CINEASTRA: REFERENCES AS IN-CONTEXT DIRECTORS FOR MULTI-SHOT VIDEO CREATION

Anonymous authors
Paper under double-blind review

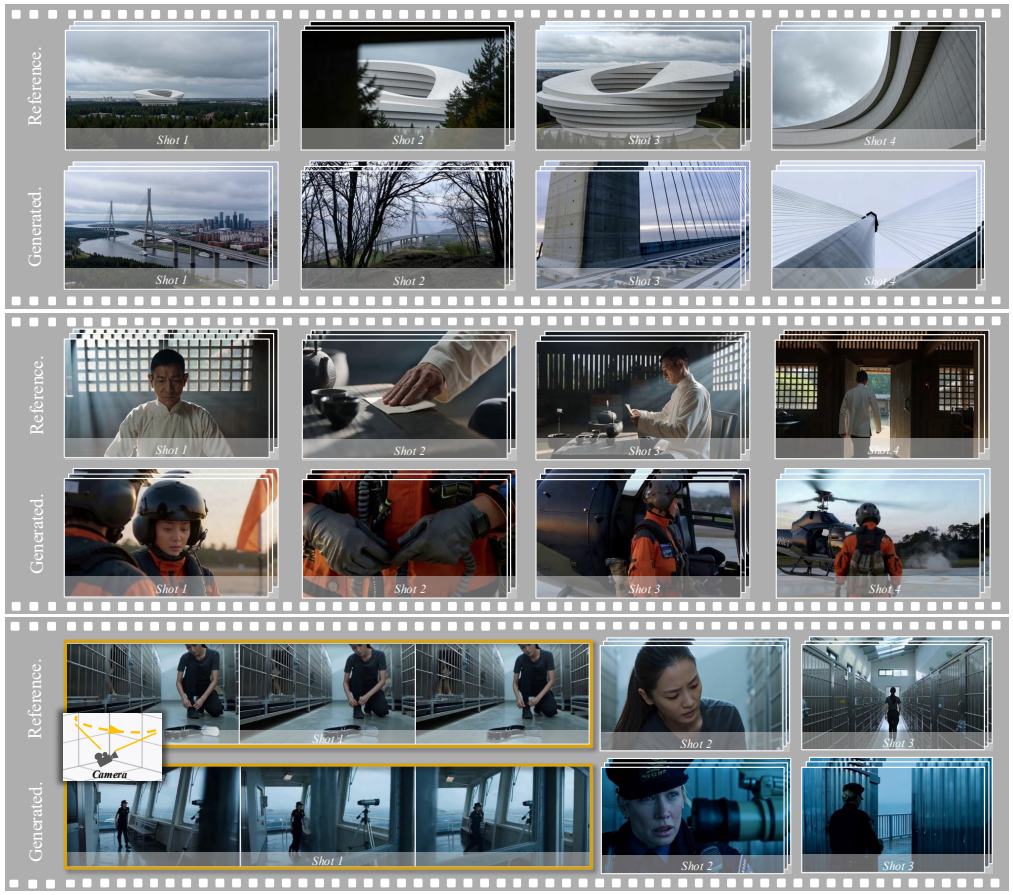

Figure 1: **Reference-guided multi-shot generation results.** Given a reference video and a target screenplay, CINEASTRA generates a sequence that faithfully emulates cinematic language by preserving viewpoint, framing, camera motion and cut timing across new subjects and scenes.

ABSTRACT

Cinematic multi-shot video generation requires both holistic control over cinematic language and versatile support for diverse creation modes. Existing methods largely rely on textual prompts or condition-specific modules, making it difficult to convey complex directorial intent and flexibly accommodate partially observed footage. To address these limitations, we introduce CINEASTRA, a unified framework for reference-guided multi-shot video creation. CINEASTRA leverages in-context learning through a Mixture-of-Transformers (MoT) architecture, where reference and target video streams interact via contextual attention to infer the reference's cinematography and narrative pacing for the target screenplay. To further support flexible creation, we propose frame-wise flow timestep conditioning, which independently manages per-frame noise levels according to assigned flow timesteps, seamlessly integrating target observations at arbitrary temporal positions. Extensive experiments demonstrate that CINEASTRA outperforms existing methods in cinematography adherence, narrative pacing alignment and visual quality, providing a unified solution for versatile multi-shot creation modes.

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.

1 INTRODUCTION

Recent advances in diffusion-based video generation (Blattmann et al., 2023; Wan et al., 2025; Seedance et al., 2026) have substantially improved visual fidelity, motion realism, and semantic alignment, enabling the synthesis of high-quality single-shot clips from textual descriptions. Building on this progress, recent work has extended video generation beyond isolated clips toward multi-shot sequences, opening new possibilities for film and short-form drama production (Meng et al., 2026; Wang et al., 2026c; Qi et al., 2025; An et al., 2026; Wu et al., 2026a;b; Kara et al., 2025; Wang et al., 2026a). To elevate immersion and expressiveness in professional filmmaking, directors orchestrate individual shots through *cinematic language* to establish atmosphere, guide viewer attention, and shape emotional responses. An ideal generative framework should therefore faithfully express such cinematic language by seamlessly integrating within-shot cinematography with cross-shot narrative pacing, rather than merely synthesizing a sequence of visually coherent clips. At the same time, real-world cinematic creation demands accommodating diverse creation modes, where creators may synthesize a sequence from scratch, extend existing footage, or complete missing segments between observed clips (Jiang et al., 2025; Cai et al., 2026). Together, these requirements point to two complementary core capabilities: **(1) holistic control over cinematic language**, which unifies within-shot cinematography (framing, viewpoint and camera motion) with cross-shot narrative pacing (shot count, duration and transitions); and **(2) versatile creation modes**, which flexibly support generation from scratch, continuation, and completion within a single framework.

Regarding the first capability, existing controllable generation methods (Guo et al., 2024; Chen et al., 2023) provide only partial support. Text-driven multi-shot models allow users to describe individual shots, but translating complex directorial intent into prompts rarely ensures precise execution (Zhao et al., 2025; Long et al., 2024). Additional conditions, such as reference images (Jiang et al., 2024), motion trajectories (He et al., 2024; Wang et al., 2024b; Bai et al., 2025a), and cut timing constraints (Kara et al., 2025), provide more direct control over specific visual attributes, yet remain largely unexplored for multi-shot sequences. Furthermore, attempting to express rich cinematic language by piecing together specialized control modules often introduces conflicting control signals and system complexity, making configuration cumbersome for creators. A promising alternative is reference-guided multi-shot video creation. Camera cloning takes an initial step in this direction by reproducing camera motion from reference footage, reducing the need for manual trajectory specification (Luo et al., 2025; Hu et al., 2024; Zhang et al., 2025; Ling et al., 2024). However, explicit trajectory extraction adds preprocessing overhead, and cross-scene scale differences can undermine motion transfer. More fundamentally, camera motion is only one aspect of cinematic language; cloning it alone fails to capture the full coordination of within-shot cinematography and cross-shot narrative pacing. This motivates a perspective shift: existing film works naturally demonstrate how cinematography and narrative pacing work together across shots, offering an intuitive demonstration of directorial intent without combining separate condition-specific controls. A more comprehensive approach should therefore interpret the reference video as an integrated visual context to guide the holistic cinematic language of the entire generated shot sequence.

Regarding the second capability, a practical creation framework must accommodate diverse creation modes with flexible target content controls. In practical scenarios, available target content may be provided at arbitrary temporal positions (Jiang et al., 2025; Cai et al., 2026). Depending on these observations, the model may need to generate a sequence from scratch, extend an initial clip, or complete missing shots between observed frames, all while faithfully preserving the cinematic language demonstrated by the reference video. However, most multi-shot generative models are tailored for specific tasks, such as text-to-video or image-to-video generation. While staged keyframe-to-video pipelines (Zhang et al., 2026; Xiao et al., 2026; Zhou et al., 2024; Zheng et al., 2024) accept keyframe conditions, independently synthesizing individual shots compromises cross-shot coherence. On the other hand, holistic sequence modeling methods improve global consistency (Meng et al., 2026; Wang et al., 2026c;a; Wu et al., 2026b; Kara et al., 2025), but their rigid formulations cannot easily accommodate partially observed content. These limitations highlight the need for a flexible conditioning mechanism that supports diverse temporal arrangements within a single model.

