

# CineOrchestra: Unified Entity-Centric Conditioning for Cinematic Video Generation

Sharath Girish<sup>1,\*</sup> Tsai-Shien Chen<sup>1,2,\*</sup> Zhikang Dong<sup>1</sup> Mukesh Singhal<sup>2</sup>

Hao Chen<sup>1</sup> Sergey Tulyakov<sup>1</sup> Aliaksandr Siarohin<sup>1</sup>

<sup>1</sup>Snap Inc. <sup>2</sup>UC Merced

Project page: [snap-research.github.io/CineOrchestra/](https://snap-research.github.io/CineOrchestra/)

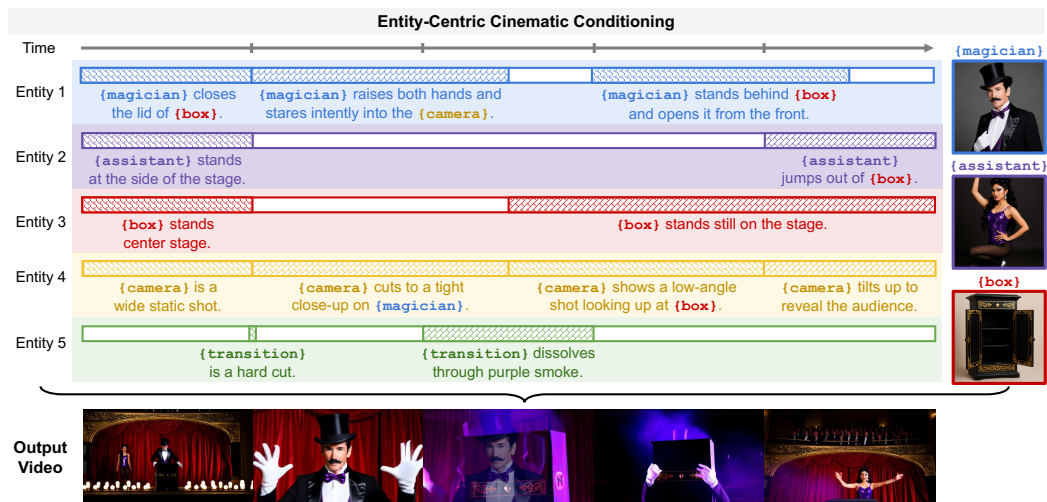


Figure 1: *CineOrchestra* generates cinematic scenes from unified conditioning. Our entity-centric conditioning represents every cinematic element (*visual subjects*, *camera*, and *shot transitions*) as a unified, timestamped expression (top), with optional reference images (top right). It enables *CineOrchestra* to generate cinematic frames in a single forward pass, jointly realizing multi-subject personalization, multi-event timing, multi-shot composition, and camera control (bottom).

## Abstract

Cinematic video depicts multiple subjects acting or interacting at specific moments, captured with deliberate camera movement, and stitched together by shot transitions. Together, these elements demand a level of fine-grained control beyond current text-to-video models. Existing work addresses each axis in isolation: multi-subject personalization, temporal control, multi-shot synthesis, or camera control; no prior framework jointly integrates all four. We present *CineOrchestra*, a unified video diffusion model that controls subjects, events, cameras, and shot transitions simultaneously. Our key insight is that these heterogeneous cinematic elements share a fundamental structure: each is an entity acting over a specific temporal interval, which can therefore all be expressed through one shared structure of entity-centric conditioning primitives, augmented with reference images for visual entities.

\*Equal contribution.

This formulation reduces the architectural challenge to a single positional encoding problem, which we solve with two parameter-free coordinated rotary embeddings: (i) an interval-sampled temporal RoPE that yields consistent attention behavior across events of dramatically varying duration, and (ii) a 2D entity-temporal cross-attention RoPE that disambiguates per-entity conditions and routes each to its corresponding spatiotemporal region. On two new benchmarks, *CineOrchestra* outperforms six per-axis specialists on dense caption following and shot-transition timing, with consistent gains in a pairwise user study and component ablations.

## 1 Introduction

A scene from a movie is rarely a *single static shot* of a *single subject* performing a *single action*. Instead, several characters and objects co-exist, each acting or interacting with others at specific moments, captured by a camera that moves with intent, and stitched together by transitions between distinct shots. This makes cinematic video generation an inherently compositional problem that remains out of reach for current text-to-video models [6, 61, 35, 72, 21, 46, 24], which typically condition on a single global caption and produce a single shot.

Recent work has begun to decompose this monolithic conditioning, with each line targeting one axis of cinematic composition: (i) multi-reference personalization models [12, 38, 17, 15] compose co-existing entities; (ii) temporal control methods [71, 19] manipulate when each event happens; (iii) multi-shot synthesis methods [70, 62, 65, 42] produce several connected shots; and (iv) camera-conditioning models [25, 68] allow user-specified viewpoint motion. Each line, however, uses bespoke architectures trained on disjoint data; no existing model jointly ingests *subjects*, *events*, *shot transitions*, and *camera* and produces a coherent cinematic video scene.

We present *CineOrchestra*, a cinematic video generation framework that controls all four axes within a single model. Our key insight is that these heterogeneous elements share a common structure: each is an entity acting over a temporal interval. We therefore propose an *entity-centric cinematic conditioning* that expresses every cinematic element as a structured tuple of (start\_time, end\_time, prompt), as illustrated in Fig. 1. Each character and object receives a unique tag (e.g., {man}, {car}), a global identity description, a set of event-level dense descriptions (e.g., “[0.1s–2.3s] {man} jumps into {car}”), and an optional reference image to provide identity details. Crucially, we extend the same structure to cinematography through two special tags, {camera} and {transition}, which carry only event-level descriptions (e.g., “[6.3s–6.4s] {transition} shows a hard cut”; “[0.0s–10.0s] {camera} pans left across {car}”). The same representation thus captures the full cinematic structure with no modality-specific design.

This unification shifts the entire technical challenge onto a single problem: positional encoding [60, 58] in our video DiT backbone [45, 43]. A single clip may carry many reference images and dozens of overlapping events whose durations span a dramatically wide range from 0.1s hard cuts to 10s sustained camera moves, which standard fixed-cadence temporal RoPE [71] cannot fairly represent. We therefore introduce two coordinated RoPE designs: a *interval-sampled temporal RoPE* that evenly samples position within each event’s interval and rescales them for a duration-invariant similarity peak, and a *2D entity-temporal cross-attention RoPE* that disambiguates entity tokens and routes each conditioning element to its corresponding spatiotemporal region.

On two newly proposed benchmarks, *CineBench* and *CineBenchSyn*, *CineOrchestra* outperforms six per-axis specialists on dense caption following and shot-transition timing, confirmed by a pairwise user study and ablations of each component. Our contributions can be summarized as follows:

- We present *CineOrchestra*, the first framework for joint cinematic video generation across subjects, events, camera, and shot transitions, built on a unified entity-centric conditioning primitive.
- We introduce two parameter-free coordinated RoPEs that handle variable-duration events, disambiguate per-entity conditions, and route each to its target spatiotemporal region.
- We introduce two benchmarks, *CineBench* and *CineBenchSyn*, on which *CineOrchestra* outperforms six per-axis specialists in both automatic metrics and a pairwise user study.

## 2 Related Work

Generating cinematic video scenes requires simultaneous control over character identity, narrative timing, shot structure, and camera movement, where prior work addresses each axis in isolation.

**Video diffusion models.** Diffusion models [56, 57, 27, 51, 29] have driven tremendous progress in text-to-video and image-to-video generation through training on Internet-scale data [26, 4, 10, 55, 35, 11]. While earlier approaches adopt U-Net [52] as the denoising backbone [5, 22, 8, 9], recent models have switched to diffusion transformers [45] (DiTs) whose scalability better handles high-resolution, long, and visually complex videos [6, 61, 72, 23, 46, 21, 24]. *CineOrchestra* builds on a video DiT [43] but further extends it to enable cinematic compositional control.

**Multi-subject personalization.** Subject personalization began with optimization-based methods [53, 18, 36] and matured into feed-forward reference-image conditioning [73, 64, 54, 14, 47]. Recent work extends the idea to multiple subjects [12, 38, 30, 15, 17, 74, 16, 13], typically by injecting reference-image tokens through attention operations [60] to render several identities in one coherent video. While these methods ground multiple identities, they treat the clip as a single global event with one caption and cannot express the time-localized scripts that cinematic video scenes require.

**Multi-event temporal control.** Videos add a temporal dimension over frames, motivating work that decomposes a video into time-localized segments. *Multi-shot* methods [70, 62, 65, 42] synthesize several connected shots separated by hard cuts, but each per-shot caption is monolithic, leaving no dense intra-shot timing. *Continuous-frame* methods [71, 19] address this by generating a single video from dense, time-stamped captions with a temporal RoPE that biases attention toward each caption’s annotated interval, and additionally accept a separate scene-cut conditioning track for shot transitions. However, this track encodes only cut *timing*, with no natural-language description of the transition (e.g., dissolve, wipe, fade). In contrast, *CineOrchestra* unifies *both* under one entity-centric primitive: dense captions handle intra-shot timing, while a {transition} tag carries a dense description of inter-shot transitions and composes with other entity tags, all in one forward pass.

**Cinematography conditioning.** Camera motion is a primary expressive tool in filmmaking, making viewpoint control an important axis of video generation. Existing methods condition generation on explicit geometric signals, such as Plücker-embedded camera trajectories [25, 2, 76, 1], joint camera-and-object pose sequences [68], or novel-view re-rendering of a source video [3]. However, such inputs demands either specialized capture rigs, 3D reconstruction pipelines, or manual pose authoring. *CineOrchestra* instead conditions camera behavior through time-stamped natural-language descriptors (e.g., “[1.5s–6.7s] {camera} pushes in slowly”; “[6.7s–9.0s] {camera} pans left to reveal {man}”), retaining directorial control without requiring any explicit pose inputs.

## 3 CineOrchestra

*CineOrchestra* reformulates cinematic video generation as conditioning a single video diffusion transformer [45, 43] (DiT) on a unified, entity-centric description of cinematic structure. Sec. 3.1 presents the overall architecture, which inherits the standard self-attention then cross-attention layout in recent multi-reference video diffusion models [12, 38, 19, 17, 15]. Our key technical novelty lies in the conditioning primitive (Sec. 3.2) and the rotary positional embeddings (Secs. 3.3 to 3.5). Lastly, Sec. 3.6 describes the data curation pipeline that supplies entity-centric annotations at scale.

### 3.1 Architecture Overview

As illustrated in Fig. 2, we build on a video DiT [43] that operates on spatiotemporally-patchified video tokens  $\mathbf{V}$  produced by a video variational autoencoder [34] (VAE). To support multi-reference image conditioning, each input image is encoded by the same VAE as a single-frame video, yielding per-entity tokens  $\{\mathbf{I}_k\}_{k=1}^K$  at the same spatial resolution as  $\mathbf{V}$ , where  $K$  is the number of entities in the clip. Following Video Alchemist [12], we apply random augmentations to each reference image before encoding to mitigate copy-paste artifacts. We then append these tokens to the video sequence and apply full self-attention over the concatenated stream  $[\mathbf{V}; \mathbf{I}_1; \dots; \mathbf{I}_K]$ .

To support text conditioning, each text prompt (including a global identity caption per entity and a dense caption per event) is independently encoded with T5 [49] and concatenated into a single

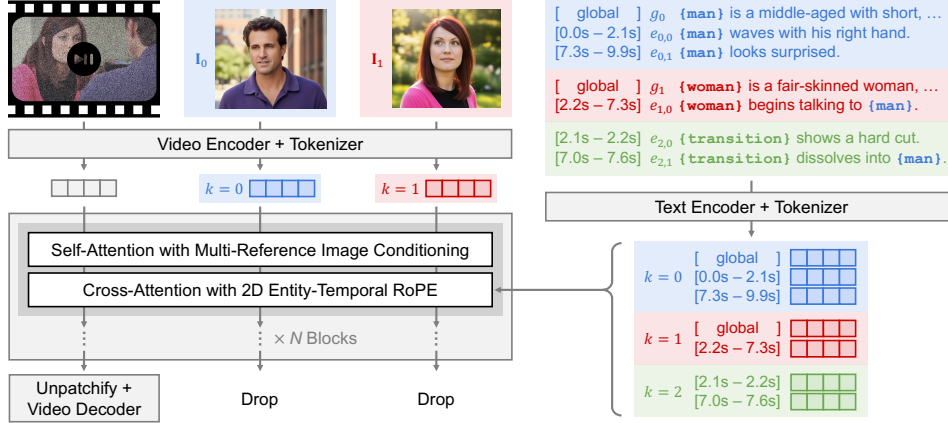


Figure 2: **Overview of CineOrchestra.** Each entity  $k$  is represented by a reference image  $\mathbf{I}_k$ , a global description  $g_k$ , and a set of event-level dense descriptions  $\{(t_{k,j}^s, t_{k,j}^e, e_{k,j})\}$  pairing temporal intervals with prompts. Reference image tokens are concatenated to the video tokens for full self-attention, while all text tokens are consumed via cross-attention. Two coordinated RoPEs, interval-sampled temporal RoPE (Sec. 3.3) and 2D entity-temporal cross-attention RoPE (Sec. 3.5), disambiguate per-entity conditions and route each to its target spatiotemporal region.

key-value bank. Each transformer block then applies cross-attention where  $[\mathbf{V}; \mathbf{I}_1; \dots; \mathbf{I}_K]$  serves as queries and this bank serves as keys and values.

