Noname manuscript No.
(will be inserted by the editor)

CineDance: Towards Next-Generation Multi-Shot Long-Form Cinematic Audio-Video Generation

Yuheng Chen[1] · **Teng Hu**[1] · **Yuji Wang**[1] · **Qingdong He**[2] · **Zhucun Xue**[3] · **Qianyu Zhou**[4] · **Xiangtai Li**[5] · **Lizhuang Ma**[1] · **Jiangning Zhang**[3] · **Dacheng Tao**[5]

Received: date / Accepted: date

Abstract The fidelity and structural diversity of training datasets fundamentally determine the capabilities of video generation models. While commercial systems show remarkable ability to generate cinematic narratives, the progress of open-source models remains limited by the scarcity of high-quality training data. To bridge this gap, we introduce CineDance-1M, a large-scale, open research Text-to-Audio-Video (T2AV) dataset designed specifically for multi-shot, long-form joint audio-video generation. Averaging 92.8 seconds and 24.2 continuous shots per video, it provides configurable, structured annotations for both audio and video modalities. This exceptional quality is achieved through a rigorous three-stage curation pipeline: *i)* diverse sourcing and comprehensive cleansing, *ii)* film-theory-inspired narrative parsing, and *iii)* hierarchical dual-modal captioning. For a comprehensive assessment, we propose CineBench, featuring a diverse prompt suite and a six-dimensional, human-aligned metric system tailored for complex narrative audio-video evaluation. Furthermore, we adapt LTX-2.3 into CineDance, which demonstrates exceptional single-modality quality alongside precise audio-video alignment and robust subject and environment consistency, effectively validating our curation strategy and the high quality of CineDance-1M. We anticipate that this work will serve as a solid foundation for accelerating future research in multi-shot, long-form joint audio-video generation. Our project page is available at https://aliothchen.github.io/projects/CineDance/.

[1] Shanghai Jiao Tong University, Shanghai, China.
[2] University of Electronic Science and Technology of China, Chengdu, China.
[3] Zhejiang University, Hangzhou, China.
[4] The University of Tokyo, Tokyo, Japan.
[5] Nanyang Technological University, Singapore, Singapore.

Keywords Text-to-Audio-Video Generation, Long-Form Video Generation, Multi-Shot Generation, Cinematic Dataset, Audio-Video Benchmark

1 Introduction

The unprecedented evolution of video generative models has catalyzed a growing demand for high-fidelity visual content across film production, immersive media, and interactive entertainment [21–23, 33, 57, 65, 87, 95]. While current works have achieved remarkable visual excellence in generating single-shot clips, the transition to multi-shot, long-form narrative generation remains largely underexplored. This progression is primarily bottlenecked by the scarcity of such large-scale open-source datasets, coupled with the limited generalizability of existing foundation models to long-form generation, leaving a vast landscape for future exploration.

However, multi-shot long-form audio-video generation is not a straightforward extension of short-clip synthesis, since it faces ***two core challenges***. Firstly, **long-horizon scalability**: Foundation models trained and optimized primarily for short clips generally struggle to generalize to extended cinematic durations. As shown in Fig. 2(a), when Wan2.2-T2V-A14B [65] is directly applied to a 30-second continuous-motion prompt, its generation quality noticeably deteriorates compared with the short-form output, exhibiting spatial blurring, reduced motion amplitude, repetitive background patterns, and near-static dynamics. This suggests that strong short-clip synthesis capability does not necessarily imply stable long-form generation. Secondly, **cross-shot semantic consistency**: Multi-shot generation requires characters, objects, and scenes to remain visually consistent and semantically identifiable across discrete shots;