These considerations raise a central question: *Can a single unified model learn holistic cinematic language from a reference video while flexibly accommodating arbitrary target observations for generation, continuation, and completion?*

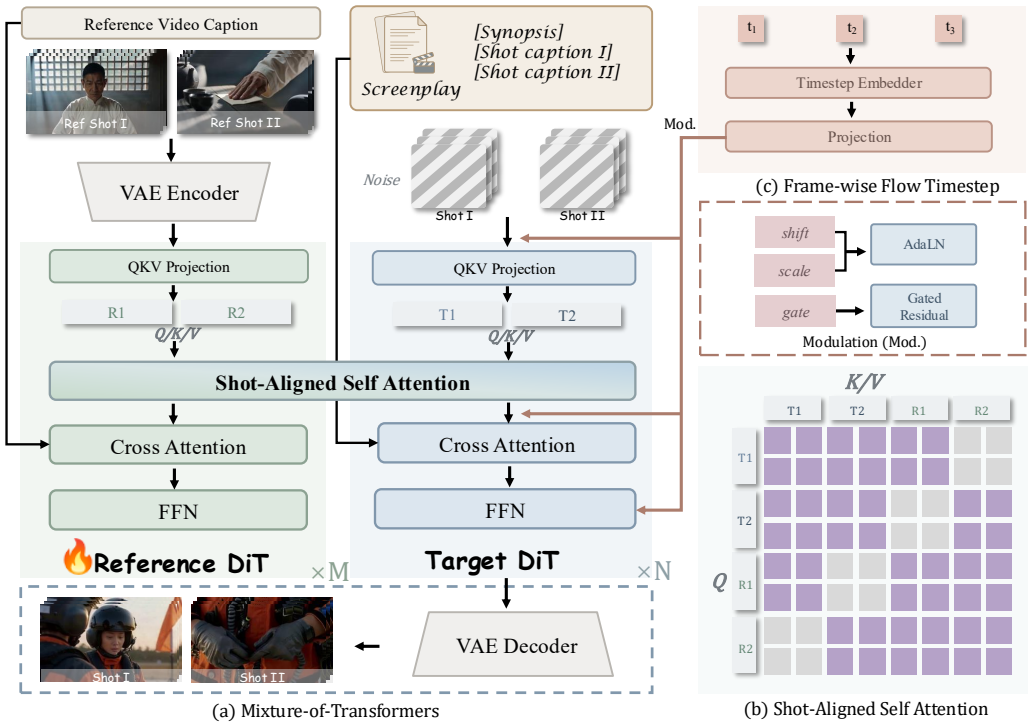

Figure 2: **Overview of CINEASTRA.** The target sequence is processed by the full target DiT with N blocks, while the reference video is processed by a reference DiT initialized from M selected blocks of the target DiT. **(a) MoT** enables in-context interactions between the reference and target streams, allowing the target to draw on reference cinematography and narrative pacing. **(b) Shot-Aligned Self Attention** associates each target shot with its corresponding reference shot while preserving global interactions among target tokens. **(c) Frame-wise Flow Timestep** independently modulate target frames, enabling flexible conditioning for generation, continuation, and completion.

To answer this question, we introduce CINEASTRA, a unified framework for reference-guided multi-shot video creation. To enable reference-guided directorial control, we leverage the in-context learning capabilities of Diffusion Transformers (Peebles & Xie, 2023) through a Mixture-of-Transformers (MoT) architecture. By jointly modeling the reference and target video streams via contextual attention, CINEASTRA infers the reference's cinematography and narrative pacing as visual guidance to direct the target sequence specified by the screenplay. To support flexible creation, we propose frame-wise flow timestep conditioning, which independently represents the generation state of each target frame according to its assigned diffusion timestep. This design naturally accommodates target observations at arbitrary temporal positions, unifying generation from scratch, video continuation, and missing-clip completion within a single model.

Our main contributions are summarized as follows:

- We introduce CINEASTRA, a unified MoT-based framework that captures holistic cinematic language by unifying within-shot cinematography and cross-shot narrative pacing from a reference video through in-context interactions.
- We propose a frame-wise flow timestep conditioning mechanism that independently manages per-frame generation states, enabling flexible multi-shot creation modes including generation from scratch, video continuation, and missing-clip completion.
- Extensive experiments demonstrate that CINEASTRA achieves superior multi-shot coherence, reference adherence, and visual quality compared to existing methods.

2 Related Work

Multi-Shot Video Generation. Multi-shot video generation aims to synthesize distinct shots while maintaining coherent subjects, scenes, and narrative progression. Existing approaches can be broadly divided into staged pipelines and native multi-shot models. Staged pipelines decompose the process into script planning, consistent keyframe generation, and independent shot synthesis (Zheng et al., 2024; Zhang et al., 2026; Zhou et al., 2024; Xiao et al., 2026). Although this formulation naturally accommodates multiple keyframes, its performance is constrained by the quality of intermediate results: errors may propagate between stages, and limited interaction among independently generated shots can weaken cross-shot consistency. Native approaches instead model multiple shots within a single generative process, allowing information to be shared across the entire sequence (Kara et al., 2025; Wu et al., 2026b; Meng et al., 2026; Wang et al., 2026c;a). While such joint modeling generally improves cross-shot coherence, these methods predominantly operate under text-to-video or image-to-video formulations and are not readily applicable to extending or completing partially observed sequences (Cai et al., 2026). Beyond coherence, constructing a compelling cinematic narrative requires expressive control over cinematic language. Existing methods generally provide dedicated controls for particular functions, such as shot arrangement (Kara et al., 2025), transition timing, subject consistency (Wang et al., 2026c), and customized content (Wang et al., 2026a). Although effective for their respective purposes, these designs rely on attribute-specific conditions or components. Extending them toward holistic cinematic control would require coordinating multiple specialized mechanisms, introducing additional training and compatibility costs while increasing the specification burden on users. Consequently, existing methods remain limited in providing a unified means of expressing how camera behavior, framing, shot timing, transitions, and narrative events should work together across a sequence.

Reference-Based Video Generation. Reference-based video generation offers an alternative to combining specialized controllers for individual attributes by using visual examples to communicate desired behavior. Existing approaches primarily focus on motion customization (Zhao et al., 2024; Ling et al., 2024) and camera-centric control (Luo et al., 2025; Zhang et al., 2025). However, these approaches commonly introduce additional trajectory-estimation steps and remain sensitive to scene structure differences between reference and target videos. More fundamentally, existing methods isolate predefined attributes (e.g., motion or camera poses) from the reference rather than leveraging the video signal as a whole, thereby capturing only selected aspects of cinematic language—a limitation that becomes pronounced in multi-shot generation where camera behavior, framing, cut timing, and transitions must be tightly coordinated. Taking a novel perspective, CineAstra treats the reference video as an integrated visual context through in-context learning to capture cinematic language, jointly modeling cinematographic patterns and shot organization across sequences.

3 Methodology

This section presents CineAstra, a unified framework for reference-guided multi-shot creation. We first formalize the problem under arbitrary target observations (Sec. 3.1). Next, we elaborate on the Mixture-of-Transformers (MoT) architecture for in-context cinematic language learning (Sec. 3.2) and frame-wise flow timestep conditioning for flexible temporal generation (Sec. 3.3). The overall framework is illustrated in Fig. 2, with detailed algorithmic procedures provided in Alg. 1.