Notably, our framework adds no learnable parameters to the underlying video DiT: every component reuses an existing module, and our two coordinated RoPEs in Secs. 3.3 and 3.5 are parameter-free modifications to positional encoding. App. B.2 completes the architectural specifications.

### 3.2 Entity-Centric Cinematic Conditioning

We represent every cinematic element through a single entity-grounded primitive. Each entity  $k \in \{1, \dots, K\}$  is associated with the following fields:

- A unique tag  $\tau_k$  that serves as a stable referent across all captions (e.g.,  $\{\text{man\_old}\}$ ,  $\{\text{car\_red}\}$ ).
- An optional reference image  $\mathbf{I}_k$  that faithfully encodes visual identity.
- An optional global identity description  $g_k$  (e.g., “ $\{\text{man\_old}\}$  is an 80-year-old man with short white hair wearing a black coat”).
- A set of event-level dense descriptions  $\mathcal{E}_k = \{\dots, (t_{k,j}^s, t_{k,j}^e, e_{k,j}), \dots\}$ , where each tuple pairs a temporal interval with a prompt (e.g., “[0.1s–1.3s]  $\{\text{man\_old}\}$  raises his hand”; “[1.3s–2.9s]  $\{\text{man\_old}\}$  then waves his hand”). The first token in the prompt identifies its primary entity.

Our conditioning primitive features two advantages. First, movie-level scenes routinely involve complex interactions between entities. Since every  $\tau_k$  is a stable referent, dense descriptions can express such interactions unambiguously by including multiple tags (e.g., “[2.9s–4.2s]  $\{\text{man\_old}\}$  opens the door of  $\{\text{car\_red}\}$ ”).

Second, in addition to visual subjects, cinematographic elements such as camera movement and shot transitions are also essential to a cinematic scene. Our primitive can be naturally extended to these via two reserved tags,  $\{\text{camera}\}$  and  $\{\text{transition}\}$ , which carry only event-level dense descriptions. More importantly, they are directly composable with visual subjects by referencing their tags (e.g., “[0.0s–7.8s]  $\{\text{camera}\}$  pans left across  $\{\text{car\_red}\}$ ”). Furthermore, unlike prior multi-shot methods [70, 62, 65, 42] that implicitly assume instantaneous cuts, our dense descriptions enable diverse transition types (e.g., “[5.7s–6.2s]  $\{\text{transition}\}$  fades to black”).

### 3.3 Interval-Sampled Temporal RoPE

Durations of cinematic events  $L$  could span a dramatically wide range: a hard cut may last 0.1s while a sustained pan or character action may last 10s. Prior multi-event temporal control methods [71]

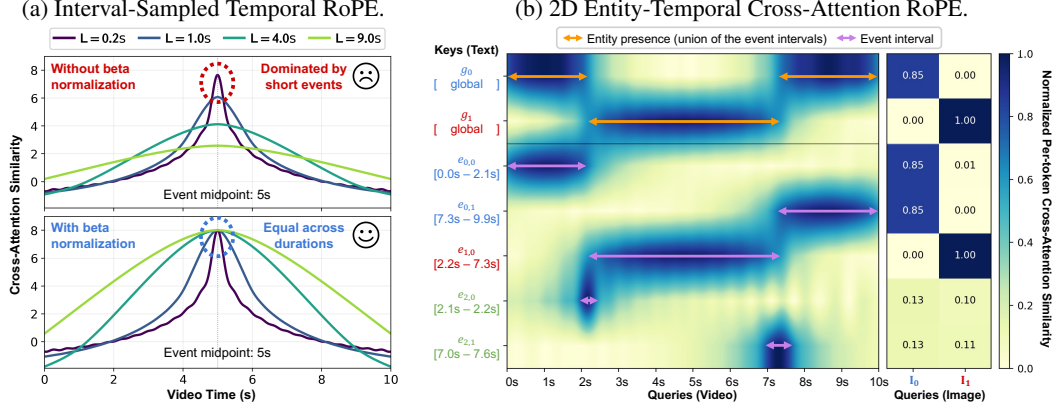


Figure 3: **Two coordinated RoPE designs.** (a) Similarity between a video token and an event token across event durations  $L$ . Our  $\beta(L)$  rescaling produces duration-invariant peaks. (b) Cross-attention similarity between video/image queries and global/dense-description keys under Tab. 1’s coordinates. Sharp peaks emerge only where query and key share the same entity and overlap in time, jointly achieving entity *disambiguation* and temporal *routing*.

encode each event with a *fixed-cadence* temporal RoPE at a constant temporal interval  $\Delta t$  (e.g., per frame or per second), which suffers two failure modes:

- Sub-cadence events falling below the encoding resolution ( $L < \Delta t$ ) become indistinguishable.
- The attention similarity against video tokens depends on  $L$ , biasing attention toward shorter events.

We address both issues through the proposed *interval-sampled temporal RoPE*. Formally, let  $R(t) \in \mathbb{R}^{d_{\text{rope}} \times d_{\text{rope}}}$  denote the standard RoPE rotation [58] at temporal coordinate  $t$ , where  $d_{\text{rope}}$  is the per-head channel dimension. For an event during  $[t_s, t_e]$ , we sample  $N = 16$  positions evenly *within* the interval and define its temporal positional encoding as the rescaled average:

$$\mathbf{P}^{\text{event}}(t_s, t_e) = \beta(L) \cdot \frac{1}{N} \sum_{i=0}^{N-1} R\left(t_s + \frac{i}{N-1}(t_e - t_s)\right), \quad (1)$$

where  $\beta(L)$  is a duration-dependent scalar specified below. By placing  $N$  positions *within* the interval rather than sampling at a fixed cadence  $\Delta t$ ,  $\mathbf{P}^{\text{event}}$  resolves the sub-cadence failure mode and jointly captures the event’s start, end, and duration in a single positional encoding.

**Similarity-peak normalization.** Setting  $\beta \equiv 1$ , however, leaves the second failure mode. The dot product between an event token and a video token undergoes *phase cancellation* across the  $N$  samples, with magnitude decaying as  $\text{sinc}(\theta_n L/2)$  for each RoPE frequency  $\theta_n$ . Consequently, wider intervals receive smaller-magnitude peaks than narrower ones (see Fig. 3(a) for visualization and App. A for the full expression). To address this, we introduce a duration-dependent scalar

$$\beta(L) = \sqrt{d_{\text{rope}}} \left/ \left\| \frac{1}{N} \sum_{i=0}^{N-1} R(t_i) \right\|_F, \quad (2)$$

so that the Frobenius norm  $\|\mathbf{P}^{\text{event}}(t_s, t_e)\|_F$  is duration-invariant. App. A derives the closed form of  $\beta(L)$  and establishes three key properties: it (i) reduces to standard RoPE at  $L = 0$ , (ii) remains well-bounded for all event durations, and (iii) yields an approximately duration-invariant peak similarity. Fig. 3(a) visualize its effect by comparing the cross-attention similarity (between a video token and an event token) without and with normalization. Sec. 4.2 further validates the design empirically.

### 3.4 Self-Attention: Multi-Reference Image Conditioning

Self-attention operates over the concatenated stream  $[\mathbf{V}; \mathbf{I}_1; \dots; \mathbf{I}_K]$ , making video and image tokens share the same positional space. In such a case, applying vanilla 3D RoPE could place each  $\mathbf{I}_k$  at coordinates that collide with arbitrary regions of  $\mathbf{V}$ , creating ambiguity between video and image tokens. We address this with separate spatial and temporal RoPE designs for image tokens.

Table 1: **Coordinates of 2D entity-temporal RoPE per cross-attention token type.**  $\mathbf{P}^{\text{event}}$  and  $\mathbf{P}^{\text{image}}$  are defined in Eqs. (1) and (3), respectively. The entity-index axis *disambiguates* per-entity conditions, while the temporal axis *routes* each to its target spatiotemporal region.

Token type	Entity-index axis	Temporal axis
Video token $\mathbf{V}$ (query)	$\frac{1}{K} \sum_{k=1}^K R_{\text{entity}}(k)$	3D-RoPE temporal coordinate
Reference image $\mathbf{I}_k$ (query)	$R_{\text{entity}}(k)$	$\mathbf{P}^{\text{image}}(k)$
Global description $g_k$ (key)	$R_{\text{entity}}(k)$	$\mathbf{P}^{\text{image}}(k)$
Dense description $e_{k,j}$ (key)	$R_{\text{entity}}(k)$	$\mathbf{P}^{\text{event}}(t_s^{k,j}, t_e^{k,j})$

**Spatial RoPE for image tokens.** Following Qwen-Image [69], we place each reference image on its own diagonal block of the spatial RoPE plane: video occupies  $[0, H) \times [0, W)$ , while image  $k$  occupies  $[kH, (k+1)H) \times [kW, (k+1)W)$ , disjoint from the video and from every other image.

**Temporal RoPE for image tokens.** Visual identity is meaningful only when an entity is on screen. For instance, the reference image of {man\_old} should influence only the frames in which he appears. We therefore anchor each reference image to the temporal extent of its events. That is,  $\mathbf{I}_k$  inherits a temporal position equal to the mean of its event-level temporal RoPEs

$$\mathbf{P}^{\text{image}}(k) = \frac{1}{|\mathcal{E}_k|} \sum_{(t_s, t_e, e) \in \mathcal{E}_k} \mathbf{P}^{\text{event}}(t_s, t_e), \quad (3)$$

which biases self-attention to route identity into the correct temporal region of the video.

### 3.5 Cross-Attention: 2D Entity-Temporal RoPE

Cross-attention injects text conditioning into the video and image queries. A single clip may present  $K$  reference images,  $K$  global descriptions, and  $\sum_{k=1}^K |\mathcal{E}_k|$  event-level dense descriptions spanning visual subjects, camera, and shot transition. With so many parallel and heterogeneous conditioning, the positional encoding must satisfy two requirements: (i) *disambiguation*: image and text tokens for different entities or events must remain positionally distinct; (ii) *routing*: each conditioning element must bias attention toward the entity and frames it describes.

**2D entity-temporal RoPE.** We meet *both* by introducing a 2D RoPE with (i) an *entity-index axis* that separates entities to enforce disambiguation and (ii) a *temporal axis* that aligns each conditioning element with its target frames for routing. Specifically, the entity-index axis is encoded by a standard RoPE rotation  $R_{\text{entity}}(k) \in \mathbb{R}^{d_{\text{entity}} \times d_{\text{entity}}}$  at integer index  $k$ , with one slot per entity. The temporal axis is encoded by the interval-sampled temporal RoPE of Sec. 3.3. Following the standard multi-axis RoPE construction [58], the per-head channel dimension is partitioned into two disjoint groups, with each axis’s rotation applied to its own group, so similarity peaks only when both axes are aligned.

**Per-token coordinates.** Given the two axes, Tab. 1 specifies the coordinates per token type:

- Video tokens  $\mathbf{V}$  (query) are averaged over all  $R_{\text{entity}}(k)$  on the entity-index axis, making them equally receptive to the conditioning of every entity present in the video.
- Reference image  $\mathbf{I}_k$  (query) and global description  $g_k$  (key) share the same coordinates since both encode entity-level rather than event-specific information. We anchor them at the entity-averaged temporal position  $\mathbf{P}^{\text{image}}(k)$  to spread their influence softly whenever entity  $k$  is on screen.
- Dense description  $e_{k,j}$  (key) instead use the precise  $\mathbf{P}^{\text{event}}(e_{k,j})$ , producing a sharp similarity peak within their event interval and thereby satisfying the routing requirement.

This assignment delivers both *disambiguation* and *routing* without any auxiliary mask. The same coordinates apply uniformly to visual subjects, camera, and shot transition, so the unified primitive of Sec. 3.2 is learned end-to-end with no entity-specific modules and no added parameters. Fig. 3(b) illustrates the 2D layout, and Sec. 4.2 ablates each axis.

### 3.6 Training Data Curation

We construct a dataset of one-minute chunks cropped from licensed movies and TV shows. Within such a window, the cast and setting stay consistent, forming a closed entity set for reliable annotation.

Table 2: **Quantitative comparison of cinematic conditioning on *CineBench*.** We compare *CineOrchestra* against specialists: (i) multi-reference personalization (top) and (ii) multi-shot synthesis (middle). We report Masked DINO for subject identity consistency, ViCLIP for dense caption following, and Qwen-VL recall for shot-transition timing. Best in **bold**, second-best underlined.

Method	Subject ID	Global Caption	Dense Caption Following (ViCLIP $\uparrow$ )				Transition Timing
	M-DINO $\uparrow$	M-CLIP $\uparrow$	Subject	Scene	Camera	Transition	Recall $\uparrow$
Phantom [38]	<b>0.509</b>	<u>0.295</u>	0.212	0.197	<u>0.178</u>	<u>0.146</u>	<u>0.431</u>
VACE [31]	0.482	<b>0.309</b>	<u>0.219</u>	<b>0.214</b>	0.171	0.136	0.340
CineTrans [70]	0.423	0.286	0.214	<u>0.208</u>	0.168	0.124	0.129
EchoShot [62]	0.399	0.286	0.202	0.193	0.164	0.129	0.094
MultiShotMaster [65]	0.383	0.286	0.207	0.207	0.172	0.126	0.343
ShotStream [42]	0.391	0.289	0.214	0.197	0.172	0.141	0.267
<i>CineOrchestra</i> (ours)	<u>0.502</u>	<u>0.295</u>	<b>0.235</b>	<u>0.208</u>	<b>0.193</b>	<b>0.150</b>	<b>0.486</b>

**Entity-centric annotation.** For each chunk, we issue a single structured-output query to Gemini 2.5 [20] that populates every field of the primitive in Sec. 3.2 and jointly extracts visual entities, camera, and shot transitions in one pass.