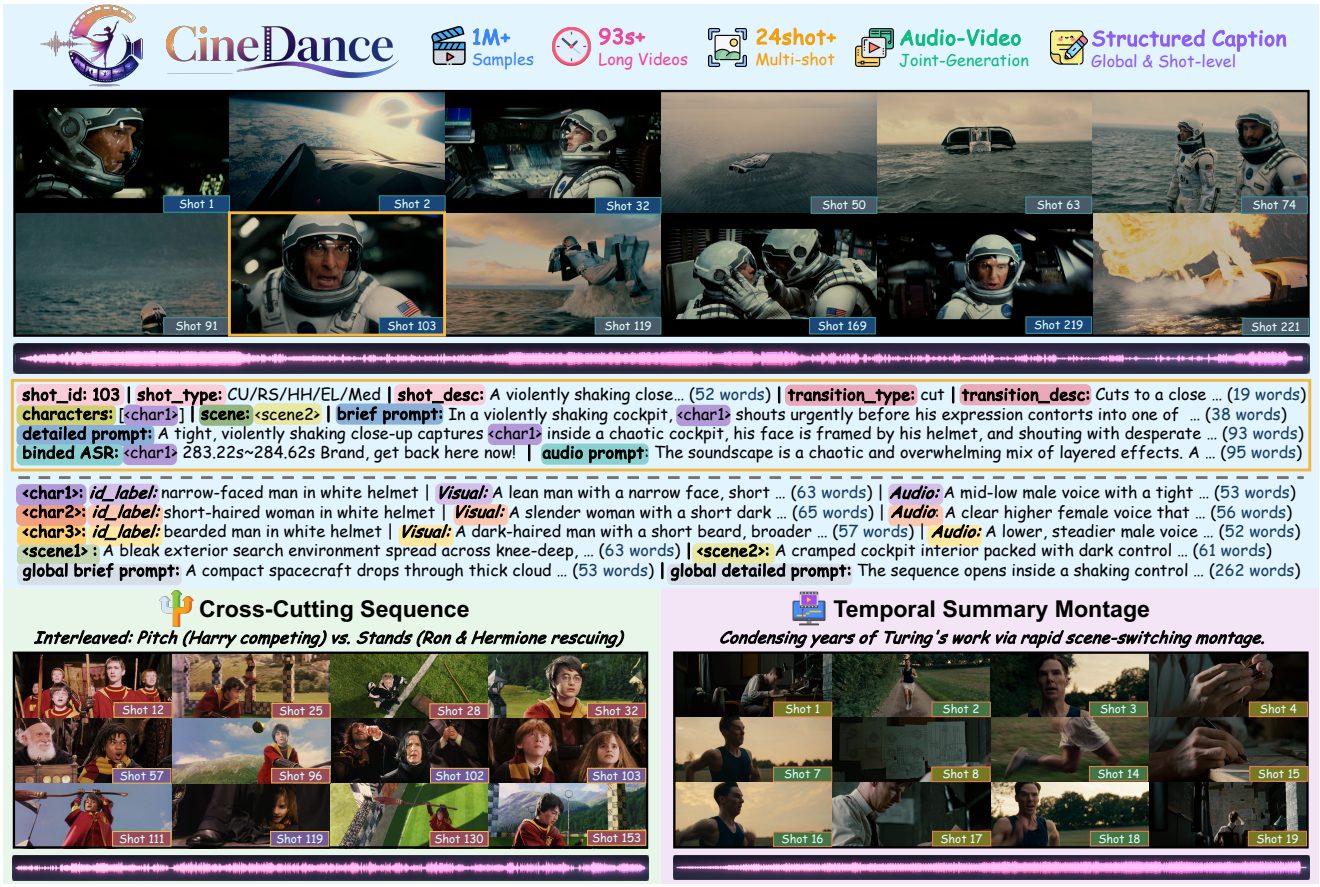


Fig. 1: CineDance-1M features **1M** unprecedented long-form (**92.8** s) and multi-shot (**24.2** shots) **audio-video** sequences (above), paired with **hierarchical structured captions** for both modalities. Compared with typical Text-To-Video (T2V) datasets, it encompasses **diverse narrative structures** (below), meeting the growing demand for cinematic, narrative-driven joint generation.

joint audio-video generation further demands audio and audio-visual consistency. Although dedicated multi-shot methods such as CineTrans [77] can introduce cinematic shot transitions, they still struggle with cross-shot entity consistency even within relatively short clips, as illustrated in Fig. 2(b). For a prompt describing the same character and scene recurring across multiple shots, CineTrans can generate plausible shot transitions, yet may fail to preserve the character's identity and state, as well as scene consistency.