3.1 Problem Formulation

Reference-guided multi-shot video creation aims to synthesize target video sequences with rich cinematic language, transferring both *cinematography* and *narrative pacing* from a reference video. Specifically, *cinematography* defines visual framing (framing, viewpoint and camera motion), whereas *narrative pacing* controls temporal structure (shot count, shot duration, cut timing, and transition patterns). Let $\mathcal{V}^r = \{\mathcal{V}_i^r\}_{i=1}^N$ and $\mathcal{V}^t = \{\mathcal{V}_i^t\}_{i=1}^N$ denote the reference and target videos, respectively, each consisting of N shots. The target video content is specified by a screenplay $\mathcal{P} = \{p_0, p_1, \ldots, p_N\}$, where p_0 provides the narrative synopsis and p_i describes the detailed content of the i-th target shot.

Beyond expressing such cinematic language from scratch, a practical creation framework must accommodate diverse creation modes such as video continuation and missing-clip completion. To

this end, we incorporate target observations provided at arbitrary temporal positions, denoted by $\mathcal{O} = \{(f, \mathbf{x}_f) \mid f \in \Omega\}$, where Ω is the set of observed frame indices and $\mathbf{x}_f$ represents the observed content at frame f. When generating entirely from scratch, $\Omega = \varnothing$. Otherwise, $\mathcal{O}$ can represent an initial video segment for continuation, isolated keyframes, or sparse clips for completion. The model must therefore integrate these observations according to their temporal positions while synthesizing the remaining content coherently. Formally, we formulate reference-guided multi-shot video creation as learning the conditional distribution $p_\theta(\mathcal{V}^t \mid \mathcal{V}^r, \mathcal{P}, \mathcal{O})$. The resulting sequence should realize the narrative specified by $\mathcal{P}$, preserve the observed target content in $\mathcal{O}$, and faithfully emulate the cinematography and narrative pacing demonstrated by $\mathcal{V}^r$.

3.2 In-Context Learning of Cinematic Language

To capture rich cinematic language, separately extracting camera trajectories and independently specifying shot composition or transitions easily disrupts inter-shot relationships. In-context learning in language modeling offers a promising alternative, demonstrating the ability to capture complex patterns directly from contextual demonstrations without explicit rule decomposition (Brown et al., 2020; Garg et al., 2022; Guo et al., 2025). Encouragingly, this paradigm is beginning to emerge in video generation, where recent video diffusion transformers (DiTs) leverage contextual interactions with reference demonstrations to capture complex visual patterns (Li et al., 2025; Deng et al., 2025). Since the DiT excel at capturing such rich contextual interactions, our key insight is that *a Mixture-of-Transformers (MoT) architecture enables in-context learning across video streams, naturally acquiring cinematic language without piecing together condition-specific controls.*

MoT Architecture. Building on this principle, CineAstra employs an MoT framework to jointly model the reference and target streams within the DiT backbone. Rather than converting the reference into explicit parameters, the model directly infers coordinated cinematography and narrative pacing through contextual interactions to guide target generation. As illustrated in Fig. 2 (a), CineAstra consists of a reference DiT and a target DiT. The reference DiT is initialized from the pretrained multi-shot DiT, with M of its N blocks ($M < N$) selected to form the reference stream for efficiency. The target stream retains the pretrained DiT's multi-shot generation capability, while the reference stream learns to encode cinematic context from the reference video. Specifically, the reference stream processes the reference video alongside its caption, while the target stream synthesizes the target video conditioned on the screenplay $\mathcal{P}$. To organize target generation according to the screenplay, we partition the target tokens $\mathcal{V}^t$ into shot-wise groups $\{\mathcal{V}_i^t\}$ according to the desired shot boundaries. On the target stream, a cross-attention mask allows each target shot $\mathcal{V}_i^t$ to attend to the sequence synopsis p_0 and its corresponding shot description p_i, while p_0 remains visible throughout the target sequence to maintain narrative context. This explicit assignment between temporal token intervals and shot descriptions determines where each shot begins and ends, providing control over shot duration and establishing the temporal structure for narrative pacing.

Shot-Aligned Self Attention. Given this screenplay-defined shot structure, we further establish shot-wise correspondence with the reference video so that each target shot draws cinematic context from its corresponding reference shot (Fig. 2 (b)). The reference and target streams retain their own projections and interact through joint self-attention. At layer l, the projected tokens from the two streams are concatenated as $\mathbf{Q}_l = [\mathbf{Q}_l^t; \mathbf{Q}_l^r]$, $\mathbf{K}_l = [\mathbf{K}_l^t; \mathbf{K}_l^r]$, and $\mathbf{V}_l = [\mathbf{V}_l^t; \mathbf{V}_l^r]$. During training, we use the known reference shot boundaries to construct an asymmetric shot-aligned attention mask $\mathbf{M}$:

$$\mathbf{M}_{qk} = \begin{cases} 0, & q \in \mathcal{V}_i^t,\ k \in \mathcal{V}^t \cup \mathcal{V}_i^r, \\ 0, & q \in \mathcal{V}_i^r,\ k \in \mathcal{V}_i^t \cup \mathcal{V}^r, \\ -\infty, & \text{otherwise.} \end{cases} \tag{1}$$

The resulting joint self-attention is formulated as

$$\mathbf{A}_l = \text{Attention}\left(\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l; \mathbf{M}\right). \tag{2}$$

This design preserves full interactions within each stream while restricting cross-stream interactions to corresponding shots. Queries from the i-th target shot $\mathcal{V}_i^t$ can access all target tokens but retrieve reference guidance only from $\mathcal{V}_i^r$; symmetrically, queries from $\mathcal{V}_i^r$ can access all reference tokens while interacting only with the corresponding target shot $\mathcal{V}_i^t$. In this way, each target shot receives the appropriate reference context without disrupting sequence-level interactions within either stream.

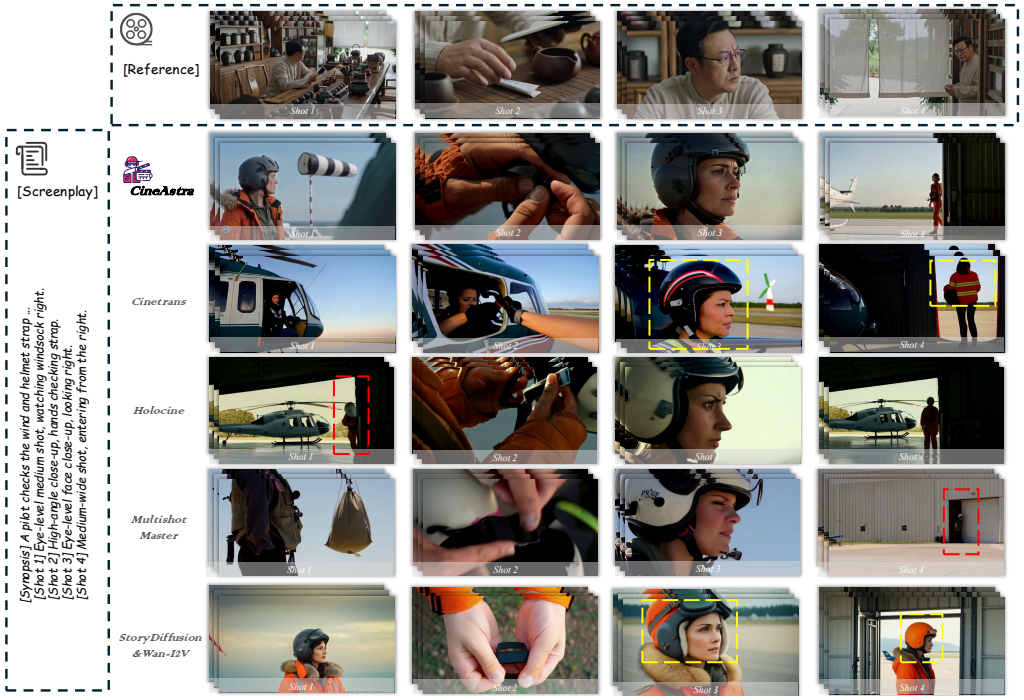

Figure 3: **Qualitative comparison.** Compared to competing methods, CINEASTRA faithfully captures reference cinematography and narrative pacing while preserving overall visual coherence. Yellow boxes highlight cross-shot inconsistencies, while red boxes indicate failure to follow reference cinematic language or screenplay.

