5306, 2025b.

Qinghe Wang, Xiaoyu Shi, Baolu Li, Weikang Bian, Quande Liu, Huchuan Lu, Xintao Wang, Pengfei Wan, Kun Gai, and Xu Jia. MultiShotMaster: A controllable multi-shot video generation framework. *arXiv preprint arXiv:2512.03041*, 2025c.

Xiaoxue Wu, Bingjie Gao, Yu Qiao, Yaohui Wang, and Xinyuan Chen. CineTrans: Learning to generate videos with cinematic transitions via masked diffusion models. In *ICLR*, 2026.

Junfei Xiao, Ceyuan Yang, Lvmin Zhang, Shengqu Cai, Yang Zhao, Yuwei Guo, Gordon Wetzstein, Maneesh Agrawala, Alan Yuille, and Lu Jiang. Captain cinema: Towards short movie generation. In *ICLR*, 2026.

Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 2025.

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. In *ICLR*, 2025.

Jiongze Yu, Xiangbo Gao, Pooja Verlani, Akshay Gadde, Yilin Wang, Balu Adsumilli, and Zhengzhong Tu. SparkVSR: Interactive video super-resolution via sparse keyframe propagation. *arXiv preprint arXiv:2603.16864*, 2026.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *NeurIPS*, 37:110315–110340, 2024.

Wentao Zhu, Yufang Huang, Xiufeng Xie, Wenxian Liu, Jincan Deng, Debing Zhang, Zhangyang Wang, and Ji Liu. AutoShot: A short video dataset and state-of-the-art shot boundary detection. In *CVPRW*, 2023.

Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. Flashvsr: Towards real-time diffusion-based streaming video super-resolution. *arXiv preprint arXiv:2510.12747*, 2025.

A Main-Experiment Reproducibility Details

This section provides the information needed to reproduce the fixed-subset Director-Gen comparison in the main paper. In particular, it records how the evaluation subset was fixed, how inputs were matched across methods, and how the reported metrics were computed.

A.1 Fixed KeyFrame-Compass Subset

The public KeyFrame-Compass short-video split contains 227 cases: 61 one-take cases and 166 multi-shot cases (Tang et al., 2026). During input validation, the public Seedance 2.0 interface rejected seven cases because of platform content or access restrictions, including copyright-related restrictions, and returned no generated video. These failures occurred before any output was available. We therefore fixed the remaining 220 cases before evaluating any method and used exactly the same case IDs for Director-Gen, the same-backbone baseline, PuSA, Frame Guidance, and Seedance 2.0. No case was removed on the basis of generated content or metric value. The exclusion rule was determined before generation and was independent of generated quality, although the retained subset may reflect Seedance-specific platform restrictions.

The complete ordered list is distributed as `keyframe_compass220_ids.txt`. The file contains 220 non-empty lines and 220 unique IDs.

A.2 Input Matching and Filtering Protocol

For each retained case, we issued one Seedance 2.0 generation request and did not retry failed or low-quality outputs. Director-Gen, the same-backbone MiniMax H3 baseline, PuSA, and Frame Guidance received the same text prompt and the same ordered keyframes. Seedance 2.0 received the corresponding prompt and ordered visual references through the controls exposed by its public interface; we did not impute controls that the interface does not support.

Every reproduced method produced one 480p video for each retained case. Locally run methods used matched video length, preprocessing, prompt formatting, keyframe inputs, and sampling settings. Output duration and frame rate were matched for hosted systems wherever their interfaces exposed the corresponding controls. A fixed seed policy was used for methods that expose seed control. The subset was never changed after generation began.

A.3 Evaluation Configuration

All reproduced outputs were evaluated with the official open-source KeyFrame-Compass protocol. We report Hit Rate (HR), Keyframe Similarity (KFS), Keyframe Order Consistency (KOC), Keyframe Position Accuracy (KPA), Static Visual Quality (SVQ), and Dynamic Visual Quality (DVQ). HR and KFS measure whether the supplied keyframes are recovered; KOC and KPA measure their order and temporal placement; and SVQ and DVQ assess appearance and motion quality. Higher is better for all six metrics.

The same-backbone baseline uses MiniMax H3 FL2VA and independently generates each interval between two adjacent keyframes before concatenating the resulting clips. Its training data, prompts, keyframes, inference settings, and evaluation protocol are identical to those of Director-Gen. Consequently, this comparison is designed to isolate the effect of joint multi-keyframe modeling under the matched local evaluation protocol. Results reproduced on the fixed 220-case subset are reported separately from values copied from the KeyFrame-Compass paper; paper-reported values are included only for context and are not treated as a controlled comparison.