**Reference image with appearance augmentation.** For each visual entity, we crop the entity from event middle frames using Gemini 3 [20] bounding boxes, filter them by CLIP-T [48] similarity to the entity’s global description, and randomly sample two of the top matches. We then pass these crops to Qwen-Image-Edit [69] to synthesize a new view of the same identity under a different pose, lighting, and facial expression. Unlike the photometric and geometric augmentations of Video Alchemist [12], this step varies appearance while preserving identity, mitigating copy-paste artifacts during training.

## 4 Experiments

We validate that *CineOrchestra* jointly controls all four cinematic axes (*subjects, events, camera, shot transitions*) in a single model. Sec. 4.1 compares against per-axis specialized baselines. Sec. 4.2 ablates the proposed components of Sec. 3. App. B provides additional training and inference details.

### 4.1 Comparisons across Four Cinematic Axes

**Baselines.** Since no prior framework jointly handles all four axes, we compare against per-axis specialists from the two axes whose baselines accept entity-centric reference inputs: (i) *multi-reference personalization* from Phantom [38] and VACE [31], and (ii) *multi-shot synthesis* from CineTrans [70], EchoShot [62], MultiShotMaster [65], and ShotStream [42]. Each baseline uses its official checkpoint, adapted to our entity-centric inputs as detailed in App. C.2.

**Benchmark datasets.** We introduce two complementary benchmarks as no existing benchmark covers all four cinematic axes: (i) *CineBench* (512 clips, 3.2k entities, 6.9k events, 1.5k reference images) contains movie and TV clips from titles unseen during training, annotated with our entity-centric primitive. (ii) *CineBenchSyn* (512 clips, 3.3k entities, 6.4k events, 1.7k reference images) includes LLM-generated prompts and Qwen-Image-generated reference images targeting under-represented edge cases. Dataset statistics are provided in App. B and the curation of *CineBenchSyn* in App. C.1. We will release *CineBenchSyn* for future research on cinematic video generation.

**Evaluation metrics.** We report seven metrics to comprehensively evaluate all four axes (see the full definitions of each metric in App. C.3):

- *Subject identity consistency*: following Video Alchemist [12], we report Masked DINO [7] similarity (M-DINO) between each reference image and the masked subject in on-screen frames.
- *Global caption following*: we report Masked CLIP [48] similarity (M-CLIP) between each per-entity global description and the masked subject.
- *Dense caption following*: we report ViCLIP [66, 67] similarity between each dense caption and the video frames inside its annotated interval. It is reported separately for four categories: subject (masked region), scene (entire frame), camera, and shot transition.

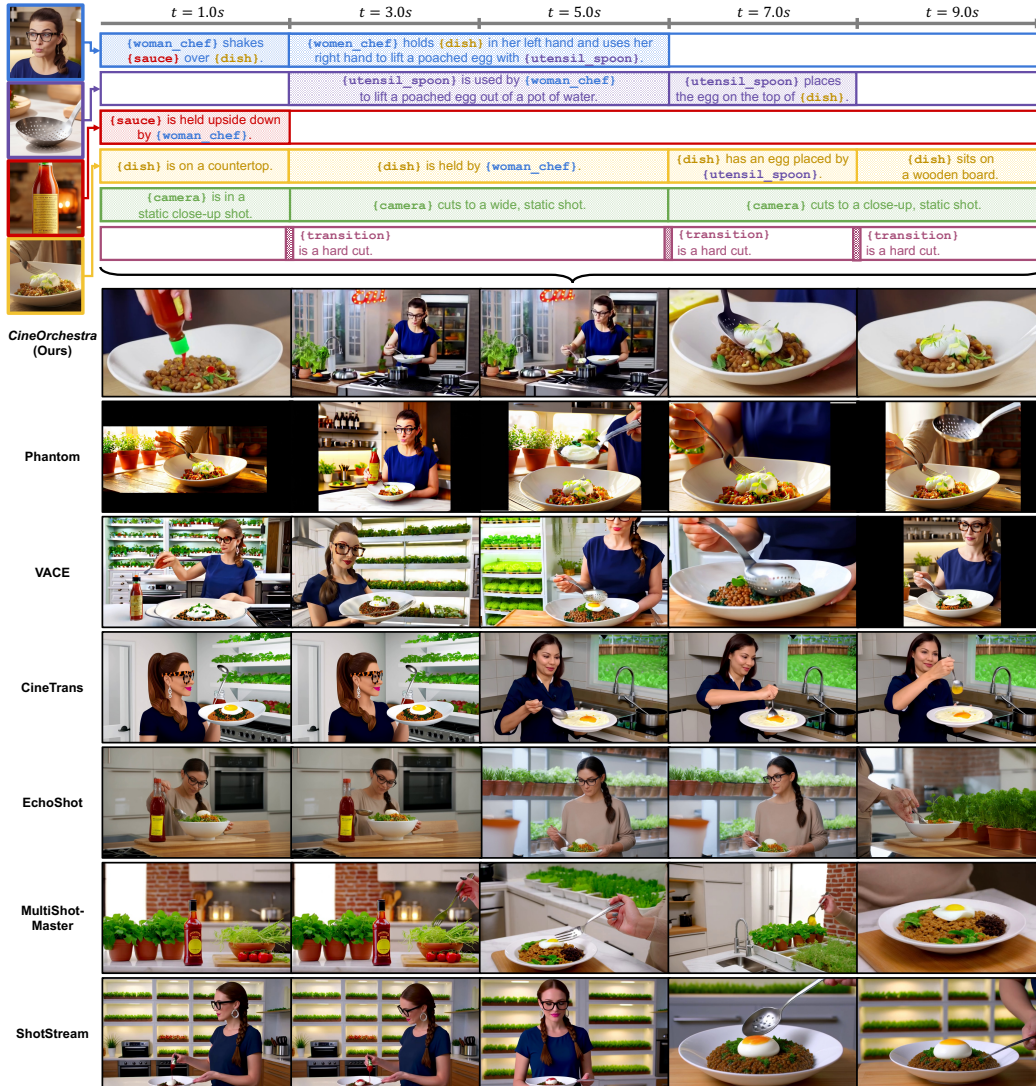


Figure 4: **Qualitative comparison of cinematic conditioning on CineBench.** Given the entity-centric conditioning (top), *CineOrchestra* (top video row) simultaneously preserves all four subject identities, follows the dense per-entity timeline, and lands three hard cuts, outperforming all existing methods. More comparisons on *CineBenchSyn* can be found in App. D.

- *Shot-transition timing*: we report a recall rate, where a Qwen2.5-VL-7B-Instruct [63] judge evaluates whether each ground-truth transition occurs within a tolerance window.

**Quantitative comparison.** Tab. 2 reports the comparison on *CineBench*, and shows a clear split between single-axis specialization and joint cinematic control. The strongest multi-subject personalization baselines remain competitive on subject identity and global appearance (*i.e.*, Phantom [38] and VACE [31] perform the best on M-DINO and M-CLIP, respectively), but these advantages do not carry over to dense caption following. The multi-shot baselines also do not dominate transition timing and remain weaker on identity preservation. In contrast, *CineOrchestra* is strongest on the axes that require routing each condition to a specific entity and temporal interval, which is precisely the regime targeted by the unified entity-centric primitive. App. D shows consistent trends on *CineBenchSyn*.

**Qualitative comparison.** The qualitative comparisons in Fig. 4 mirrors the pattern in the quantitative study. Personalization baselines [38, 31] better preserve the identity but collapse the timeline into a single shot, while multi-shot baselines [70, 62, 65, 42] place cuts but drift on identity and dish

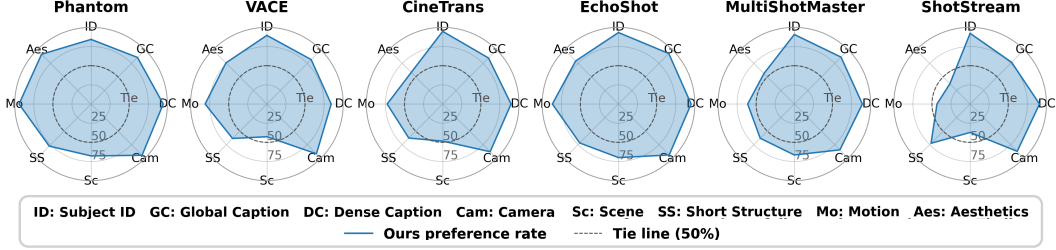


Figure 5: **User study on CineBench.** Pairwise preference of *CineOrchestra* against six baselines on eight dimensions (one radar panel per baseline). Each axis reports  $\text{pref} = W/(W + L)$ , the share of *decisive* votes that favoured *CineOrchestra*, where  $W$  counts wins,  $L$  counts losses, and ties are excluded. The dashed circle marks the 50% tie line: points outside favour *CineOrchestra*, points inside favour the baseline. *CineOrchestra* wins on every entity-, text-, and structure-related axis against all six baselines, while perceptual axes (*motion*, *overall quality*, *scene*) tie or favour the strongest perceptual baselines.

Table 3: **Ablation of two coordinated RoPEs on CineBench.** From top to bottom, (a) applies AlcheMinT’s 3-point WeRoPE [19]; (b) replaces WeRoPE with our  $N=16$  interval-sampled temporal RoPE; (c) adds the entity-index axis to image refs and global caption tokens; (d) further restricts the entity axis on global caption to its visible-time intervals; (e) extends the entity axis to video tokens at visible entity rows; (f) promotes dense-event tokens into entity slots without duration-dependent rescaling. Best in **bold**.

Variant	Subject	Global Cap.	Dense Cap. Following (ViCLIP↑)			Trans.	
	M-DINO↑	M-CLIP↑	Subject	Scene	Camera	Trans.	Recall↑
<i>Interval-Sampled Temporal RoPE</i>							
(a) WeRoPE [19] (1 pos and 2 neg), temp-axis on $\mathbf{V}$	0.455	0.285	<b>0.236</b>	0.206	0.192	0.145	0.399
(b) Ours interval-sampled ( $N=16$ ), temp-axis on $\mathbf{V}$	0.477	0.289	0.234	0.208	0.191	0.147	0.406
<i>2D Entity-Temporal RoPE</i>							
(c) + entity-axis on $\mathbf{I}_k$ and $g_k$	0.485	0.287	0.234	0.207	0.191	0.147	0.422
(d) + temp-axis on $\mathbf{I}_k$ and $g_k$	0.484	0.288	0.233	0.206	0.190	0.148	0.394
(e) + entity-axis on $\mathbf{V}$	0.489	0.289	0.234	0.207	0.190	0.145	0.416
<i>Similarity-Peak Normalization</i>							
(f) Full 2D entity-temp and no rescaling ( $\beta(L) \equiv 1$ )	0.477	0.289	0.235	0.207	0.190	0.147	0.412
<b>Full method</b>	<b>0.502</b>	<b>0.295</b>	0.235	<b>0.208</b>	<b>0.193</b>	<b>0.150</b>	<b>0.486</b>

continuity across them. *CineOrchestra* is the only method that simultaneously preserves all four identities and hits three hard cuts. App. D provides more additional comparisons on *CineBenchSyn*.

**User study.** Since automatic metrics are coarse on perceptual properties, we complement them with a pairwise human evaluation on the same *CineBench*, where raters score *CineOrchestra* against each baseline on eight dimensions (see App. C.4 for the full protocol). Fig. 5 shows that raters prefer *CineOrchestra* on every entity-, text-, and structure-related dimension across all six baselines, and on perceptual dimensions (*motion*, *overall quality*, *scene*) against most baselines.

## 4.2 Ablation Study

Tab. 3 and Fig. 6 ablates the two core RoPE designs of *CineOrchestra* on *CineBench*.

- From (a) to (b), replacing WeRoPE [19] with interval-sampled temporal RoPE improves identity consistency and transition recall, while leaving dense-caption following largely stable.
- From (c) to (e), adding the entity-temporal RoPE gives additional gains, although the partial variants are not uniformly better across all metrics, suggesting that entity disambiguation alone is insufficient without the complete coordinate design.
- Comparing (f) with the full method, removing the  $\beta(L)$  rescaling weakens identity consistency and transition timing, supporting the role of  $\beta(L)$  for events with different temporal spans.



Figure 6: **Visual ablation of two coordinated RoPEs.** Only the full method (top video row) routes each entity to its annotated interval and lands all four hard cuts at the specified times.

- Fig. 6 further verifies that only the full method (top video row) routes every entity to its annotated interval and lands all four hard cuts at the prescribed timestamps.

## 5 Conclusion

We have presented *CineOrchestra*, the first video diffusion framework to jointly control four cinematic axes within a single model: *subjects*, *events*, *camera*, and *shot transitions*. At its core is an entity-centric conditioning primitive that represents every cinematic element as a unified, structured expression. To realize this primitive inside a video DiT, we introduce two coordinated rotary embeddings that handle duration-varying events and route each per-entity condition to its target spatiotemporal region. On *CineBench* and *CineBenchSyn*, *CineOrchestra* outperforms six per-axis specialists on dense caption following and shot-transition timing, with consistent gains in a pairwise user study and ablations of the two coordinated RoPEs. By unifying disparate cinematic controls, *CineOrchestra* reconciles fine-grained controllability with long-horizon narrative coherence, opening the door to movie-level video generation directed by full script rather than single prompt.