3.3 FLEXIBLE CREATION VIA FRAME-WISE FLOW TIMESTEPS

To enable versatile multi-shot creation, the model must flexibly accommodate diverse target observations in $\mathcal{O}$. While staged keyframe-to-video pipelines can accept keyframe conditions, independently synthesizing individual shots compromises cross-shot coherence and produces temporal discontinuities. Conversely, conventional flow models assign a single shared timestep to all sequence frames, making it impossible to explicitly distinguish newly synthesized target shots from observed historical shots or anchor keyframes provided by creators. Recent advances in sequence and video models explore non-uniform, frame-dependent timestep and noise modulations, demonstrating that assigning flexible denoising states across temporal dimensions greatly enhances sequence controllability (Song et al., 2025; Chen et al., 2024; Liu et al., 2026a;b; Xie et al., 2025). Building on this principle, we introduce frame-wise flow timestep conditioning (Fig. 2 (c)) to independently specify the generation state of each frame. Specifically, at denoising step k, the global flow timestep t_k for frame f is modulated by a multiplier $m_f \in [0, 1]$ via $t_{k,f} = m_f t_k$. For an observed latent frame $\mathbf{z}_f^0$, its noisy representation is formulated as:

$$\mathbf{z}_{k,f} = \alpha(t_{k,f})\mathbf{z}_f^0 + \sigma(t_{k,f})\boldsymbol{\epsilon}_f, \qquad \boldsymbol{\epsilon}_f \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

where $\alpha(\cdot)$ and $\sigma(\cdot)$ govern the flow noise schedule. Each $t_{k,f}$ is independently embedded to derive shift, scale, and gate parameters $(\mathbf{b}_{k,f}, \mathbf{s}_{k,f}, \mathbf{g}_{k,f}) = \mathrm{Proj}(\mathrm{TEmbed}(t_{k,f}))$, yielding frame-level AdaLN modulation:

$$\mathrm{AdaLN}_{k,f}(\mathbf{h}_{k,f}) = (1 + \mathbf{s}_{k,f}) \odot \mathrm{LN}(\mathbf{h}_{k,f}) + \mathbf{b}_{k,f}. \tag{4}$$

Through this mechanism, the target DiT directly perceives local generation states from $t_{k,f}$, while the reference stream maintains a fixed timestep t_{ref} for stable context. The multiplier vector $\mathbf{m} = [m_1, \ldots, m_F]$ serves as a unified control interface: unobserved shots undergo full flow denoising ($m_f = 1.0$), whereas observed shots or anchor keyframes remain fully clean ($m_f = 0.0$). By configuring $\mathbf{m}$, a single model seamlessly handles generation from scratch, video continuation, and missing-clip completion while preserving context consistency under reference guidance.

Table 1: **Quantitative evaluation.** We evaluate performance across cinematic language, visual consistency, and aesthetic quality against state-of-the-art baselines (Zhou et al., 2024; Wan et al., 2025; Wu et al., 2026b; Wang et al., 2026c; Meng et al., 2026). (↑: higher is better. Bold/underline indicate best/second-best. ✗: unsupported conditioning.)

Method	Cinematic Language ↑			Visual Consistency ↑			Quality ↑
	Cine. Adhere.	Edit. Align.	Rhythm	Intra-shot	Inter-shot	Observed-ctx	Aes. Score
Staged Keyframe Pipeline							
StoryDiffusion + Wan-I2V	0.631	0.615	0.361	0.775	0.383	<u>0.432</u>	5.15
Holistic Modeling							
CineTrans	0.449	0.159	0.449	0.755	**0.614**	✗	<u>5.24</u>
MultiShotMaster	0.768	0.810	<u>0.875</u>	<u>0.780</u>	0.425	0.391	5.01
HoloCine	<u>0.795</u>	<u>0.873</u>	0.865	**0.794**	0.478	✗	4.99
CINEASTRA (Ours)	**0.844**	**0.889**	**0.908**	**0.794**	<u>0.491</u>	**0.573**	**5.28**

4 EXPERIMENTS

4.1 DATASET AND IMPLEMENTATION

Dataset Construction. Learning reference-guided cinematic control requires paired multi-shot sequences that share underlying cinematic language while depicting distinct narrative content. To this end, we construct a high-quality multi-shot video dataset comprising approximately 50K reference–target pairs, with each sequence consisting of 2 to 5 shots. The dataset covers diverse framing, viewpoints, camera motions, and temporal cut rhythms. To build paired training data, we aggregate real-world multi-shot cinematic clips from feature movies (Wu et al., 2025) and segment them according to shot boundaries. We then leverage a state-of-the-art model, Seedance2.0 (Seedance et al., 2026), to generate corresponding target sequences. In this process, we substitute the original subjects, actions, and environments in the prompt while preserving the underlying shot structure, camera behavior, and transition patterns. Pairs with inconsistent shot boundaries, severe visual artifacts, or identity leakage from the reference video are systematically filtered out. Each video is annotated using Qwen3-VL (Bai et al., 2025b) with a narrative synopsis p_0 and detailed shot-level descriptions p_i for each shot i. Detailed annotation templates are provided in the Appendix A.2.

Training Strategy. CINEASTRA builds upon the Wan2.2 (Wan et al., 2025) architecture as its backbone and adopts a two-stage training strategy using the constructed multi-shot dataset. In Stage I, we adapt the backbone via LoRA to support multi-shot generation under versatile creation modes. During training, we sample observation sets $\mathcal{O}$ by varying target content visibility, allowing either entire shots or arbitrary frames within shots to be observed, alongside unobserved sequences ($\mathcal{O} = \varnothing$). Observed and noisy regions are modulated using frame-wise flow timesteps (Sec. 3.3) under a unified objective. In Stage II, we introduce the reference stream to enable in-context guidance. It is constructed as a depth-distributed partial replica initialized from uniformly spaced layers of the target DiT, resulting in one-quarter of its depth. We optimize only the reference stream while keeping the target stream fixed. More details are provided in Appendix A.3.

4.2 EXPERIMENTAL SETUP

Baselines. We compare CINEASTRA against representative state-of-the-art multi-shot baselines, including holistic modeling (HoloCine (Meng et al., 2026), MultiShotMaster (Wang et al., 2026c), and CineTrans (Wu et al., 2026b)) and a staged keyframe-to-video pipeline combining StoryDiffusion with Wan2.2-I2V (Zhou et al., 2024; Wan et al., 2025). For models supporting target conditioning, we evaluate video continuation given initial clips, assessing both internal inter-shot consistency and observed-context consistency. For purely text-driven baselines, we perform generation from scratch across the entire sequence.

Evaluation. We construct a diverse evaluation set of 100 reference cases, each containing 2–4 shots and annotated with cinematography attributes, editing relations and shot durations. Firstly, we evaluate how well the generated multi-shot videos follow the reference in terms of **Cinematography** and **Narrative Pacing**. We use TransNetV2 (Soucek & Lokoc, 2024) to detect and seg-

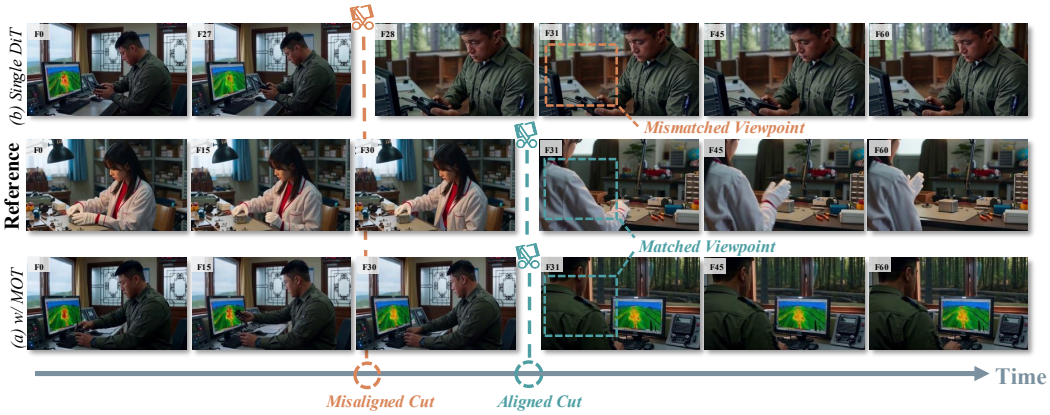

Figure 4: **Ablation of MoT.** While (b) Single DiT exhibits misaligned cuts and mismatched viewpoints, (a) CINEASTRA (w/ MoT) faithfully captures the reference's cinematic language.