These challenges further expose ***two fundamental gaps*** in the current open research ecosystem. The first is a **foundation-model gap**. Although recent T2V and T2AV foundation models achieve impressive short-form generation quality, they still suffer from fragmented capabilities and limited responsiveness to multi-shot prompts when extended to long-form multi-shot audio-video generation. As shown in Fig. 3(a), representative open video-only models and native audio-video models exhibit substantial degradation when transferring from 5-second generations to 30-second generations. Except for motion smoothness, which remains relatively high partly because the generated videos often become less dynamic, most other dimensions decrease substantially, including visual quality, audio quality, prompt alignment, and audio-video synchronization, consistent with the observation in Fig. 2. This suggests that current foundation models still struggle to maintain holistic audio-video quality over extended cinematic sequences. The second is a **data-and-benchmark gap**. *For training*, existing video datasets are typically limited by several factors: short average clip duration, weak narrative complexity, coarse annotation granularity, or the lack of native audio tracks. Recent efforts such as LVD-2M [81] and MiraData [30] mark important progress toward longer-duration and multi-shot video data, yet they still lack native audio tracks and shot-level dual-modal dense annotations. Moreover, their narrative complexity remains somewhat limited due to their parsing and curation strategies. As summarized in Fig. 3(b), exist-

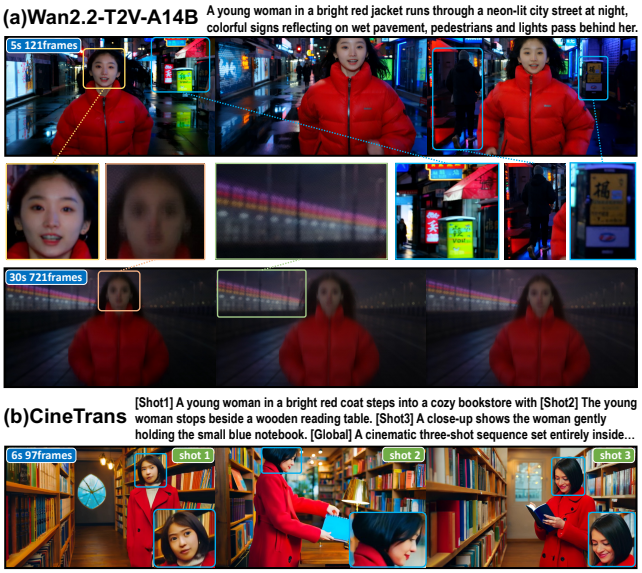
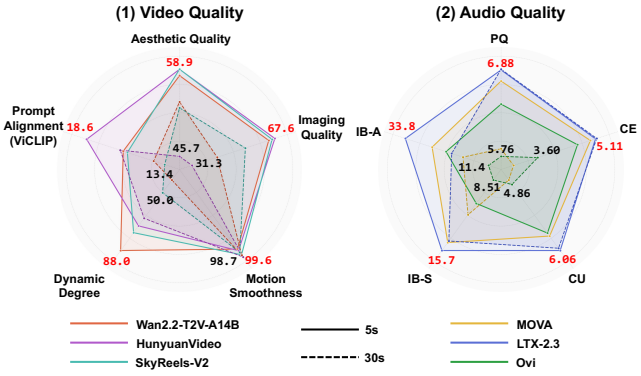
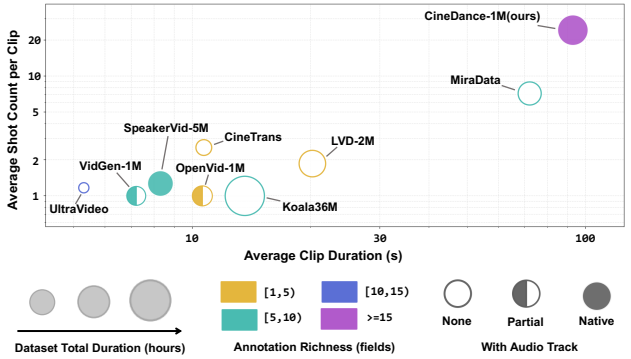


Fig. 2: Diagnostic examples illustrating two core challenges in multi-shot long-form generation. (a) Wan2.2-T2V-A14B produces plausible 5-second results but degrades at 30 seconds, with heavy spatial blurring, low-amplitude motion, and near-static dynamics. (b) CineTrans generates three-shot transitions, but character identity is poorly preserved across shots. *Shortened prompt summaries are shown for readability.*



(a) Quality comparison between 5s and 30s generated videos from popular foundation models.