B Portability Evaluation on the Proprietary Production Model

We further evaluate SmartDirector in a proprietary production setting. This study provides complementary evidence that the design can be instantiated in a substantially different backbone and deployment pipeline and can perform strongly in that setting. It is not a controlled transfer or component-isolation experiment. Because this implementation differs from the MiniMax H3 FL2VA

model used in the main paper, its results are reported separately and are not used as measurements of the main-paper implementation.

B.1 IMPLEMENTATION

The proprietary implementation uses the same Director-SR configuration as the main-paper system, so we report only the implementation-specific Director-Gen and benchmark settings here. Its Director-Gen model uses a 32B internal diffusion backbone with the Wan-2.2-VAE and is trained for 20,000 steps on 40 NVIDIA GPUs with a learning rate of 2×10^{-5}. The settings are summarized in Table 5.

Table 5: Director-Gen and benchmark settings for the proprietary-model portability study.

Item	Setting
Director-Gen backbone	32B internal diffusion model
Director-Gen VAE	Wan-2.2-VAE
Director-Gen training	20k steps on 40 GPUs
Director-Gen learning rate	2×10^{-5}
Benchmark	250 single-shot and 250 multi-shot videos
Duration and frame rate	3–15 seconds at 24 FPS
Video resolution	Native resolution of at least 1080p
External baseline	Dreamina Multiframes

B.2 CINEMATIC BENCHMARK

To facilitate a comprehensive quantitative evaluation, we construct a diverse benchmark from movies, television series, and animations. The benchmark comprises 250 single-shot videos and 250 multi-shot videos, with durations ranging from 3 to 15 seconds. All videos are rendered at 24 FPS and have a native resolution of at least 1080p. For each video, we randomly sample a set of keyframes as conditioning signals. To ensure compatibility with the causal structure of the temporal VAE, the number of frames in each chunk satisfies $4n + 1$, where n is a non-negative integer.

We compare the proprietary SmartDirector implementation with Dreamina Multiframes, a closed-source system that supports multi-keyframe conditioning. The benchmark assesses both local keyframe adherence and broader cinematic behavior, including identity consistency, narrative progression, camera transitions, and visual quality.

B.3 FVD AND GEMINI EVALUATION PROTOCOL

We evaluate the generated videos from two complementary automatic perspectives. First, FVD measures distributional fidelity between the generated videos and the real videos in the corresponding single-shot or multi-shot benchmark split. Second, Gemini-3-Pro (Gemini Team et al., 2023) assesses high-level semantic adherence to the prompt and reference keyframes, together with perceptual and cinematic quality that is not captured by distributional statistics alone.

For every generated video, the Gemini evaluator receives the full text prompt, the reference keyframes, and the generated sequence. The model name and method identity are withheld. The evaluator analyzes the complete sequence rather than isolated frames and returns a score from 0 to 5 together with a textual justification for each of the five dimensions in Table 6.

The dimension-specific prompts follow the same four-part structure used in the original evaluation: task definition, evaluation criteria, scoring criteria, and output format. Instruction following considers character, emotion, scene, and spatial consistency. Narrative coherence considers plot progression and action logic. Physical consistency considers gravity, collision and clipping, object permanence, fluid behavior, inertia, and momentum. Video quality focuses on technical artifacts, while aesthetic quality focuses on lighting, color, and composition.

All five prompts use the same six-point rubric: 0 denotes an unacceptable result that completely fails the criterion; 1 denotes a very poor result; 2 a poor result with major violations; 3 a fair result with noticeable flaws; 4 a good result with only minor defects; and 5 an excellent result with near-perfect

Table 6: Dimensions used in the proprietary-model Gemini evaluation.

Dimension	Evaluation focus
Instruction following	Fidelity to the prompt and to the semantic and spatial information anchored by the reference keyframes.
Narrative coherence	Logical event order, narrative completeness, causal progression, and continuity of actions.
Physical consistency	Gravity, contact, collision, object permanence, structural integrity, and motion dynamics.
Video quality	Flicker, blur, noise, geometric distortion, color instability, and loss of fine detail.
Video aesthetic quality	Lighting, color harmony, composition, framing, visual focus, and cinematic appearance.

execution. The raw score $s \in [0, 5]$ is converted to a 0–100 scale as $20s$. The reported average is the arithmetic mean of the five converted dimension scores.