## 6 Limitations and Societal Impacts

We identify the following limitations of our work:

**Coarse-grained cinematographic control.** *CineOrchestra* conditions camera motion and shot transitions through natural-language descriptors, which trade fine-grained geometric precision for accessibility. Applications demanding exact viewpoint repeatability or downstream 3D reconstruction are better served by trajectory-based methods [25, 1, 76, 68] or source-video re-rendering [3]. We leave combining our entity-centric prompts with explicit pose conditioning to future work.

**Bounded clip length.** *CineOrchestra* generates an entire scene in a single forward pass, which delivers strong cross-shot coherence but inherits the context-length limits of the underlying DiT. Multi-shot pipelines [70, 62, 65, 42] reach longer durations by synthesizing shots independently and stitching them post hoc, but suffer from weaker subject and lighting continuity. We see extending our primitive to long-form generation via autoregressive models with shared entity tokens as a next step.

**No audio modality.** *CineOrchestra* is purely visual, yet cinematic experience is inseparable from dialogue and ambient sound. Integrating an audio branch that respects the same entity-temporal conditioning is a promising future direction.

**Impact statement.** *CineOrchestra* lowers the barrier to producing cinematic video by letting users direct subjects, events, camera moves, and shot transitions through natural-language scripts and reference images. This can benefit independent filmmakers, educators, and accessibility-driven content creation. Faster previsualization may also reduce waste in physical productions.

At the same time, controllable identity-preserving video generation carries real risks of misuse shared with the broader class of generative video models, including non-consensual portrayal of real people, fabrication of harmful scenarios involving minors, and large-scale visual disinformation.

We recommend that any deployed system combine (i) consent verification and watermarking, (ii) reference-image policies blocking public figures and known minors, (iii) prompt-level filters for sexual, violent, or politically deceptive content, and (iv) gated weight release with abuse-investigation logging.

## References

- [1] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *CVPR*, 2025. 3, 10
- [2] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. VD3d: Taming large video diffusion transformers for 3d camera control. In *ICLR*, 2025. 3
- [3] 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 *ICCV*, 2025. 3, 10
- [4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendeleevitch, 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. 3
- [5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 3
- [6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. Technical report, OpenAI, 2024. 2, 3
- [7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 7
- [8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 3
- [9] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, 2024. 3
- [10] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023. 3
- [11] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, 2024. 3
- [12] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. In *CVPR*, 2025. 2, 3, 7
- [13] Tsai-Shien Chen, Aliaksandr Siarohin, Guocheng Gordon Qian, Kuan-Chieh Jackson Wang, Egor Nemchinov, Moayed Haji-Ali, Riza Alp Guler, Willi Menapace, Ivan Skorokhodov, Anil Kag, et al. Omni-attribute: Open-vocabulary attribute encoder for visual concept personalization. In *CVPR*, 2026. 3
- [14] Yusuf Dalva, Guocheng Gordon Qian, Maya Goldenberg, Tsai-Shien Chen, Kfir Aberman, Sergey Tulyakov, Pinar Yanardag, and Kuan-Chieh Jackson Wang. Canvas-to-image: Compositional image generation with multimodal controls. *ACM TOG*, 2026. 3
- [15] Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. MAGREF: Masked guidance for any-reference video generation with subject disentanglement. In *ICLR*, 2026. 2, 3
- [16] Yuwei Fang, Willi Menapace, Aliaksandr Siarohin, Tsai-Shien Chen, Kuan-Chieh Wang, Ivan Skorokhodov, Graham Neubig, and Sergey Tulyakov. VIMI: Grounding video generation through multi-modal instruction. In *EMNLP*, 2024. 3
- [17] Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025. 2, 3
- [18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 3
- [19] Sharath Girish, Viacheslav Ivanov, Tsai-Shien Chen, Hao Chen, Aliaksandr Siarohin, and Sergey Tulyakov. Alchemint: Fine-grained temporal control for multi-reference consistent video generation. In *CVPR*, 2026. 2, 3, 9
- [20] Google. Gemini. <https://aistudio.google.com/models/gemini-2-5-flash-image>, 2025. 7, 17, 18
- [21] Google. Veo 3. <https://deepmind.google/models/veo/>, 2025. 2, 3
- [22] 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. In *ICLR*, 2024. 3
- [23] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *ECCV*, 2024. 3

- [24] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. [2](#), [3](#)
- [25] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *ICLR*, 2025. [2](#), [3](#), [10](#)
- [26] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. [3](#)
- [27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. [3](#)
- [28] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2021. [19](#)
- [29] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. [3](#)
- [30] Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025. [3](#)
- [31] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *ICCV*, 2025. [7](#), [8](#), [20](#), [22](#), [23](#)
- [32] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *CVPR*, 2024. [19](#)
- [33] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. [17](#)
- [34] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. [3](#)
- [35] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. [2](#), [3](#)
- [36] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. [3](#)
- [37] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. [19](#)
- [38] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *ICCV*, 2025. [2](#), [3](#), [7](#), [8](#), [20](#), [22](#), [23](#)
- [39] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. [20](#)
- [40] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. [19](#)
- [41] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. [19](#)
- [42] Yawen Luo, Xiaoyu Shi, Junhao Zhuang, Yutian Chen, Quande Liu, Xintao Wang, Pengfei Wan, and Tianfan Xue. Shotstream: Streaming multi-shot video generation for interactive storytelling. *arXiv preprint arXiv:2603.25746*, 2026. [2](#), [3](#), [4](#), [7](#), [8](#), [10](#), [20](#), [22](#), [23](#)
- [43] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *CVPR*, 2024. [2](#), [3](#)
- [44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. [20](#)
- [45] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. [2](#), [3](#), [17](#)
- [46] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. [2](#), [3](#)
- [47] Guocheng Gordon Qian, Ruihang Zhang, Tsai-Shien Chen, Yusuf Dalva, Anujraaj Argo Goyal, Willi Menapace, Ivan Skorokhodov, Meng Dong, Arpit Sahni, Daniil Ostashev, et al. Layercomposer: Interactive personalized t2i via spatially-aware layered canvas. *arXiv preprint arXiv:2510.20820*, 2025. [3](#)
- [48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. [7](#), [17](#), [20](#)
- [49] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. [3](#), [17](#), [20](#)
- [50] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryal, 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. [20](#)
- [51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. [3](#)

- [52] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3
- [53] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 3
- [54] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024. 3
- [55] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023. 3
- [56] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3
- [57] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 3
- [58] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 2, 5, 6, 15
- [59] Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026. 20
- [60] A Vaswani. Attention is all you need. In *NeurIPS*, 2017. 2, 3
- [61] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwei Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3, 17, 20
- [62] Jiahao Wang, Hualian Sheng, Sijia Cai, Weizhan Zhang, Caixia Yan, Yachuang Feng, Bing Deng, and Jieping Ye. Echoshot: Multi-shot portrait video generation. In *NeurIPS*, 2025. 2, 3, 4, 7, 8, 10, 20, 22, 23
- [63] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 8, 21
- [64] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 3
- [65] 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 *CVPR*, 2026. 2, 3, 4, 7, 8, 10, 20, 22, 23
- [66] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2024. 7, 20
- [67] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 7, 20
- [68] 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 *SIGGRAPH*, 2024. 2, 3, 10
- [69] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 6, 7, 17, 19
- [70] 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. 2, 3, 4, 7, 8, 10, 20, 22, 23
- [71] Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. Mind the time: Temporally-controlled multi-event video generation. In *CVPR*, 2025. 2, 3, 4
- [72] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025. 2, 3
- [73] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3
- [74] Zhenghao Zhang, Junchao Liao, Xiangyu Meng, Long Qin, and Weizhi Wang. Tora2: Motion and appearance customized diffusion transformer for multi-entity video generation. In *ACM MM*, 2025. 3
- [75] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023. 19
- [76] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024. 3, 10

# **CineOrchestra: Unified Entity-Centric Conditioning for Cinematic Video Generation**

## Supplementary Material

### A Derivation and Properties of $\beta(L)$

This appendix derives the closed form of the duration-dependent scaling  $\beta(L)$  from Eq. (2) and proves its three properties: (P1) zero-duration limit  $\beta(0) = 1$ , (P2) monotonic growth toward a bounded asymptote, and (P3) approximate duration-invariance of the peak similarity  $\max_{t_v} s(t_v; e)$ . Fig. 7 summarizes all three.

**Setup.** Standard RoPE [58] in  $d_{\text{rope}}$  dimensions is the block-diagonal rotation

$$R(t) = \text{diag}(R_1(\theta_1 t), R_2(\theta_2 t), \dots, R_{d_{\text{rope}}/2}(\theta_{d_{\text{rope}}/2} t)), \quad (4)$$

where each  $R_n(\alpha) \in \text{SO}(2)$  is a planar rotation by angle  $\alpha$  and  $\{\theta_n\}_{n=1}^{d_{\text{rope}}/2}$  are the per-channel RoPE frequencies. Each block is orthogonal, so  $\|R(t)\|_F^2 = d_{\text{rope}}$  for every  $t$ . The dot product between a video token at time  $t_v$  and an event token is

$$s(t_v; e) = \frac{\beta(L)}{N} \sum_{i=0}^{N-1} \mathbf{q}^\top R(t_i - t_v) \mathbf{k}, \quad (5)$$

where  $\mathbf{q}, \mathbf{k} \in \mathbb{R}^{d_{\text{rope}}}$  are the query and key projections of the video and event-caption tokens.

**Closed form.** Identifying each  $\text{SO}(2)$  block with the complex unit circle via  $R_n(\alpha) \leftrightarrow e^{j\alpha}$ , the  $n$ -th block of the unnormalized average  $\frac{1}{N} \sum_{i=0}^{N-1} R(t_i)$  corresponds to

$$z_n(L) = \frac{1}{N} \sum_{i=0}^{N-1} e^{j\theta_n t_i}. \quad (6)$$

Writing  $\bar{t} = (t_s + t_e)/2$ , the offsets  $t_i - \bar{t}$  are evenly spaced over  $[-L/2, L/2]$ , so the Dirichlet-kernel identity yields

$$|z_n(L)| = \frac{1}{N} \left| \frac{\sin(N\theta_n L/[2(N-1)])}{\sin(\theta_n L/[2(N-1)])} \right| \xrightarrow{N \rightarrow \infty} |\text{sinc}(\theta_n L/2)|, \quad (7)$$

with  $\text{sinc}(x) := \sin(x)/x$ . We adopt the continuous- $N$  limit below; for  $N = 16$  and the RoPE frequencies considered here, the discrete and continuous expressions agree to within 1%.

The squared Frobenius norm of the unnormalized average is the sum of the per-block contributions  $\|R_n\|_F^2 = 2$  scaled by  $|z_n(L)|^2$ :

$$\left\| \frac{1}{N} \sum_i R(t_i) \right\|_F^2 = 2 \sum_{n=1}^{d_{\text{rope}}/2} \text{sinc}^2(\theta_n L/2). \quad (8)$$

Imposing  $\|\mathbf{P}^{\text{evt}}(e)\|_F^2 = \beta(L)^2 \cdot 2 \sum_n \text{sinc}^2(\theta_n L/2) = d_{\text{rope}}$  yields the closed form

$$\beta(L) = \left( \frac{2}{d_{\text{rope}}} \sum_{n=1}^{d_{\text{rope}}/2} \text{sinc}^2(\theta_n L/2) \right)^{-1/2}. \quad (9)$$

Eq. (9) depends only on  $L$  and the RoPE spectrum  $\{\theta_n\}$  and can be precomputed once.

**Property 1: zero-duration limit.** At  $L = 0$ , all sample positions coincide ( $t_i = t_s$  for every  $i$ ), so  $\text{sinc}(0) = 1$  and the sum in Eq. (9) equals  $d_{\text{rope}}/2$ . Substituting gives  $\beta(0) = \left( \frac{2}{d_{\text{rope}}} \cdot \frac{d_{\text{rope}}}{2} \right)^{-1/2} = 1$ , and  $\mathbf{P}^{\text{evt}}(e)$  collapses to the single rotation  $R(t_s)$ , recovering standard single-position RoPE (green marker in Fig. 7(a)).