Table 2: **Ablation of CINEASTRA components.** The **w/o MoT** variant adapts a single DiT via supervised fine-tuning (SFT), whereas **w/o Frame-wise Timestep** relies on I2V guidance.

Ablation Variant	Cinematic Language ↑			Consistency ↑
	Cine. Adhere.	Edit. Align.	Rhythm	
w/o MoT	0.688	0.508	0.579	0.460
w/o Frame-wise Timestep	0.821	0.714	0.771	0.442
w/o Shot-Aligned Self Attention	0.786	0.634	0.681	0.490
Ours	**0.844**	**0.889**	**0.908**	**0.491**

ment shots in generated videos. For automated reference-following assessment, we employ GPT-6-Astra (OpenAI, 2026) as the evaluator, conditioned on human-annotated exemplars and in-context visual prompts pairing reference videos with their annotations. Accordingly, we report three metrics: *(i) Cinematography Adherence (Cine. Adhere.)* measures whether generated shots reproduce the reference cinematography across 12 fine-grained categories along three axes: *framing* (environmental, subject-level, and local), *view point* (low-angle, eye-level, and high-angle), and *camera motion* (static, approach, recede, horizontal, vertical, and compound). Order-preserving temporal Intersection-over-Union (tIoU) matching aligns generated and reference shot intervals, penalizing missing or superfluous shots. *(ii) Editing Relation Alignment (Edit. Align.)* measures the hit rate of pre-annotated target editing relations across eight canonical types: *cut-in*, *cut-out*, *cutaway*, *reverse shot*, *reaction*, *eyeline*, *match-on-action*, and *angle change*. Transitions without a successfully aligned shot boundary receive zero. *(iii) Narrative Rhythm Consistency (Rhythm)* evaluates temporal pacing as one minus half the L_1 distance between the reference and generated distributions of shot-duration proportions. Secondly, we evaluate **Visual Consistency**. *Intra-shot Consistency* follows VBench (Huang et al., 2024) and is computed within detected shots, while *Inter-shot Consistency* measures the mean ViCLIP (Wang et al., 2024a) similarity across generated shots. To evaluate performance on **Continuation and Completion**, we further report *Observed-context Consistency*, which measures ViCLIP similarity between each generated shot and its annotated corresponding observed clip; methods that do not support target-observation conditioning are marked as unsupported. Finally, we evaluate **Aesthetic Quality** using the LAION (Schuhmann, 2022) aesthetic score. Detailed metric computation procedures are provided in the Appendix A.4.

4.3 QUALITATIVE AND QUANTITATIVE COMPARISONS

Qualitatively (Fig. 3), CINEASTRA consistently captures reference viewpoint and framing while strictly adhering to the target screenplay, whereas baseline methods exhibit noticeable limitations. Specifically, yellow boxes highlight cross-shot inconsistencies across generated frames, while red boxes indicate failures to follow the reference cinematic language or screenplay text descriptions. See Fig. 5 and Fig. 8 for more qualitative visual results.

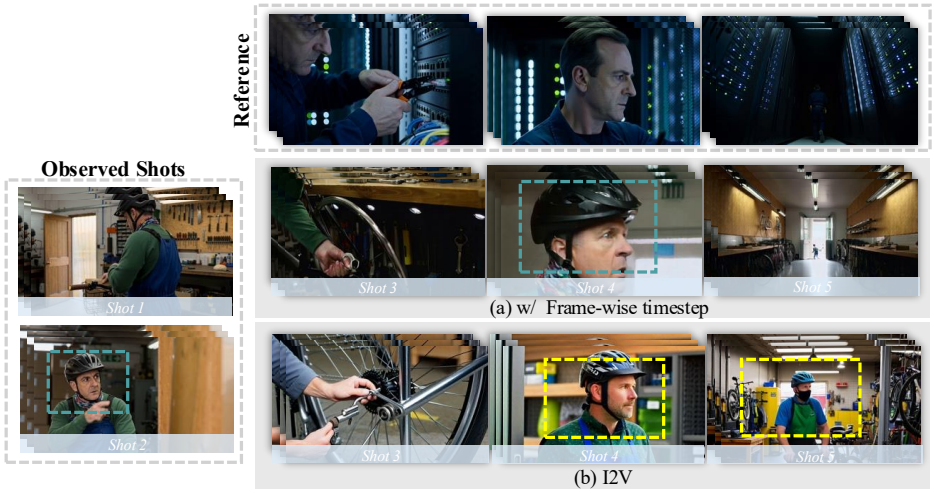

Figure 5: **Ablation of frame-wise flow timesteps.** Given observed historical shots, we perform reference-guided multi-shot video continuation. Blue boxes highlight cross-shot consistency, while yellow boxes indicate inconsistencies.

Quantitatively (Tab. 1), CINEASTRA achieves superior performance across cinematography adherence, editing relation alignment, and narrative rhythm. Regarding visual consistency, while CineTrans yields a high inter-shot score, this artifact stems from its weak cinematic language understanding, which tends to repetitively generate static content across shots. In contrast, CINEASTRA achieves strong inter-shot coherence across shots with distinct framings and viewpoints without compromising reference cinematography. Furthermore, staged keyframe-to-video baselines suffer from lower inter-shot consistency due to error accumulation across independently synthesized shot keyframes. Finally, CINEASTRA substantially outperforms competing models in observed-context consistency, demonstrating the efficacy of frame-wise timesteps for target-conditioned creation.

4.4 ABLATION STUDY

MoT. Qualitatively (Fig. 4), a single DiT under SFT struggles to capture the reference's cinematic language, producing mismatched viewpoints and slight timing offsets at cut frames. Without MoT, the single-stream backbone fails to reproduce the demonstrated cinematography, whereas incorporating the dedicated reference stream enables accurate reference following across framing and viewpoint. Quantitatively (Tab. 2), we exploit the redundancy of DiT blocks (Wang et al., 2026b; Bian et al., 2026) by constructing the reference DiT from uniformly spaced layers of the target backbone, using only $1/4$ of its depth. Replacing MoT with a single-stream shared DiT leads to noticeable performance degradation. More architectural and parameter details are provided in Appendix A.1.

Frame-wise Flow Timestep. To evaluate the variant without frame-wise timesteps, we assess video continuation with observed content by selecting an observed frame to generate the sequence via image-to-video. Quantitatively (Tab. 2), relying solely on I2V leads to a noticeable performance degradation. Qualitatively (Fig. 5), while the I2V baseline exhibits clear inconsistencies with the observed footage, our frame-wise flow timestep modulation allows arbitrary initial footage to be flexibly inserted, seamlessly guiding video continuation and synthesis without visual discontinuities while adhering to the reference cinematic language.

Shot-Aligned Self Attention. As shown in Tab. 2, applying the shot-aligned mask $\mathbf{M}$ eliminates interference from non-corresponding reference shots, markedly boosting both cinematography adherence and editing relation alignment. This confirms that restricting target shots to attend strictly to their corresponding reference counterparts enforces precise shot-to-shot directorial alignment.

5 CONCLUSION

In this paper, we presented CINEASTRA, a unified framework for reference-guided multi-shot video creation. By combining a MoT architecture with frame-wise flow timestep conditioning, CINEASTRA captures holistic cinematic language in context and flexibly handles versatile creation modes. Extensive evaluations confirm its state-of-the-art performance in cinematic adherence, narrative pacing, and cross-shot coherence.