B.4 Human Evaluation Protocol

To assess perceptual quality, we conduct a blind pairwise evaluation between videos generated by the proprietary SmartDirector implementation and Dreamina. Each completed assignment contains ten pairs: five single-shot and five multi-shot examples. The evaluation interface randomizes the anonymous labels A and B for every pair, so participants are not shown the model identity.

For each pair, participants are required to provide an Overall Quality rating. They may additionally assess the following three dimensions:

- **Identity Consistency:** consistency of character identity, attire, accessories, scene settings, and salient objects across time and shot transitions.
- **Narrative Pacing:** coherence and natural rhythm of narrative progression, including the timing of events, scene transitions, and the temporal flow between keyframes.
- **Keyframe Adherence:** accuracy in reconstructing the content, composition, and poses supplied by the reference keyframes.

The retained aggregate contains 380 mandatory Overall Quality responses, equally divided into 190 single-shot and 190 multi-shot responses. Since each assignment contains ten pairs, this corresponds to 38 completed assignments. The optional dimensions contain 319 valid Identity Consistency responses, 317 Narrative Pacing responses, and 316 Keyframe Adherence responses; percentages for these dimensions therefore use separate denominators.

Participants use a five-point comparison scale: (1) A significantly better than B, (2) A slightly better than B, (3) similar quality, (4) B slightly better than A, and (5) B significantly better than A. After resolving the randomized labels, we collapse the responses into SmartDirector wins, ties, and Dreamina wins. The Good/Same/Bad statistic is

$$\mathrm{GSB} = \frac{\mathrm{Wins} - \mathrm{Losses}}{\mathrm{Wins} + \mathrm{Losses} + \mathrm{Ties}}. \tag{8}$$

B.5 Results

As shown in Table 7, SmartDirector achieves substantially lower FVD than Dreamina in both scenarios. In the single-shot setting, FVD decreases from 226.85 to 41.12. The difference is also pronounced in the multi-shot setting (251.83 versus 65.65), where scene transitions and camera cuts introduce additional temporal complexity.

The semantic evaluation shows the same trend. In the single-shot setting, SmartDirector improves the average Gemini score from 83.87 to 91.30, with the largest gain in Narrative Coherence. In the multi-shot setting, the average increases from 59.32 to 88.48, and the improvements extend across all five dimensions.

Table 7: FVD and Gemini-based semantic evaluation on the cinematic benchmark. All SmartDirector values are from the proprietary production implementation.

Method	FVD ↓	Inst.	Narr.	Phys.	Qual.	Aest.	Avg. ↑
Single-shot							
Dreamina	226.85 ± 3.44	89.45	78.46	85.73	79.81	85.90	83.87
Proprietary SmartDirector	**41.12 ± 1.01**	**92.97**	**91.02**	**91.24**	**90.44**	**90.85**	**91.30**
Multi-shot							
Dreamina	251.83 ± 5.87	64.60	57.23	60.52	46.46	67.78	59.32
Proprietary SmartDirector	**65.65 ± 2.46**	**89.70**	**86.08**	**88.75**	**88.18**	**89.69**	**88.48**

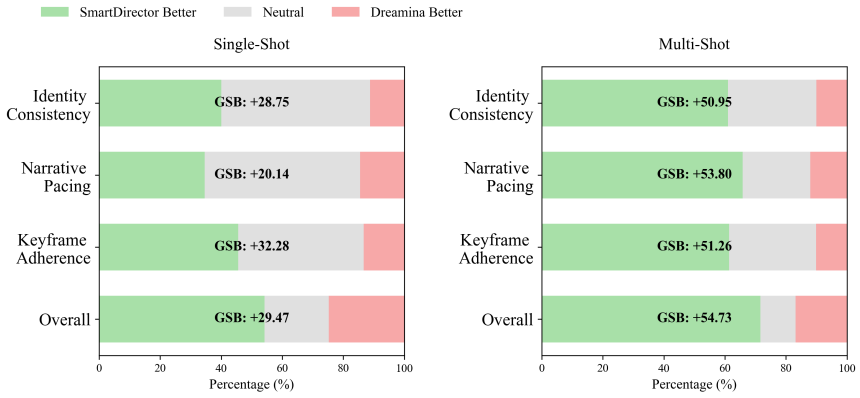

Figure 5: Human evaluation results on the cinematic benchmark using the proprietary SmartDirector implementation.