(b) Dataset comparison in scale, average duration, shot count, annotation richness, and audio availability.

Fig. 3: (a) Representative T2V and joint audio-video foundation models show clear quality degradation when extended from 5s to 30s. (b) Existing datasets remain limited in average clip duration, narrative complexity, annotation granularity, and audio-track availability.

ing datasets therefore provide limited supervision for long-form multi-shot T2AV generation. *For evaluation*, a similar limitation exists. Most previous benchmarks focus on single-shot video-only evaluation, as shown in Tab. 1. Although recent multi-shot benchmarks such as MSVBench [58] begin to evaluate cross-shot narrative coherence, they remain video-centric, relatively small in scale, and do not provide unified joint audio-video evaluation. Together, these data and benchmark limitations motivate CineDance-1M, a large-scale structured dataset for multi-shot T2AV generation, and CineBench, a unified benchmark for evaluating long-form multi-shot audio-video narratives.

To bridge this gap, we introduce **CineDance-1M**, the first large-scale T2AV dataset designed to catalyze the paradigm shift towards multi-shot long-form generation. Derived from premium cinematic media, it comprises **one million** sequences (averaging 92.8 s and 24.2 shots) at a minimum 1080p resolution, as shown in Fig. 1. Crucially, it provides the first **natively structured, configurable, and shot-wise annotations for dual modalities**. The exceptional fidelity of CineDance-1M stems from a rigorous three-stage curation pipeline, which encapsulates our core contributions: **1) Diverse Sourcing and Comprehensive Cleansing.** We collect diverse 1080p videos to guarantee pristine generative priors. Following rigorous spatio-temporal cropping and subtitle removal, we comprehensively evaluate visual quality, audio fidelity, and cross-modal alignment to compile a structured metadata dictionary for versatile downstream filtering (Sec. 3.1). **2) Film-Theory-Inspired Narrative Parsing.** Moving beyond conventional scene-cutting, we introduce a state-based, bottom-up grouping algorithm. By formalizing cinematic syntax into core parsing and merging rules, we guide Qwen-3.5-27B [85] to assemble discrete TransNetV2 [59] shots into coherent, long-form narratives (Sec. 3.2). **3) Configurable Dual-Modal Annotation.** We propose a hierarchical captioning paradigm using anchor tokens to rigidly bind global subject definitions with granular shot-wise references. We deploy specialized MLLMs (Qwen3.5-35B-A3B [85] for video and Qwen3-Omni-30B-A3B [82] for audio) using a task-decomposition strategy, explicitly enabling cross-modal binding while reducing hallucinations (Sec. 3.3). To systematically evaluate complex cinematic

Table 1: Comparison of representative generation benchmarks (one per category from Sec. 2.4). ✓/❍/✗ denote explicit/partial/absent coverage. AV-Sync: audio-video synchronization or cross-modal alignment; MultiShot: multi-shot structure; Narr.Cont.: cross-shot narrative continuity; Struct.Prompt: structured shot-level prompt.