AI USE DISCLOSURE

In this work, we used generative AI tools for tasks with required disclosure, specifically to generate synthetic data sets, clean and reformat dataset, and assist with translation. We have not used generative AI tools for implementing methods, developing theoretical conceptual frameworks, or formulating mathematical claims. All AI-assisted artifacts and text were thoroughly verified and manually checked for correctness by the authors. The authors take full responsibility for the final content, claims, and artifacts produced in this work.

REFERENCES

Zhaochong An, Menglin Jia, Haonan Qiu, Zijian Zhou, Xiaoke Huang, Zhiheng Liu, Weiming Ren, Kumara Kahatapitiya, Ding Liu, Sen He, et al. Onestory: Coherent multi-shot video generation with adaptive memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16173–16184, 2026.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025a.

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*, 2025b.

Yuxuan Bian, Xin Chen, Zenan Li, Tiancheng Zhi, Shen Sang, Linjie Luo, and Qiang Xu. Video-as-prompt: Unified semantic control for video generation. In *International Conference on Learning Representations*, volume 2026, pp. 114558–114583, 2026.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Minghong Cai, Qiulin Wang, Zongli Ye, Wenze Liu, Quande Liu, Weicai Ye, Xintao Wang, Pengfei Wan, Kun Gai, and Xiangyu Yue. Videocanvas: Unified video completion from arbitrary spatiotemporal patches via in-context conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4475–4485, 2026.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. *arXiv preprint arXiv:2305.13840*, 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Teng Hu, Jiangning Zhang, Ran Yi, Yating Wang, Hongrui Huang, Jieyu Weng, Yabiao Wang, and Lizhuang Ma. Motionmaster: Training-free camera motion transfer for video generation, 2024.

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: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6689–6700. IEEE, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M Rehg, and Tobias Hinz. Shotadapter: Text-to-multi-shot video generation with diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28405–28415. IEEE, 2025.

Baolu Li, Yiming Zhang, Qinghe Wang, Liqian Ma, Xiaoyu Shi, Xintao Wang, Pengfei Wan, Zhenfei Yin, Yunzhi Zhuge, Huchuan Lu, et al. Vfxmaster: Unlocking dynamic visual effect generation via in-context learning. *arXiv preprint arXiv:2510.25772*, 2025.

Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026a.

Yaofang Liu, Yumeng Ren, Aitor Artola, Yuxuan Hu, Xiaodong Cun, Xiaotong Zhao, Alan Zhao, Raymond Chan, Suiyun Zhang, Rui Liu, et al. Pusa v1. 0: Unlocking temporal control in pretrained video diffusion models via vectorized timestep adaptation. In *International Conference on Learning Representations*, volume 2026, pp. 16191–16217, 2026b.

Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videostudio: Generating consistent-content and multi-scene videos. In *European Conference on Computer Vision*, pp. 468–485. Springer, 2024.

Yawen Luo, Xiaoyu Shi, Jianhong Bai, Menghan Xia, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Camclonemaster: Enabling reference-based camera control for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–10, 2025.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Shuailei Ma, Yixuan Li, Cheng Chen, et al. Holocine: Holistic generation of cinematic multi-shot long video narratives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 461–471, 2026.

OpenAI. GPT-6 Astra. https://developers.openai.com/api/docs/models/gpt-6-astra, 2026. Accessed: 2026-09-26.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Tianhao Qi, Jianlong Yuan, Wanquan Feng, Shancheng Fang, Jiawei Liu, SiYu Zhou, Qian He, Hongtao Xie, and Yongdong Zhang. Maskˆ 2dit: Dual mask-based diffusion transformer for multi-scene long video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18837–18846, 2025.

Christoph Schuhmann. Improved aesthetic predictor. GitHub repository, 2022. URL https://github.com/christophschuhmann/improved-aesthetic-predictor.

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.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.

Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11218–11221, 2024.

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.

Jiahao Wang, Hualian Sheng, Sijia Cai, Weizhan Zhang, Caixia Yan, Yachuang Feng, Bing Deng, and Jieping Ye. Echoshot: Multi-shot portrait video generation. *Advances in Neural Information Processing Systems*, 38:22058–22090, 2026a.

Jiangshan Wang, Zeqiang Lai, Jiarui Chen, Jiayi Guo, Hang Guo, Xiu Li, Xiangyu Yue, and Chunchao Guo. Elastic diffusion transformer. *arXiv preprint arXiv:2602.13993*, 2026b.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16268–16278, 2026c.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2024, pp. 42055–42079, 2024a.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b.

Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. Moviebench: A hierarchical movie level dataset for long video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28984–28994. IEEE, 2025.

Xiaoxue Wu, Xinyuan Chen, Yaohui Wang, and Yu Qiao. Shotdirector: Directorially controllable multi-shot video generation with cinematographic transitions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2079–2089, 2026a.

Xiaoxue Wu, Bingjie Gao, Yu Qiao, Yaohui Wang, and Xinyuan Chen. Cinetrans: Learning to generate videos with cinematic transitions via masked diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 60198–60230, 2026b.

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 *International Conference on Learning Representations*, volume 2026, pp. 83349–83375, 2026.

Desai Xie, Zhan Xu, Yicong Hong, Hao Tan, Difan Liu, Feng Liu, Arie Kaufman, and Yang Zhou. Progressive autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6312–6322. IEEE, 2025.

David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2050–2062. IEEE, 2025.

Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. Stage: Storyboard-anchored generation for cinematic multi-shot narrative. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 659–669, 2026.

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 *International Conference on Learning Representations*, volume 2025, pp. 50060–50090, 2025.

Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 273–290. Springer, 2024.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, et al. Videogen-of-thought: Step-by-step generating multi-shot video with minimal manual intervention. *arXiv preprint arXiv:2412.02259*, 2024.

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

A APPENDIX

Overview. In this appendix, we provide supplementary materials to support the main manuscript, structured as follows: Sec. A.1 details algorithmic procedures and reference stream architecture; Sec. A.2 presents dataset construction details and directorial data example visualizations; Sec. A.3 outlines training setups, progressive stages, and compute resource allocation; Sec. A.4 formulates quantitative evaluation metrics and details the GPT-6-Astra evaluator implementation; Sec. A.5 describes the user study setup and visual preference findings.

A.1 ALGORITHMIC & REFERENCE STREAM ARCHITECTURE

Algorithm 1 outlines the unified training and inference procedures of CINEASTRA, detailing the frame-wise timestep adaptation, MoT reference stream optimization, and reference-guided generation steps.