The human evaluation in Fig. 5 and Table 8 likewise favors SmartDirector. The margin is largest in the multi-shot setting: 71.58% of the mandatory Overall Quality responses prefer SmartDirector, compared with 16.85% for Dreamina and 11.58% ties. Preferences in favor of SmartDirector are also observed across all three optional dimensions. Because participation in these dimensions was optional, their response counts use separate denominators and should not be interpreted as a causal attribution of the Overall Quality preference.

Table 8: Detailed human evaluation breakdown (percentage) on the cinematic benchmark. S and D denote the proprietary SmartDirector implementation and Dreamina, respectively.

Scenario	Dimension	S significantly better	S slightly better	Neutral	D slightly better	D significantly better
All	Overall Quality	39.21	23.68	16.32	8.16	12.63
	Identity Consistency	25.39	25.08	38.87	4.39	6.27
	Narrative Pacing	29.02	21.14	36.59	7.57	5.68
	Keyframe Adherence	26.27	27.22	34.81	5.38	6.33
Single	Overall Quality	28.42	25.79	21.05	11.58	13.16
	Identity Consistency	14.38	25.62	48.75	5.00	6.25
	Narrative Pacing	14.47	20.13	50.94	10.06	4.40
	Keyframe Adherence	17.72	27.85	41.14	6.33	6.96
Multi	Overall Quality	50.00	21.58	11.58	4.74	12.11
	Identity Consistency	36.48	24.53	28.93	3.77	6.29
	Narrative Pacing	43.67	22.15	22.15	5.06	6.96
	Keyframe Adherence	34.81	26.58	28.48	4.43	5.70

C Data and Annotation Details

This section specifies the annotation record and processing order used for the non-redistributable company-internal cinematic corpus. It complements the data pipeline overview in Section 3 without repeating the collection and training discussion from the main paper.

C.1 Annotation Schema

The annotation record separates temporal segmentation, local visual analysis, and global narrative aggregation. As summarized in Table 9, the same record can be used to construct both single-shot and multi-shot training examples.

Table 9: Fields and tools used in the proprietary-model annotation pipeline.

Annotation field	Primary tool	Stored content
Shot boundaries	AutoShot	Start/end indices and candidate shot segments.
Character and object attributes	Qwen3-VL and SAM2	Appearance, identity-bearing regions, tracked characters, and salient objects.
Camera and motion cues	VGGT and Qwen3-VL	Camera movement, viewpoint change, and coarse motion description.
Per-shot content	Qwen3-VL	Entities, actions, spatial relations, and local scene details.
Holistic narrative	Gemini-3-Pro	Ordered event summary and cross-shot temporal relationships.
Training record	Hierarchical aggregation	Single-shot caption, multi-shot caption, ordered keyframes, and shot-time metadata.

C.2 Processing Order

The record is constructed in the following order:

1. AutoShot (Zhu et al., 2023) identifies candidate shot boundaries and produces temporally ordered shot segments.
2. A vision-language model compares consecutive shots and groups visually and semantically compatible segments into a coherent multi-shot sequence.
3. VGGT (Wang et al., 2025b) and Qwen3-VL (Bai et al., 2025) describe camera motion, while SAM2 (Ravi et al., 2024) tracks characters and Qwen3-VL records their appearance and the content of each shot.
4. After all local records are available, Gemini-3-Pro aggregates them into a holistic narrative while retaining the shot-level time ranges, actions, active characters, transitions, and camera descriptions.
5. The local and global annotations are packaged with the ordered keyframes to form the final single-shot or multi-shot training record.

Generating the holistic description only after the per-shot records allows the global summary to condition on complete local information and reduces the risk of losing identities, actions, and camera cues that are required for keyframe-conditioned training. The aggregation stage may also enrich these local records with a higher-level narrative interpretation of the full sequence.

C.3 Example Annotation

The following is a real two-shot annotation produced by the pipeline. It illustrates how the final text combines artistic style, character identity, a holistic story summary, shot-time ranges, transition type, camera motion, and shot-level actions in one training record.

Example Multi-Shot Annotation

This two-shot video, rendered in a Pixar-style 3D animation with delicate fur and soft lighting, follows a hungry orange cat Character[1] scouting from a tree as its mood shifts from curious exploration to intense longing.
Shot 1 [0.0–2.0 s]. The camera stays static as only the right front paw of Character[1] is visible, gripping the tree trunk tightly.
Shot 2 [2.0–6.0 s]. After a hard cut, the camera slowly pans right and zooms in as Character[1] peeks from the left of the trunk with both paws, looking off-screen to the right with a curious expression.