**Property 2: monotonic growth and bounded asymptote.** Each  $\text{sinc}^2(\theta_n L/2)$  is non-increasing on  $[0, 2\pi/\theta_n]$ , the regime in which the corresponding frequency component has not yet completed its

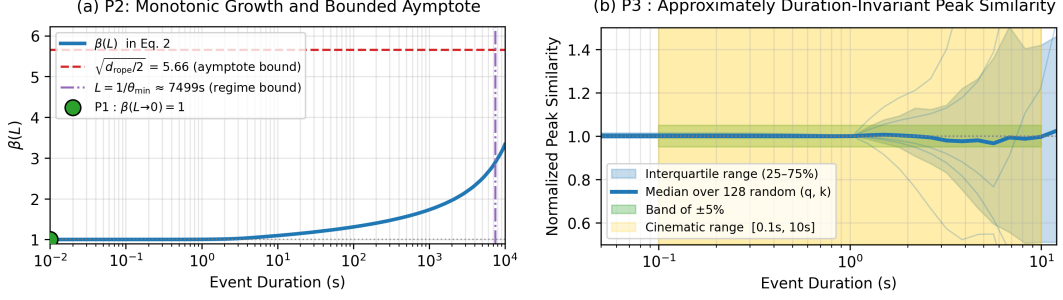


Figure 7: **Closed-form  $\beta(L)$  and its three properties.** (a) shows  $\beta(L)$  from Eq. (2) across event durations  $L$  in log-scale. The green marker verifies **P1** ( $\beta(0) = 1$ ). The curve is monotone (**P2**), upper-bounded by  $\sqrt{d_{\text{rope}}/2}$  (red dashed) and approaches this asymptote only past  $L \approx 1/\theta_{\text{min}}$  (purple dashed), so the cinematic range sits well within the bounded regime. (b) verifies **P3**: normalized peak similarity  $\max_{t_v} s(t_v; e)$  (median and 25–75% range over 128 random  $(\mathbf{q}, \mathbf{k})$  pairs) stays within  $\pm 5\%$  (green band) across  $[0.1\text{s}, 10\text{s}]$  (yellow region).

first arc. Across this range  $\sum_n \text{sinc}^2(\theta_n L/2)$  is non-increasing, so  $\beta(L)$  is non-decreasing in  $L$ . As  $L$  grows past this range, high-frequency components ( $\theta_n L \gtrsim 2\pi$ ) average out and only the slowest frequencies contribute meaningfully to the sum. In the practical regime  $\theta_{\text{min}} L \ll 1$  (satisfied for all event durations encountered in this paper given standard RoPE frequencies),  $\beta(L)$  is therefore upper-bounded by

$$\beta(L) \leq \sqrt{d_{\text{rope}}/2}, \quad (10)$$

attained when only the smallest RoPE frequency contributes the full  $|\text{sinc}(\theta_{\text{min}} L/2)| \approx 1$ . Fig. 7(a) plots  $\beta(L)$  (blue) against this asymptote (red dashed) and the regime bound  $L \approx 1/\theta_{\text{min}}$  (purple dashed), showing the cinematic range sits well inside the bounded regime.

**Property 3: approximate duration-invariance of the peak similarity.** Decomposing  $\mathbf{q}, \mathbf{k}$  into per-block 2D components  $\mathbf{q}_n, \mathbf{k}_n \in \mathbb{R}^2$  and writing  $\mathbf{q}_n^T R_n(\alpha) \mathbf{k}_n = a_n \cos(\alpha + \phi_n)$  for amplitudes  $a_n = \|\mathbf{q}_n\| \|\mathbf{k}_n\|$  and phases  $\phi_n$ , the similarity profile of Eq. (5) expands as

$$s(t_v; e) = \beta(L) \sum_{n=1}^{d_{\text{rope}}/2} a_n \text{sinc}(\theta_n L/2) \cos(\theta_n [\bar{t} - t_v] + \phi_n). \quad (11)$$

At  $t_v = \bar{t}$  the cosine factor reduces to  $\cos(\phi_n)$  for every  $n$ , and bounds on  $\max_{t_v} s(t_v; e)$  follow from how the per-frequency amplitudes  $\{a_n\}$  are distributed across the spectrum.

(i) *Spectrally diffuse  $\mathbf{q}, \mathbf{k}$ .* When the energy  $\{a_n^2\}$  is spread over multiple frequencies, the Cauchy–Schwarz inequality gives

$$|s(\bar{t}; e)| \leq \beta(L) \sqrt{\sum_n a_n^2} \sqrt{\sum_n \text{sinc}^2(\theta_n L/2)}. \quad (12)$$

Substituting Eq. (9) cancels the second square-root factor exactly, leaving

$$|s(\bar{t}; e)| \leq \sqrt{d_{\text{rope}}/2} \sqrt{\sum_n a_n^2}, \quad (13)$$

which is independent of  $L$ . The peak similarity is therefore *exactly*  $L$ -invariant in the spectrally diffuse regime.

(ii) *Spectrally concentrated  $\mathbf{q}, \mathbf{k}$ .* When most of the energy concentrates on a single frequency  $\theta_{n^*}$ ,

$$|s(\bar{t}; e)| \approx a_{n^*} \beta(L) |\text{sinc}(\theta_{n^*} L/2)|, \quad (14)$$

and invariance holds up to a multiplicative factor that depends on whether  $\theta_{n^*}$  lies in the slow- or fast-decaying part of the spectrum. The factor is bounded above by  $\sqrt{d_{\text{rope}}/2}$  in either case (Property 2).

Learned attention patterns sit between regimes (i) and (ii) in practice. Fig. 7(b) confirms empirically that  $\max_{t_v} s(t_v; e)$  stays within  $\pm 5\%$  across the full cinematic range  $L \in [0.1\text{s}, 10\text{s}]$ .

## B Implementation Details

**Data samples.** Our training corpus consists of videos ranging from one minute to two hours with wide-varying camera angles/shots/multiple entities across multiple scenes. Each title is split into non-overlapping one-minute chunks that serve as atomic training samples. We run a YOLO v11 person detector [33] on one frame every 10 s and discard chunks whose majority of probe frames have either zero people or a count above a fixed clutter threshold. A 2% random sample of training chunks (with overlapping source titles) is held out only as a loss-monitoring validation set. Our test benchmark *CineBench* is built from a disjoint pool of source titles with no character or location overlap with training; we manually curate 512 one-minute chunks from this pool.

### B.1 Annotation of Entity-Centric Cinematic Conditioning

Each chunk is annotated through entity-centric captioning followed by reference-image curation with appearance augmentation.

**Entity-centric captioning.** A single Gemini-2.5-Pro [20] call (prompt in Fig. 8, fixed JSON schema) returns a unique curly-brace tag per distinct entity (e.g., {girl\_young}, {scene\_museum}), a one-sentence global appearance description per visual subject, and a temporally dense per-entity timeline of `<mm:ss.ff>`-stamped events whose descriptions begin with the primary entity tag. Two cinematography entities are tracked alongside the visual subjects: {camera} (pans, zooms, tracks, handheld) and {transition} (hard cuts, fades, dissolves, animated wipes).

**Reference-image curation with appearance augmentation.** For each visual subject we sample one frame at the centre of each event and prompt Gemini-3-Flash-Preview [20] for a bounding box, reusing the global description as a disambiguating clue (entities with no returned box are dropped). We crop each box, encode it with CLIP ViT-L/14 [48], and rank crops by cosine similarity against the entity’s global description; we keep the top four and sample two without replacement with probability proportional to score. The pair is passed jointly into Qwen-Image-Edit [69] with an entity-aware prompt that places the subject in a new background, lighting, and pose (also varying expression for people-like entities). The resulting image is the appearance-augmented reference  $\mathbf{I}_k$  used during training;

**Per-chunk caption statistics of *CineBench*.** Each chunk averages 4.6 visual subject entities and 13.4 events (8.7 subject, 2.7 camera, 2.0 transitions). Event durations are heavily right-skewed (median 2.5 s, 10<sup>th</sup> pct. 0.7 s, 90<sup>th</sup> pct. 10.2 s) with a long tail of  $L=0.1$  shot-transition events. This combination of many entities, many events per entity, and per-event durations spanning four orders of magnitude motivates the duration-aware  $\beta(L)$  rescaling in Sec. 3.3.

### B.2 Model Architecture

We use the Wan 2.1 video autoencoder [61] (frozen,  $8\times 8$  spatial and  $4\times$  temporal compression, 16 latent channels) to encode video, and apply it identically to reference images as 1-frame clips. The backbone is a pre-LN diffusion transformer [45] that operates on the video latent concatenated with reference-image tokens; each block applies self-attention over the joint sequence, cross-attention to the text-only streams, and a gated MLP, with time-step and scalar conditioning projected to per-block modulation weights. All captions are encoded with T5-XXL [49] (4096-dim/token); per chunk we encode up to 16 entities (with 128-token global descriptions) and up to 32 dense events (with 64-token event descriptions). The two streams stay separate so the entity-temporal RoPE (Sec. 3.3) can address them under different coordinate conventions, and empty slots are zero-padded with attention masking. Each chunk additionally carries up to  $K=4$  appearance-augmented references  $\mathbf{I}_k$  (App. B.1), one per dominant entity; unused slots are zero-filled and masked.

### Entity-Centric Captioning Prompt

You are an expert system for *entity-centric, temporally dense event captioning*.

#### # Instructions (follow strictly)

1. Watch the entire video.
2. Identify all distinct entities (e.g., people, animals, objects, scenes) appearing in the video and create an *exclusive entity tag* (enclosed by *curly brackets*) for each entity.
  - a) For the entity tags of scenes which are part of background/foreground, use "scene" as the first word of the tags (e.g., {scene\_room\_dark}, {scene\_sky}, {scene\_car\_interior}).
3. For each entity, produce *one detailed description of its appearance* and *multiple temporally dense, meaningful events*:
  - a. Each appearance and event description **ONLY** focuses on **ONE** primary entity.
  - b. The first word of each description **MUST BE** the entity tag of the primary entity (i.e., use it as the subject of the sentences).
  - c. [IMPORTANT] Generate events **ONLY** when the primary entity is **VISIBLE** in the video (including background but non-/partially-occluded appearances).
  - d. Start a new event whenever primary entity shows a noticeable change in:
    - i. *Action* (e.g., start/stop/do something new)
    - ii. *Interaction* (e.g., touch, hand over, look at, gesture to)
    - iii. *Movement* (e.g., direction/speed change, enters/exits frame)
    - iv. *Pose* (e.g., raise/lower arms, turn head, sit/stand)
    - v. *Object state* (e.g., open/close, appear/disappear, move location)
    - vi. *Expression* (e.g., smiles, frowns, squints, opens mouth)
  - e. Use the format of <mm:ss.ff> for the timestamps.
4. Camera events:
  - a. Treat camera behavior as a separate entity.
  - b. No need to generate appearance description for the camera entity.
  - c. *Include camera movements* when they occur (e.g., "camera pans left", "slow zoom in", "handheld shake", "camera tracks forward").
5. Shot transition events:
  - a. Treat shot transition as a separate entity.
  - b. Identify every *shot boundary discontinuity*, including but not limited to
    - i. *Hard Cut*: an instantaneous shift from one shot to another.
    - ii. *Fade*: a gradual transition from a shot to a solid color (usually black) or vice versa.
    - iii. *Dissolve/Cross-fade*: a gradual transition where one shot overlaps and replaces another.
    - iv. *Animation*: an object/image animation commonly used to make slides like powerpoint or keynotes.
    - v. Many others...
  - c. **ONLY** label the transitions with discontinuous shots. **DO NOT** label the transitions caused by camera movement, zooms, obstructions, or lighting changes.
  - d. To label a *hard cut*, use the same start\_time and end\_time.

#### # Output JSON example (all entities need to be packed by brackets)

```
[
  {
    "entity": "{man_young}",
    "appearance_description": "{man_young} is a young and tan-skinned man. ...",
    "timeline": [
      {
        "start_time": "<00:00.00>",
        "end_time": "<00:03.50>",
        "event_description": "{man_young} jumps into a car."
      },
      {
        "start_time": "<00:03.50>",
        "end_time": "<00:05.00>",
        "event_description": "{man_young} talks to {man_old} ..."
      }
    ]
  },
  ...
  {
    "entity": "{camera}",
    "timeline": [
      {
        "start_time": "<00:01.00>",
        "end_time": "<00:02.50>",
        "event_description": "{camera} is static."
      }
    ]
  },
  {
    "entity": "{transition}",
    "timeline": [
      {
        "start_time": "<00:01.00>",
        "end_time": "<00:01.00>",
        "event_description": "{transition} shows a hard cut."
      }
    ]
  }
]
```

Figure 8: **Entity-centric captioning prompt.** A single structured-output query to Gemini-2.5-Pro [20] returns entity tags, per-entity global descriptions, and dense <mm:ss.ff>-stamped event timelines for visual subjects, {camera}, and {transition} in one pass.

### B.3 Training and Inference

**Pretraining and fine-tuning.** We first pretrain the backbone on a large corpus of generic video-caption pairs at the same architecture and latent space, with the entity-temporal RoPE and reference-image inputs disabled, then fine-tune *CineOrchestra* from this checkpoint on the entity-annotated chunks of App. B. All hyperparameters below refer to the fine-tuning stage. Each step draws a 10-second window from a one-minute chunk and decodes it at the native 15 fps, yielding 153-frame clips at  $288 \times 512$ .

**Optimizer, schedule, and objective.** We use AdamW [41] ( $\beta_1 = 0.9$ ,  $\beta_2 = 0.99$ ,  $\epsilon = 10^{-8}$ , weight decay 0.01); most parameters train at  $3 \times 10^{-5}$ , while the cross-attention layers reorganised under the 2D entity-temporal RoPE (Sec. 3.5) train at  $1 \times 10^{-4}$ . The schedule is constant after a 1,000-step linear warmup with global gradient-norm clip 1.0. Mixed precision uses `bfloat16` parameters and `float32` gradient reductions, sharded via FSDP-2 [75] with per-block activation checkpointing; an EMA copy at  $\beta_{\text{ema}} = 0.9999$  (after 1,000-step warmup) is used for all reported numbers and samples. We train rectified flow [37, 40] with  $\sigma_{\text{data}} = 1$ ,  $\sigma_{\text{noise}} = 2$ , logit-normal timesteps (location 0, scale 1,  $\epsilon = 10^{-3}$ , logit-shift 3.0) and the EDMv2-normalised regression [32] with image/video weights both 1.0. To enable separate inference-time guidance scales, we apply Bernoulli CFG-dropout [28] with probability 0.1 to each of the reference-image stream, the joint global-and-dense caption stream, and the auxiliary scalar conditions (resolution, dataset id, sampling frame rate); the two text streams are dropped synchronously. Training lasts 25,000 iterations on 32 NVIDIA H100 GPUs at per-GPU batch size 2.