Algorithm 1 CINEASTRA Training and Inference Procedure

Require: Pretrained Multi-Shot DiT backbone Θ_0; Reference video $\mathcal{V}^r = \{\mathcal{V}_i^r\}_{i=1}^N$ with text description c^r; Screenplay $\mathcal{P} = \{p_0, p_1, \ldots, p_N\}$; Target observation mask Ω and values $\mathbf{x}_f$.
Ensure: Optimized CINEASTRA parameters $\Theta_{\text{target}}, \Theta_{\text{ref}}$; Synthesized multi-shot target video $\mathcal{V}^t = \{\mathcal{V}_i^t\}_{i=1}^N$.
// Stage I: Frame-Wise Timestep Adaptation via LoRA
Initialize target backbone with LoRA weights: $\Theta_{\text{target}} \leftarrow \Theta_0 + \Delta\Theta_{\text{LoRA}}$.
for each training iteration in Stage I **do**
  Sample target video $\mathcal{V}^t$ and screenplay $\mathcal{P}$.
  Sample observation set $\mathcal{O} = \{(f, \mathbf{x}_f) \mid f \in \Omega\}$ by varying visibility (including $\Omega = \varnothing$).
  Construct frame-wise multiplier vector $\mathbf{m} = [m_1, m_2, \ldots, m_F]$, where $m_f = 0.0$ for $f \in \Omega$ else 1.0.
  Sample global flow timestep $t \sim \mathcal{U}(0, 1)$ and compute local noise levels: $t_f = m_f \cdot t, \forall f$.
  Apply frame-wise flow noising: $\mathbf{z}_{t,f} = \alpha(t_f)\mathbf{z}_f^0 + \sigma(t_f)\boldsymbol{\epsilon}_f$, $\boldsymbol{\epsilon}_f \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
  Compute frame-level AdaLN modulations: $(\mathbf{b}_f, \mathbf{s}_f, \mathbf{g}_f) = \text{Proj}(\text{TEmbed}(t_f))$.
  Update Θ_{target} by minimizing flow matching loss on unobserved frames ($m_f > 0$).
end for
// Stage II: MoT Reference Stream Optimization with Shot-Aligned Mask
Construct Reference Stream Θ_{ref} by sampling M uniformly spaced layers from Θ_{target}.
Freeze target stream parameters Θ_{target}.
for each training iteration in Stage II **do**
  Sample aligned pair $(\mathcal{V}^r, \mathcal{V}^t)$ and screenplay $\mathcal{P}$.
  Construct asymmetric shot-aligned attention mask $\mathbf{M}_{qk}$:
  $\mathbf{M}_{qk} = 0$ if $(q \in \mathcal{V}_i^t, k \in \mathcal{V}^t \cup \mathcal{V}_i^r)$ or $(q \in \mathcal{V}_i^r, k \in \mathcal{V}_i^t \cup \mathcal{V}^r)$; else $-\infty$.
  Forward joint self-attention across streams: $\mathbf{A}_l = \text{Attention}([\mathbf{Q}_l^t; \mathbf{Q}_l^r], [\mathbf{K}_l^t; \mathbf{K}_l^r], [\mathbf{V}_l^t; \mathbf{V}_l^r]; \mathbf{M})$.
  Update reference stream parameters Θ_{ref} to learn in-context cinematic language.
end for
// Inference: Reference-Guided Versatile Multi-Shot Creation
function CINEASTRAINFERENCE($\mathcal{V}^r, \mathcal{P}, \mathcal{O}, \mathbf{m}$)
  Set fixed reference timestep t_{ref} for $\mathcal{V}^r$.
  for $k = K, \ldots, 1$ **down to** 1 **do**
    Compute frame-wise noise schedule: $t_{k,f} = m_f \cdot t_k, \forall f \in [1, F]$.
    Modulate target latent features using $\text{AdaLN}_{k,f}(\mathbf{h}_{k,f}) = (1 + \mathbf{s}_{k,f}) \odot \text{LN}(\mathbf{h}_{k,f}) + \mathbf{b}_{k,f}$.
    Perform joint MoT attention under shot-aligned mask $\mathbf{M}$ to predict velocity fields.
    Update target noise latents $\mathbf{z}_{k-1}$ via flow matching step (keep $m_f = 0.0$ frames clean).
  end for
  return Synthesized multi-shot target video $\mathcal{V}^t$.
end function

Reference Stream Architectural Details. To inject reference context into the generative trajectory without interfering with the target stream's generation priors, the reference stream is constructed as a depth-distributed partial replica of the target DiT backbone. Specifically, we sample $M = 10$ uniformly spaced layers (indices $\{0, 4, 8, \ldots, 36\}$ out of $N = 40$ backbone blocks) to form sparse interaction nodes across shallow, intermediate, and deep layers. In each selected layer, the clean reference video is encoded via VAE into latent tokens $\mathbf{z}^r$, which remain clean and unnoised throughout the denoising process. To drive feature modulation within the reference stream, we assign a dedicated fixed noise level t_{ref} derived from the final timestep of the flow scheduler. At these interaction

Director Template: Low-Angle Eyeline Reveal

Global Narrative Synopsis (p_0): A quiet tea house setting. The tea master notices the kettle whistling has stopped, picks up a folded note, and walks out toward the small courtyard.

Shot 1: Establishing	Shot 2: Eyeline Cut	Shot 3: Reaction	Shot 4: Wide Reveal
Framing: Subject-level	**Framing:** Local	**Framing:** Local	**Framing:** Environmental
Viewpoint: Low-angle	**Viewpoint:** Eye-level	**Viewpoint:** Eye-level	**Viewpoint:** High-angle
Motion: Static	**Motion:** Approach	**Motion:** Static	**Motion:** Recede
Function: Establishes character & tea room spatial relation.	**Function:** Cuts to kettle area, advancing action via eyeline.	**Function:** Focuses on character reaction & picking up note.	**Function:** Reveals courtyard spatial context & action result.

Reference Stream ($\mathcal{V}^r$):	Provides $N = 4$ shot transitions with low-angle eyeline reveal camera trajectories, narrative pacing, and canonical editing relations (*eyeline*, *reaction*, *angle change*).
Target Stream ($\mathcal{V}^t$):	Guided by Target Screenplay $\mathcal{P} = \{p_0, p_1, p_2, p_3, p_4\}$, synthesizing novel scenes while faithfully adhering to $\mathcal{V}^r$'s 3-axis cinematography and canonical editing relations.

Figure 6: **Data Example Visualization for Low-Angle Eyeline Reveal.** Illustration of our structured 4-shot directorial creation example aligned with the 3-axis cinematography specification (*framing*, *viewpoint*, *camera motion*) and 8 canonical editing relations (*e.g.*, *cut-in*, *cut-out*, *cut-away*, *reverse shot*, *reaction*, *eyeline*, *match-on-action*, *angle change*). Both reference and target streams share the $N = 4$ shot structure, transferring directorial style while adhering to the target screenplay.

nodes, the target and reference streams independently process their AdaLN modulations, text cross-attention, and FFNs, while temporarily concatenating their queries, keys, and values $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ to perform joint self-attention under the shot-aligned mask $\mathbf{M}$. Through this sparse, multi-level interaction, the reference stream effectively transfers holistic cinematic patterns, providing unified in-context guidance for target video synthesis.

A.2 Dataset Construction & Data Examples

Our dataset consists of ∼50k high-quality paired reference–target multi-shot sequences, where each video clip is resampled to 480×832 resolution at 16 fps.

Director Data Example Visualization. Fig. 6 illustrates a representative data creation example from our dataset. It demonstrates how our 4-shot directorial structure maps $N = 4$ shot transitions, 3-axis cinematographic specifications (*framing*, *viewpoint*, *camera motion*), and canonical editing relations (*e.g.*, *eyeline*, *reaction*, *cut-in*) between reference and target streams.

A.3 Training Details & Compute Resources

We initialize CineAstra from a 14B pretrained multi-shot Diffusion Transformer backbone (Wan et al., 2025; Meng et al., 2026). The MoT reference stream comprises $M = 10$ sparse interaction blocks (representing ∼3.5B parameters) sampled from the backbone, while the target stream remains frozen to preserve stable multi-shot generation priors.

During training, the reference video is encoded into clean latents without noise injection, receiving a fixed near-zero noise level timestep solely to generate its AdaLN temporal modulation signals, whereas the target stream receives varying noise timesteps according to its denoising schedule.

The model is trained in two progressive stages (high-noise and low-noise refinement) over ∼6k total optimization steps. Training is conducted on a distributed cluster of 64 NVIDIA H100 GPUs (80GB VRAM) with a global batch size of 32 and gradient accumulation of 4. We use the AdamW optimizer with a constant learning rate of 1×10^{-5} and a weight decay of 1×10^{-4}. Loss is

computed exclusively on the denoising predictions of unobserved target frames, leaving the clean reference video unconstrained.