Benchmark	Year	Size	Video	Audio	AV-Sync	MultiShot	Narr.Cont.	Struct.Prompt
VBench-series [25, 26]	2024	1600	✓	✗	✗	❍	✗	✗
FETV [41]	2023	619	❍	✗	✗	✗	✗	✗
T2V-CompBench [61]	2025	1400	❍	✗	✗	✗	✗	❍
AVGen-Bench [101]	2026	235	✓	✓	✓	✗	✗	✗
MSVBench [58]	2026	20	❍	✗	✗	✓	✓	✓
CineBench (ours)	2026	1000	✓	✓	✓	✓	✓	✓

synthesis, we introduce **CineBench**, a comprehensive evaluation suite featuring difficulty-stratified prompt tiers and novel annotation-grounded metrics designed to accurately measure cross-shot narrative continuity and joint audio-video alignment. To verify the effectiveness of CineDance-1M, we extend the LTX-2.3 model [16] to **CineDance**, establishing a strong baseline that exhibits robust capabilities in multi-shot, long-form joint audio-video generation. In summary, our core contributions are fourfold:

1) Addressing the critical scarcity of narrative-driven data, we introduce CineDance-1M, the first large-scale, 1080p dataset dedicated to multi-shot long-form joint audio-video generation, effectively bridging the gap between single-shot clips and complex cinematic synthesis.

2) Empowered by a highly automated, structure-aware processing pipeline, our approach integrates film-theory-inspired narrative parsing with hierarchical dual-modal captioning to guarantee premium caption quality, unprecedented semantic density, and precise cross-modal alignment.

3) To systematically quantify generative capabilities in this new paradigm, we establish CineBench, a comprehensive evaluation suite that stratifies benchmark instances by theme and complexity-based difficulty, also employing a robust six-dimensional metric system for holistic assessment.

4) By fine-tuning the LTX-2.3 backbone on our dataset, we establish CineDance as a robust baseline that demonstrates significant superiority over existing models in maintaining spatio-temporal consistency, identity preservation, and precise cross-modal synchronization.

2 Related Works

2.1 Open-Source Dataset

The evolution of generative models has been fundamentally propelled by the availability of large-scale datasets. Early pioneering collections such as HowTo100M [49], WebVid-10M [3], InternVid [71], and Panda-70M [8] prioritized massive scale and highly diverse web sourcing, laying a crucial foundation that significantly accelerated the development of early models. Recent datasets including OpenHumanVid [34], Koala-36M [68], OpenVid [51], VideoUFO [70], and UltraVideo [84] utilize advanced structured captioning and higher spatial resolutions to enhance visual fidelity. Concurrently, collections like MiraData [30] and LVD-2M [81] explore long-form video generation. However, constrained by their data sourcing and basic shot-merging strategies, their videos remain relatively short and predominantly single-shot. In the audio-visual domain, large-scale datasets such as VGGSound [7] and SpeakerVid-5M [96] provide valuable cross-modal pairs for general alignment and human-centric generation, with another line of works focusing on highly specialized talking-head scenarios like VoxCeleb [50] and LRS3 [1]. Despite their respective contributions, the aforementioned open-source datasets suffer to varying degrees from critical limitations, including brief durations, single-shot dynamics, the absence of the acoustic modality, and a lack of configurable structured annotations.

2.2 Foundation Models and Joint Audio-Video Generation

Video foundation models. Recent video foundation models have substantially advanced open-domain text-to-video (T2V) and image-to-video (I2V) generation. Representative systems such as HunyuanVideo [33, 73], Wan [65], SkyReels [5], CogVideoX [87] and OpenSora [37] improve visual fidelity, motion quality, and prompt following through stronger diffusion backbones, large-scale training data, and post-training strategies. These models provide powerful short-form visual priors, but most remain video-only and are optimized for short clips rather than structured multi-shot narratives. When directly extended to long-form cinematic generation, they often suffer from temporal degradation,

weakened motion dynamics, and cross-shot identity or scene drift, indicating that strong short-clip synthesis does not automatically translate into long-form narrative control.

Joint audio-video generation. Recent works further explore audio-video generative modeling beyond silent videos. MMAudio [9] synthesizes synchronized audio from video and optional text conditions via multimodal joint training. JavisDiT introduces a joint audio-video diffusion transformer with hierarchical spatio-temporal synchronization priors [39], while Ovi [43] models audio and video as a unified generative process through twin DiT backbones and blockwise cross-modal fusion. Other systems such as Harmony [20], UniVerse [66], and LTX-2.3 [16] further study synchronized speech, sound effects, and video generation in increasingly unified frameworks. Despite this progress, existing works are still mostly evaluated on short clips or local audio-video alignment, and rarely address long-form multi-shot cinematic narratives that require recurring speaker binding, voice timbre consistency, ambient sound continuity, and shot-level structured control.