**Multi-condition CFG and inference.** *CineOrchestra* takes two grouped conditioning modalities: text  $c_{\text{txt}}$  (covering  $g_k$  and  $e_{k,j}$ ) and image  $c_{\text{img}}$  (covering  $\mathbf{I}_k$ ). We combine them as

$$\begin{aligned} \tilde{f}_\theta(z_t, c_{\text{img}}, c_{\text{txt}}) = & f_\theta(z_t, c_{\text{img}}, c_{\text{txt}}) \\ & + \lambda_{\text{txt}} \cdot (f_\theta(z_t, c_{\text{img}}, c_{\text{txt}}) - f_\theta(z_t, c_{\text{img}}, \emptyset)) \\ & + \lambda_{\text{img}} \cdot (f_\theta(z_t, c_{\text{img}}, c_{\text{txt}}) - f_\theta(z_t, \emptyset, c_{\text{txt}})) \\ & + \lambda_{\text{joint}} \cdot (f_\theta(z_t, c_{\text{img}}, c_{\text{txt}}) - f_\theta(z_t, \emptyset, \emptyset)), \end{aligned}$$

with  $\emptyset$  the fully-unconditional pass and  $\lambda_{\text{joint}} = 5$ ,  $\lambda_{\text{img}} = 3$ ,  $\lambda_{\text{txt}} = 0$ . We sample with a 40-step rectified-flow sampler at the same  $288 \times 512 / 153$ -frame (10 s at 15 fps) shape used during training; in practice we can also generate 40 s videos at  $288 \times 512$  or 10 s videos at  $720 \times 1280$  as we show in our supplementary material and in App. D. Timesteps are warped with a time-shifting factor of 5.66 to spend more steps in the higher-noise region.

## C Evaluation Details

### C.1 Curation of *CineBenchSyn*

*CineBenchSyn* complements the real-footage benchmark with 512 hand-authored 10.2s scenarios. Each scenario emits the same JSON schema as the real-footage annotations (App. B.1) — `global_entities` covering `{camera}`, `{transition}`, `{scene_*}`, characters, animals, and props with their visible-time intervals, and `dense_entities` listing fine-grained timestamped events — so the same downstream pipeline serves both data sources. Authoring is parameterised across shot count, character count, and event density; zero-width hard cuts and the  $\leq 16$  entity /  $\leq 64$  event caps match training conventions, and a small validator enforces structural invariants (canonical scene names, no overlapping scenes, all referenced names resolve, intervals in  $[0, 10.2]$ ). For every non-camera, non-transition, non-scene entity we generate a single reference image with the publicly released Qwen-Image text-to-image model [69] at  $1024 \times 1024$ , 50 inference steps, true-CFG-scale 4.0, fixed seed, and a quality-oriented negative prompt. The image-generation prompt is built from the entity’s appearance description by stripping curly braces, replacing underscores with spaces, choosing the relative pronoun *who* or *which* from a hand-curated person/animal token list, reading singular/plural number off the description’s leading copula, and emitting “An image of *a/an* *<natural name>* *{who|which}* *{is|are}* *<description>*”. The PNG path is written back into the JSON’s `ref_image_path` field so training, evaluation, and user-study pipelines locate it identically to a real-footage reference. Across the evaluated slice, scenarios average 4.5 entities, 1.6 events per entity, 3.1 camera events, 2.1 transitions, and 3.5 reference images (max 6), with per-event durations spanning 0.1s to 10.2s.

## C.2 Baselines

Each *CineOrchestra* caption carries persistent entity descriptions, per-entity time intervals, per-entity reference images, and dense per-shot captions; adapting them to each baseline relies on four mechanisms. *Shot boundaries* are derived from transition events ( $N$ -shot samples have  $N-1$  transitions): hard cuts collapse to a single timestamp, while long fades are split at their midpoint and allocated half to each adjacent shot. *Annotation trimming* aligns the rendered duration: events past the rendered window are dropped, span-crossing events are clipped, and orphaned entities are removed. ShotStream [42], MultiShotMaster [65], VACE [31], and Phantom [38] render the full sample and need no trimming; CineTrans [70] (5.06 s) and EchoShot [62] (7.81s) are trimmed to their fixed durations before prompt construction. *Active-entity filtering* avoids T5’s [49] 512-token cap (silently truncated by Wan-family pipelines [61]): the per-shot context for MultiShotMaster, EchoShot, VACE, and Phantom keeps only entities whose intervals overlap the current shot, while ShotStream and CineTrans retain their released schema (a flat sample-level global of brace-stripped concatenated entity descriptions). *Reference-image captioning* is needed for the four text-only baselines (ShotStream, MultiShotMaster, CineTrans, EchoShot — MultiShotMaster had only released text-to-video weights at writing): we generate one- to two-sentence captions of each reference image with Qwen3.5-35B-A3B [59] (chain-of-thought disabled) and inline them as the entity description; VACE and Phantom consume reference images directly.

All baselines run with their public weights at  $832\times 480$  resolution and 16 fps (MultiShotMaster 15 fps), 50 denoising steps (ShotStream 4, using its distilled checkpoint). VACE and Phantom only support single-shot generation, so we run inference per shot and concatenate; CineTrans and EchoShot are restricted to fixed 5.06s and 7.81s outputs, while the rest match the per-shot lengths in our annotations.

## C.3 Evaluation Metrics

We evaluate every generated video on *CineBench* (and *CineBenchSyn*) against its conditioning annotations (entity tags, appearance descriptions, dense event timelines, transitions); the pipeline produces five metrics averaged across the 512 videos — two for entity identity, and one each for global-caption alignment, dense-caption alignment, and transition timing.

**Per-entity mask extraction.** For every non-camera, non-transition entity we merge visible-time intervals into disjoint segments and sample three keyframes per segment (at the 20/50/80<sup>th</sup> percentiles). We run Grounding-DINO (grounding-dino-base) [39] on each keyframe with the entity’s natural-language tag (box threshold 0.25, text threshold 0.20), re-score the boxes with CLIP ViT-B/32 [48], drop boxes with CLIP-text score below 0.1, and feed the survivors to SAM2 (sam2-hiera-large) [50] which produces both per-keyframe masks and a forward/backward-tracked mask video. The same crops, masks, and per-entity mask videos are reused by every downstream metric.

**Subject identity consistency (DINO, M-DINO).** We encode each entity’s reference image  $I_k$  with DINOv2 [44] and compute cosine similarity to every kept crop, taking the max per disjoint interval, the mean per entity, and the mean per video and benchmark. DINO uses raw bounding-box crops; M-DINO replaces the box background with black using the SAM2 keyframe mask. We report M-DINO as the headline subject-identity number.

**Global caption following (CLIP, M-CLIP).** The same crops are encoded with CLIP ViT-B/32 and scored against the CLIP text embedding of the entity tag in natural language (e.g. {man\_young}  $\rightarrow$  “man young.”). The unmasked / SAM-masked variants follow the same max-over-crops, mean-over-intervals, mean-over-entities aggregation; M-CLIP-text is reported.

**Dense caption following (ViCLIP).** For every dense event we extract the corresponding sub-clip, sample eight uniformly-spaced frames, and compute cosine similarity between the ViCLIP-B/16 video embedding [66, 67] and the ViCLIP text embedding of the event description (with entity tags expanded to natural language). Events are bucketed into *subject*, *scene*, *camera*, and *transition* categories, and the mean-over-events-within-video, mean-over-videos statistic is reported per category to form the dense-caption block of Tabs. 2 and 3.

**Shot-transition timing (VLM recall).** For each transition event we cut a sub-clip spanning  $[start - \delta_l, end + \delta_r]$ , with each side padded outward by up to 1.0 s and clipped to maintain at least 0.5 s of

separation from the next adjacent transition (or clip boundary). The sub-clip is passed to Qwen2.5-VL-7B-Instruct [63] with the event description and a yes/no prompt asking (i) whether any shot transition is visible (PRESENT) and (ii) whether the type/style matches the description (MATCHES). The reported number is the per-video *presence recall* averaged across the benchmark; the stricter *match recall* is computed as a sanity check.

#### C.4 User Study Protocol

Automatic metrics are coarse, especially for motion plausibility and aesthetic quality where neither CLIP- nor VLM-based evaluators are well calibrated. We therefore complement App. C.3 with a head-to-head human evaluation on *CineBench*; results are visualised in Fig. 5 and tabulated in Tab. 4.

**Setup and rating.** For every prompt, raters see two anonymised side-by-side videos (LEFT/RIGHT, side randomised per item) generated from the same conditioning bundle: per-entity reference images, global appearance descriptions, the dense event timeline, and the special {camera} / {transition} tracks. They judge against the prompt alone (no ground-truth shown) on the five-point scale (1=strongly Left, 3=tie, 5=strongly Right); ratings are required for every dimension and aggregated by averaging per dimension, with a method’s win-rate reported as the fraction of items whose average favours that side.

**Evaluation dimensions.** Eight dimensions are scored independently:

- *Subject/entity consistency* judges face, body, and core object identity for entities with a reference image (clothing/accessory drift is not penalised; items with no reference images default to 3).
- *Global-description following* scores per-entity appearance attributes (clothing, hairstyle, accessories, body type) excluding face identity and {scene\_\*} entities.
- *Dense-caption following* judges whether each event happens, with the right subject and roughly the right timing (excluding {camera} and {transition} events).
- *Shot structure* scores shot count, transition position and *type* (a hard cut for a requested dissolve is penalised), and absence of spurious cuts.
- *Camera angles and movement* scores pan/zoom direction, static-vs-moving, and continuous-move-vs-cut against the {camera} timeline.
- *Scene-description following* scores {scene\_\*} coverage and timing.
- *Overall motion quality* judges smoothness and physical plausibility independent of caption following.
- *Overall video quality* judges sharpness, composition, lighting, aesthetics, and frame-level cleanness.

**Annotation interface.** The task uses a video-annotation tool (Fig. 9) where dense events appear as time-track captions and global descriptions in a side panel; we extend the same labelling interface used to collect the underlying entity-centric annotations, so labellers see captions in the same layout they are rating against. Side assignment is randomised per item and method identity is hidden from raters.

Table 4: **Pairwise human-evaluation preferences on CineBench (%)**. Each cell reports raters’ wins / ties / losses rate for *CineOrchestra* against the indicated baseline along the corresponding dimension. Underline marks dimensions where wins exceed losses. We report the original numbers in Tab. 4

Method	Subject ID	Global Cap.	Dense Cap.	Camera	Scene	Shot Structure	Motion	Overall
Phantom [38]	<u>62/27/11</u>	<u>54/37/9</u>	<u>82/13/5</u>	<u>70/26/4</u>	<u>44/35/21</u>	<u>55/29/16</u>	<u>79/16/5</u>	<u>79/14/7</u>
VACE [31]	<u>81/9/9</u>	<u>65/21/14</u>	<u>73/14/14</u>	<u>68/26/6</u>	28/33/38	<u>37/42/21</u>	<u>65/20/15</u>	<u>66/13/21</u>
CineTrans [70]	<u>80/16/4</u>	<u>68/20/12</u>	<u>75/16/9</u>	<u>54/39/7</u>	30/38/32	<u>26/59/15</u>	<u>52/29/19</u>	<u>46/21/33</u>
EchoShot [62]	<u>84/10/6</u>	<u>78/16/6</u>	<u>86/10/4</u>	<u>68/28/4</u>	<u>48/31/21</u>	<u>40/44/16</u>	<u>68/22/10</u>	<u>65/18/17</u>
MultiShotMaster [65]	<u>76/16/8</u>	<u>68/22/10</u>	<u>75/16/9</u>	<u>55/34/10</u>	<u>41/39/21</u>	<u>36/42/21</u>	<u>45/27/28</u>	<u>44/22/34</u>
ShotStream [42]	<u>80/13/7</u>	<u>60/23/18</u>	<u>74/19/7</u>	<u>58/34/8</u>	25/34/42	<u>49/32/19</u>	29/33/38	32/14/54



Figure 9: **User study interface**. Side-by-side rating UI shown to in-house raters. Each video is scored on a 1 – 5 scale across eight axes spanning visual quality (overall quality, motion realism), identity preservation (reference ID consistency), and prompt adherence (global description, dense caption, shot structure, camera, and scene). App. C.4 shows full instructions and per-axis question.

Table 5: **Quantitative comparison of cinematic conditioning on *CineBenchSyn***. Best in **bold**, second-best underlined.

Method	Subject ID	Global Caption	Dense Caption Following (ViCLIP↑)				Transition Timing
	DINO↑	CLIP↑	Subject	Scene	Camera	Transition	Recall↑
Phantom [38]	<b>0.601</b>	<u>0.311</u>	0.237	<u>0.240</u>	0.215	0.125	0.114
VACE [31]	<u>0.562</u>	<b>0.315</b>	0.239	<b>0.243</b>	0.220	0.128	0.115
CineTrans [70]	0.489	0.295	<u>0.241</u>	0.227	<u>0.232</u>	<u>0.145</u>	0.230
EchoShot [62]	0.371	0.283	0.224	0.226	0.205	0.124	0.053
MultiShotMaster [65]	0.426	0.290	0.239	0.238	0.231	0.140	<u>0.258</u>
ShotStream [42]	0.466	0.290	0.224	0.206	0.206	0.132	0.165
<i>CineOrchestra</i> (ours)	0.556	0.310	<b>0.245</b>	<u>0.240</u>	<b>0.251</b>	<b>0.146</b>	<b>0.360</b>

## D Additional Results

We report the quantitative evaluation on *CineBenchSyn* in Tab. 5, and provide additional qualitative comparisons at the default resolution  $288 \times 512$  in Figs. 10 and 11 (*CineBench*) and Figs. 12 to 15 (*CineBenchSyn*). Please refer to our project website for video comparisons and additional results.