A.4 Evaluation Metrics & Evaluator

To quantitatively benchmark cinematic language adherence, we establish three evaluation metrics covering cinematographic attribute alignment, editing relation preservation, and narrative pacing:

Cinematography Adherence (*Cine. Adhere.*). This metric evaluates shot-level cinematographic fidelity across three primary axes: framing, camera elevation, and camera motion. For each case with K reference shots and N generated shots, we map corresponding shots based on temporal IoU maximization over normalized video timelines. Let C denote the number of matched shots where the dominant cinematographic label matches the reference. To penalize shot-count mismatches, $\mathrm{Score}_{\mathrm{case}}$ incorporates a quadratic length penalty:

$$\mathrm{Score}_{\mathrm{case}} = \frac{C}{\max(K,N)} \times \frac{\min(K,N)}{\max(K,N)}. \tag{5}$$

The overall dataset score is computed via reference shot-weighted averaging: $\mathrm{Cine.\,Adhere.} = \frac{\sum_i K_i \cdot \mathrm{Score}_i}{\sum_i K_i}$.

Editing Alignment (*Edit. Align.*). Editing Alignment quantifies the preservation rate of pre-annotated cross-shot transition relations, such as *cut-in*, *cut-out*, *reaction*, *eyeline*, *match-on-action*, and *angle change*:

$$\mathrm{Edit.\,Align.} = \frac{\text{Number of Correctly Preserved Target Relations}}{\text{Total Number of Pre-Annotated Target Relations}}. \tag{6}$$

A strict shot-count gating protocol is applied: if the number of generated shots $N \neq K$, the case receives a score of 0 while remaining in the evaluation denominator.

Narrative Rhythm (*Rhythm*). Narrative rhythm assesses shot duration pacing across the sequence. Let p_j and q_j represent the normalized relative duration proportions ($\sum p_j = \sum q_j = 1$) of the j-th reference and generated shots, respectively. The rhythm similarity is formulated as 1 minus the Total Variation Distance:

$$\mathrm{Rhythm}_{\mathrm{case}} = 1 - \frac{1}{2} \sum_{j=1}^{\min(K,N)} |q_j - p_j|. \tag{7}$$

A strict shot-count gating rule assigns a score of 0 if $N \neq K$. The final reported Rhythm score is averaged across all test cases and normalized to $[0, 1]$.

A.4.1 Implementation of GPT-6-Astra Shot Attribute Evaluator.

To objectively evaluate shot-level cinematographic attributes without human bias, we employ GPT-6-Astra (OpenAI, 2026) as an automated judge. Conditioned on human-annotated exemplars and in-context visual prompts, GPT-6-Astra predicts discrete class labels alongside step-by-step visual rationales.

Evaluated Axes & Taxonomy. For each detected shot, the evaluator predicts dominant labels across three orthogonal cinematographic dimensions (12 fine-grained categories in total): *(i) Framing*: *environmental*, *subject-level*, and *local*; *(ii) Viewpoint*: *low-angle*, *eye-level*, and *high-angle*; *(iii) Camera Motion*: *static*, *approach*, *recede*, *horizontal*, *vertical*, and *compound* (plus *unknown* for unclassifiable cases).

Temporal Sampling & Zero-Shot Evaluation Protocol. For each extracted shot, we sample frame triplets at 20%, 50%, and 80% temporal progression marks to capture intra-shot visual dynamics. To eliminate confirmation bias, GPT-6-Astra operates strictly under a zero-shot evaluation regime, directly inspecting the sampled frame sequences without receiving generation prompts, ground-truth labels, or pairwise comparisons.

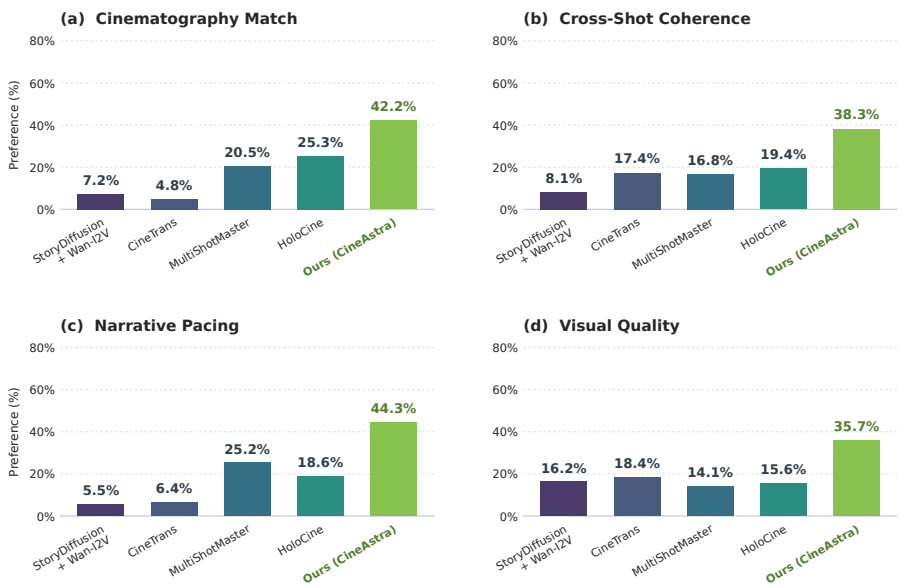

Figure 7: **User study preference results.**

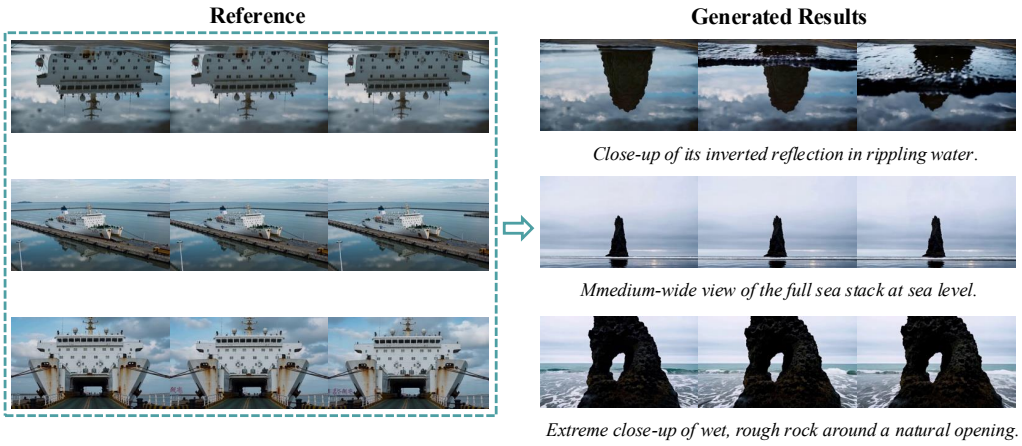

Figure 8: **Qualitative frame sequences.**

Detailed Rubric & Edge-Case Guidelines. To resolve ambiguous visual boundary conditions, the evaluator is conditioned on a detailed rubric specifying edge-case disambiguation rules:

- *Perspective vs. Posture*: A seated character viewed from a slightly elevated position is classified as *eye-level* unless explicit vanishing-point perspective cues indicate a true top-down tilt (*high-angle*).
- *Workplane Close-ups*: Tabletop close-ups displaying hand operations with a visible downward working surface are categorized as *high-angle*, whereas near-horizontal side views of tools remain *eye-level*.
- *Non-Human Scenes*: Empty environments or isolated object/hand interactions are assigned *N/A* for subject-level attributes to prevent false orientation predictions.
- *Foreground Occlusion*: Foreground furniture partially masking a character does not alter framing to a tight cut; the dominant overall spatial composition governs classification.

A.5 USER STUDY

To evaluate human visual preference, we conducted a user study with 30 computer vision and generative AI researchers and creators. Participants evaluated 20 randomly sampled creation cases across four criteria: (1) *Cinematography Match*, (2) *Cross-Shot Coherence*, (3) *Narrative Pacing*, and (4) *Visual Quality*. We performed pairwise AB testing, where synthesized videos were presented in randomized order. As shown in Fig. 7, CINEASTRA consistently outperforms baseline models, achieving a clear preference in *Cinematography Match* and *Narrative Pacing*. Furthermore, we present a qualitative frame sequence example in Fig. 8 to visually illustrate our superior generation quality, shot-to-shot consistency, and precise adherence to complex cinematic language.