Figs. 16 and 17 report comparisons at  $720 \times 1280$  for 10s long videos even when our model is trained only up to 5 seconds. The four control axes the framework targets transfer cleanly to the higher resolution: per-subject reference identity is preserved across shot boundaries, camera primitives match their textual specification, shot transitions occur at the prescribed timestamps with the prescribed type (hard cut, cross-dissolve, wipe), and dense per-event captions are reflected in the relevant frames. By generating more tokens at 720p, the model achieves much higher quality, characterized by richer detail, enhanced aesthetic appeal, and smoother motion dynamics compared to lower-resolution baselines.

Figs. 18 and 19 extend the evaluation to 40s clips with substantially more scripted shot transitions and dense, time-localized events spanning the full duration. Identity, scene grounding, and dense-caption alignment remain stable over the longer horizon; entities reappear consistently after intervening shots, camera primitives resolve to the requested framings, and shot transitions trigger at the right moments with the right type. We observe minimal qualitative degradation maintaining high levels of entity caption reference binding as the duration is scaled from the 10s training horizon to 40s at inference.



Figure 10: Additional qualitative comparison on *CineBench*.

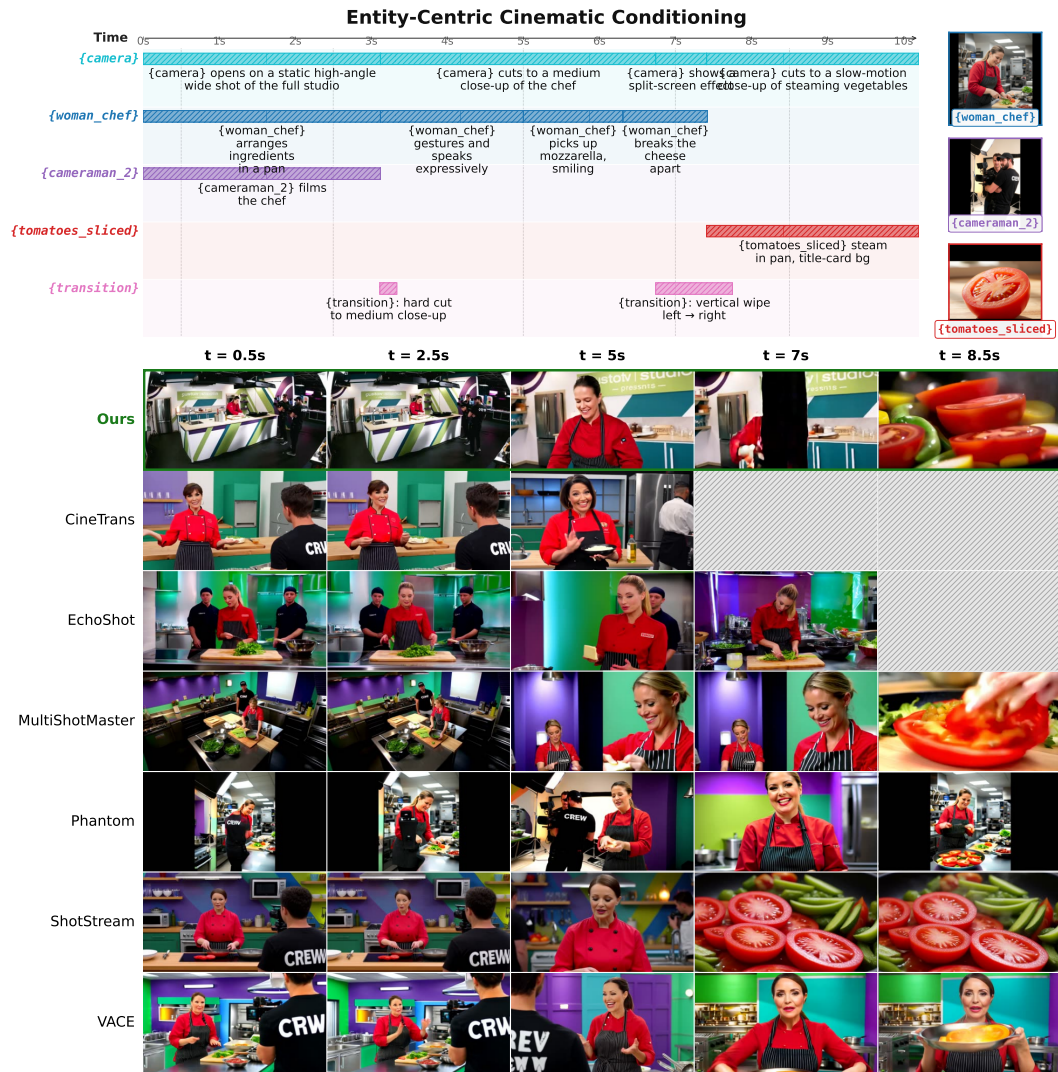


Figure 11: Additional qualitative comparison on *CineBench*.

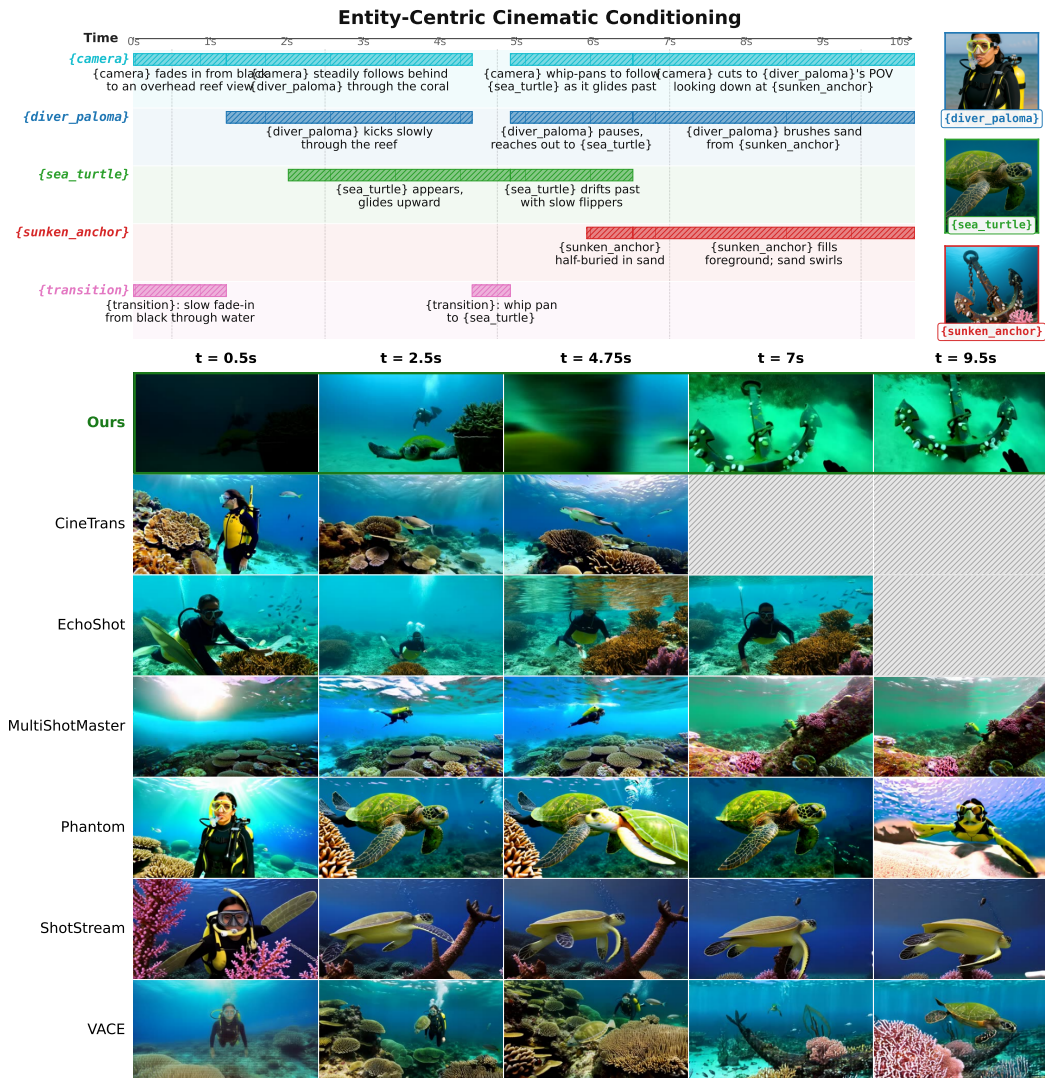


Figure 12: Additional qualitative comparison on *CineBenchSyn*.

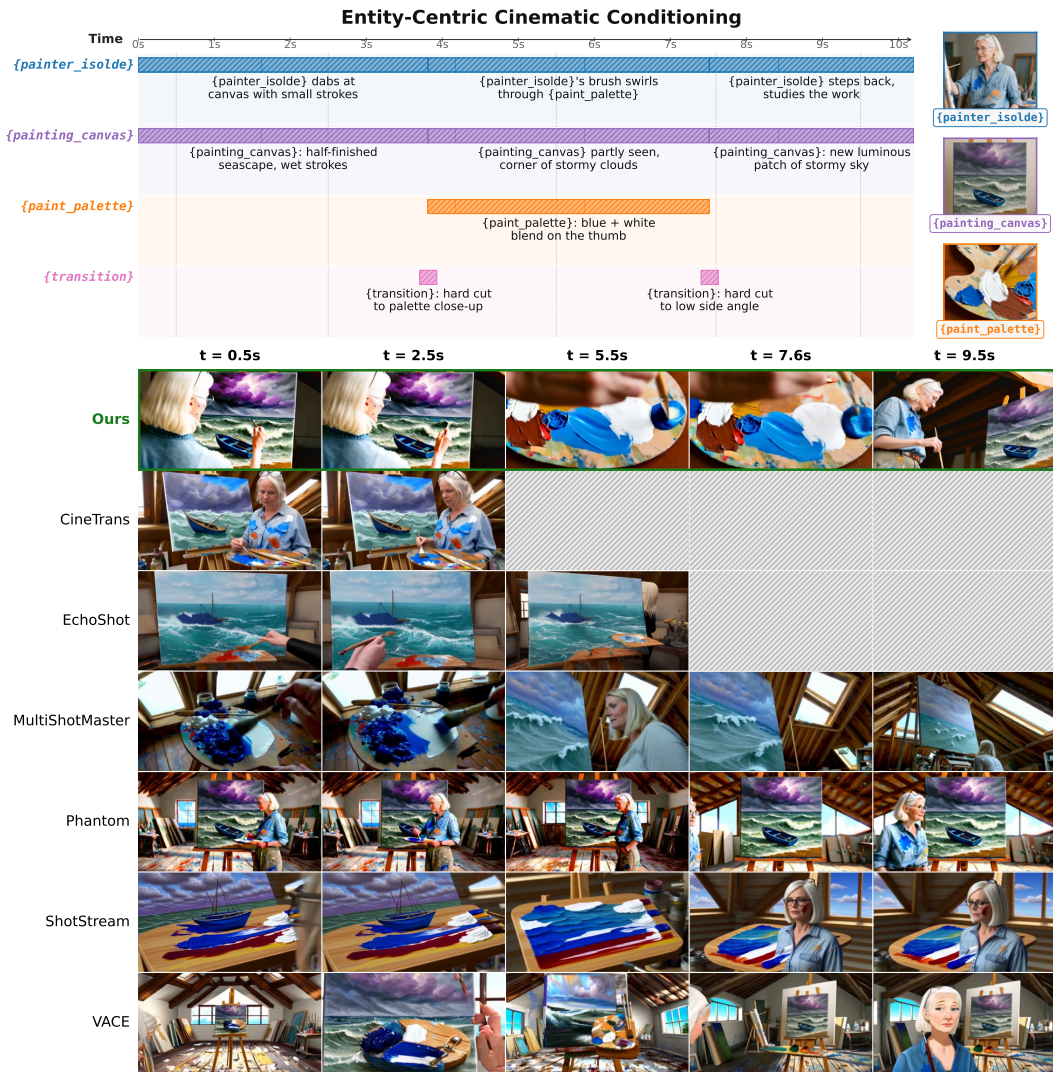


Figure 13: Additional qualitative comparison on *CineBenchSyn*.



Figure 14: Additional qualitative comparison on *CineBenchSyn*.

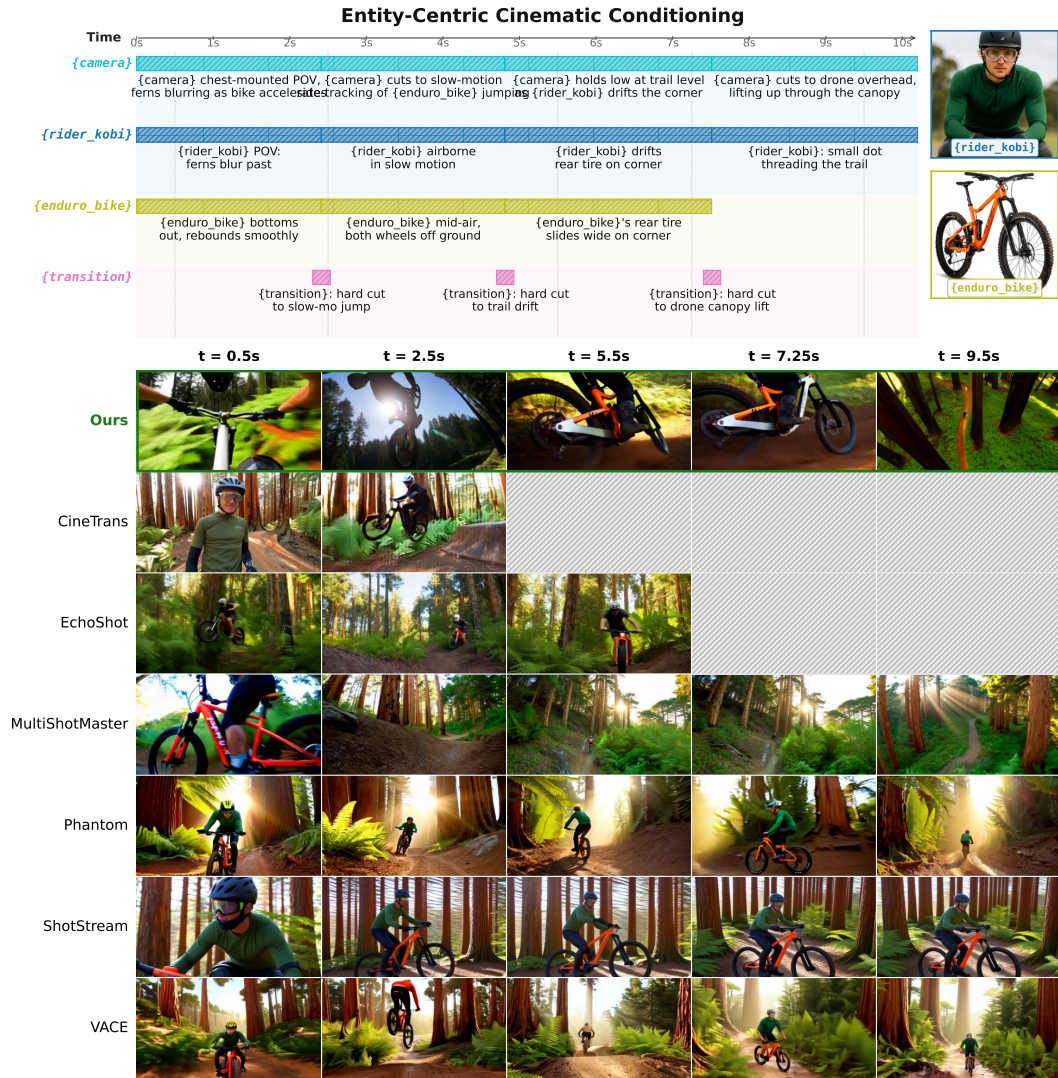


Figure 15: Additional qualitative comparison on *CineBenchSyn*.



Figure 16: Qualitative comparison at 720p resolution for *CineOrchestra* on *CineBench*.

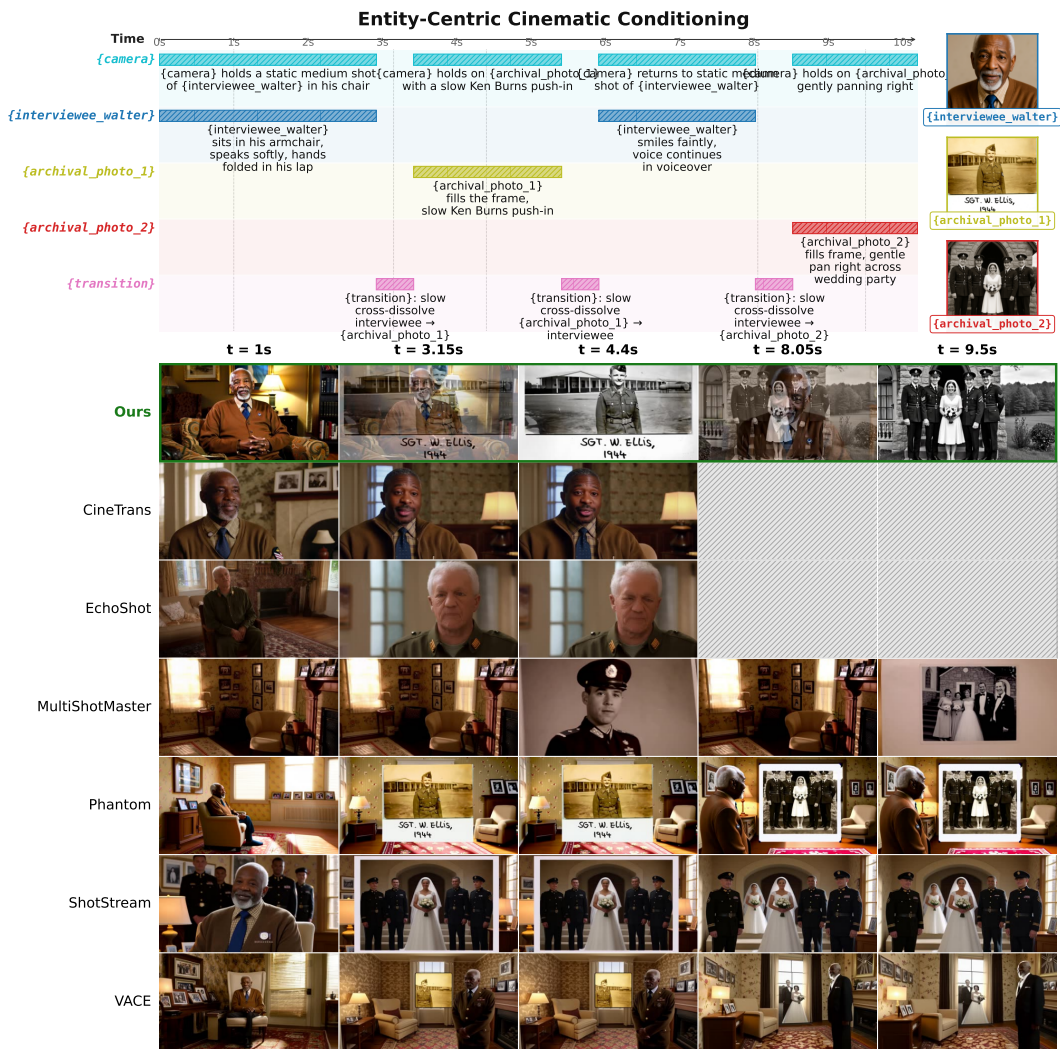


Figure 17: Qualitative comparison at 720p resolution for *CineOrchestra* on *CineBenchSyn*.

### Entity-Centric Cinematic Conditioning — long videos

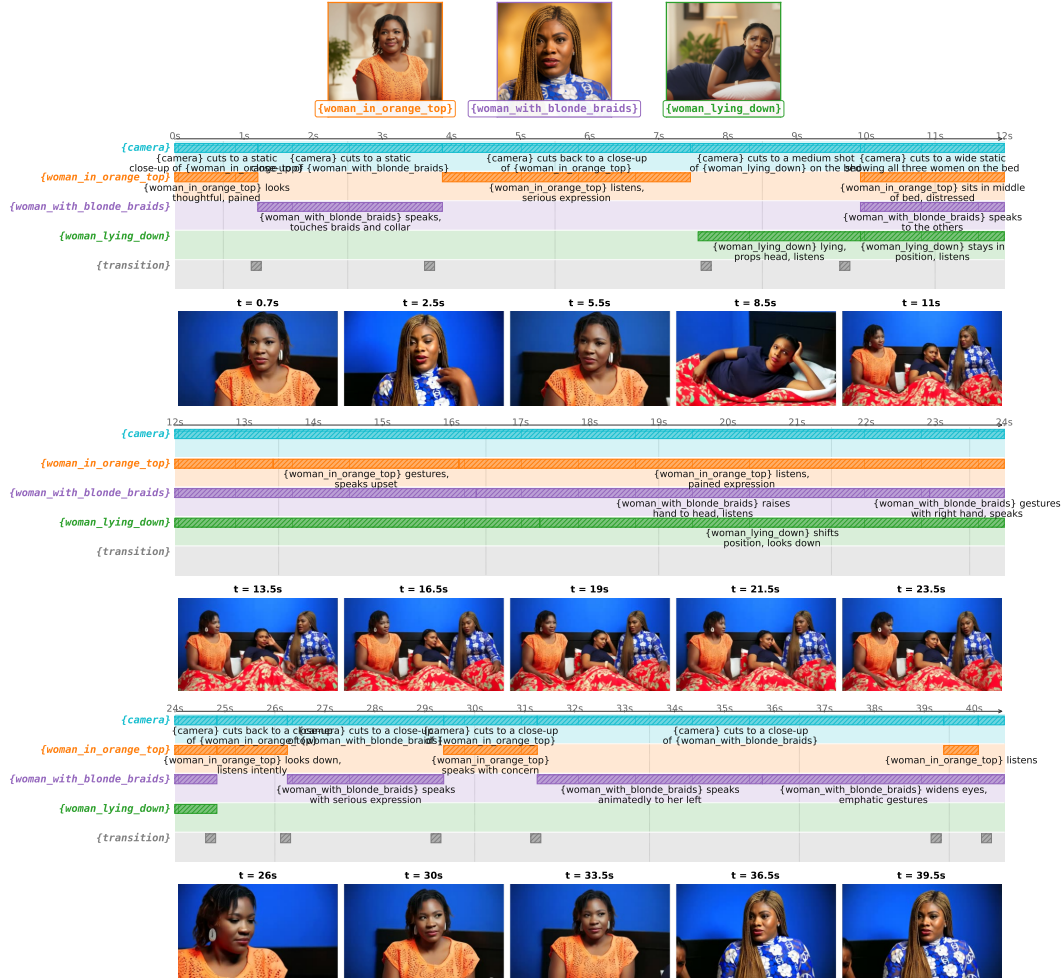


Figure 18: Long video generation (40s) from *CineOrchestra* on *CineBench*.

### Entity-Centric Cinematic Conditioning — long videos

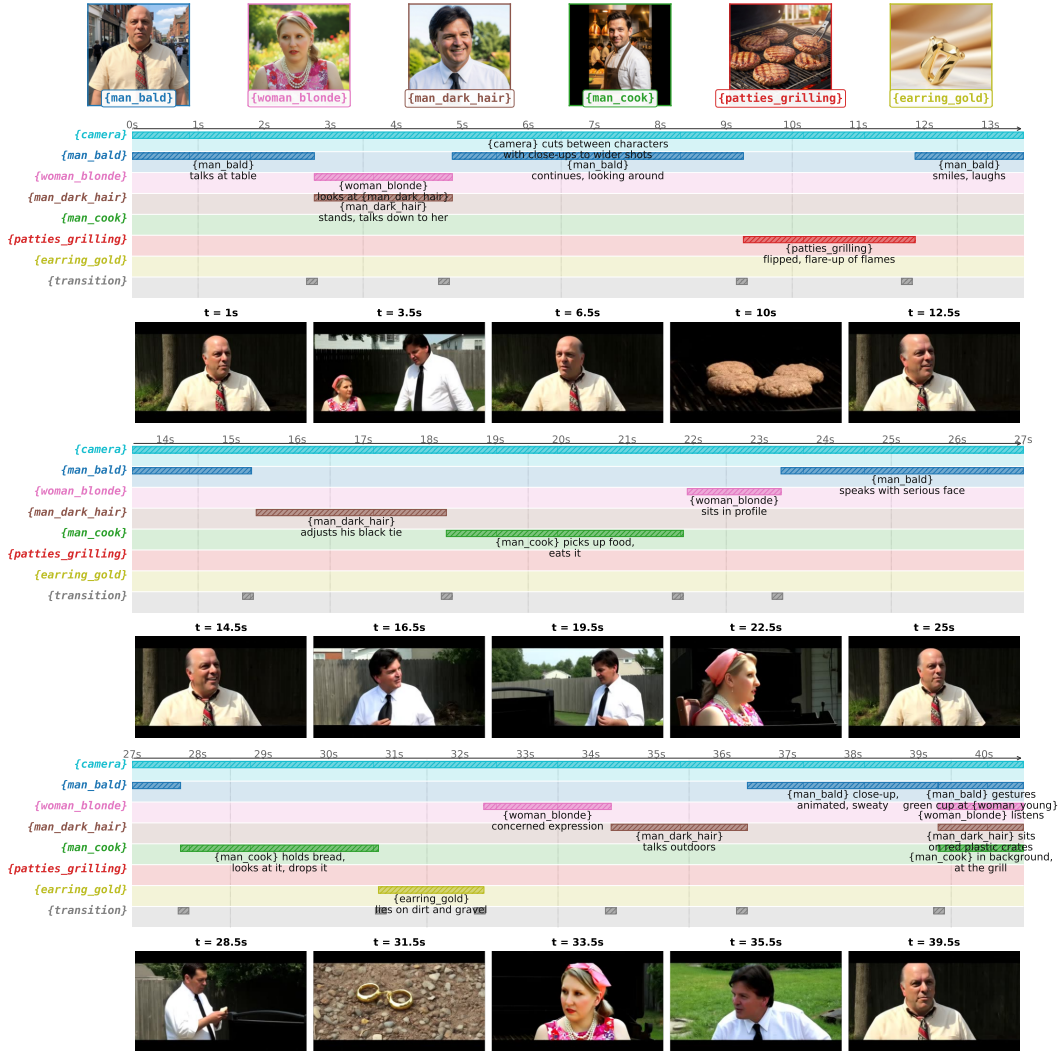


Figure 19: Long video generation (40s) from *CineOrchestra* on *CineBench*.