2.3 Multi-Shot Long-Form Video Generation

When extending single-shot foundation models to multi-shot scenarios, prevalent strategies involve employing masked attention mechanisms [31, 55, 77, 79] to explicitly isolate and distinguish distinct shots, alongside structural modifications like Rotary Position Embeddings (RoPE) [60] to enforce temporal boundaries [67, 86]. Transitioning to long-form generation, shot-by-shot paradigms [44, 92] have been extensively explored to synthesize individual segments within a single forward pass. Aimed at efficiently modeling global consistency under strict computational budgets, sparse attention patterns are actively integrated [29, 47], while emerging autoregressive frameworks [28, 44, 90] have concurrently garnered widespread attention for their ability to generate continuous sequences. Furthermore, to explicitly correlate distinct shots and maintain long-term visual consistency and semantic coherence, keyframe-conditioned generation [2, 53, 54, 76, 79, 80, 93, 97, 99] is widely utilized, frequently coupled with Vision-Language Models (VLMs) or Multimodal Large Language Models (MLLMs) to orchestrate narrative progression. More generally, memory-driven approaches [28, 44, 45, 79, 92] preserve critical feature representations from preceding shots, leveraging them as conditional anchors to seamlessly guide subsequent generation steps. Despite these advances, existing approaches still largely rely on external structural priors to impose multi-shot organization at inference time.

2.4 Benchmarks for Video and Audio-Video Generation

1) *General video generation evaluation.* Benchmarks such as VBench/VBench++ [25, 26, 98] evaluate video generation through hierarchical dimensions including visual quality, temporal consistency, motion, subject consistency, and prompt alignment. EvalCrafter [40], VideoScore [18, 19], FETV [41], and AIGCBench [13] further introduce diverse prompts, human-aligned scoring, fine-grained prompt taxonomies, or controllability-oriented metrics. However, they mainly target short-form visual generation rather than long-form multi-shot audio-video narratives. ***2)*** *Compositional and prompt-centric evaluation.* T2V-CompBench [61] evaluates compositional prompt following, including attribute binding, spatial relations, actions, interactions, and numeracy. Other benchmarks study specific capabilities such as text rendering, temporal metamorphosis, motion perception, video dynamics, or real-user prompt distributions [15, 36, 38, 69, 89]. These diagnostics are valuable, but mostly remain short-form or single-shot and lack cross-shot narrative evaluation. ***3)*** *T2AV evaluation.* TAVGBench [46] and AVGen-Bench [101] evaluate text-to-audio-video generation with emphasis on cross-modal alignment, unimodal quality, and semantic controllability. MTAVG-Bench [100], VidAudio-Bench [94], and VABench [24] further cover multi-speaker dialogue, video-to-audio, video-text-to-audio, or synchronous-audio video generation. They advance audio-video evaluation, but mainly focus on local AV alignment or speech coherence rather than long-form cinematic narratives. ***4)*** *Multi-shot and narrative evaluation.* MSVBench [58] evaluates multi-shot generation with hierarchical scripts, reference images, and cross-shot metrics; EntityBench [17] studies long-range consistency of characters, objects, and locations; and MuSS [91] targets multi-shot subject-to-video narrative evaluation. These benchmarks approach narrative generation, but remain largely video-centric or subject-to-video oriented.

3 Curating CineDance-1M for Cinematic Audio-Video Generation

While recent datasets predominantly focus on the massive scale of discrete short clips, CineDance-1M prioritizes high-fidelity, narrative-driven cinematic sequences. As intuitively outlined in Fig. 4, our rigorous data curation pipeline consists of three core stages: ***1)*** Data Preparation and Quality Assessment (Sec. 3.1), ***2)*** Bottom-Up Narrative Sequence Parsing via State-Based Shot Grouping (Sec. 3.2), and ***3)*** Configurable Structured Dual-Modal Annotation (Sec. 3.3). To rigorously validate the

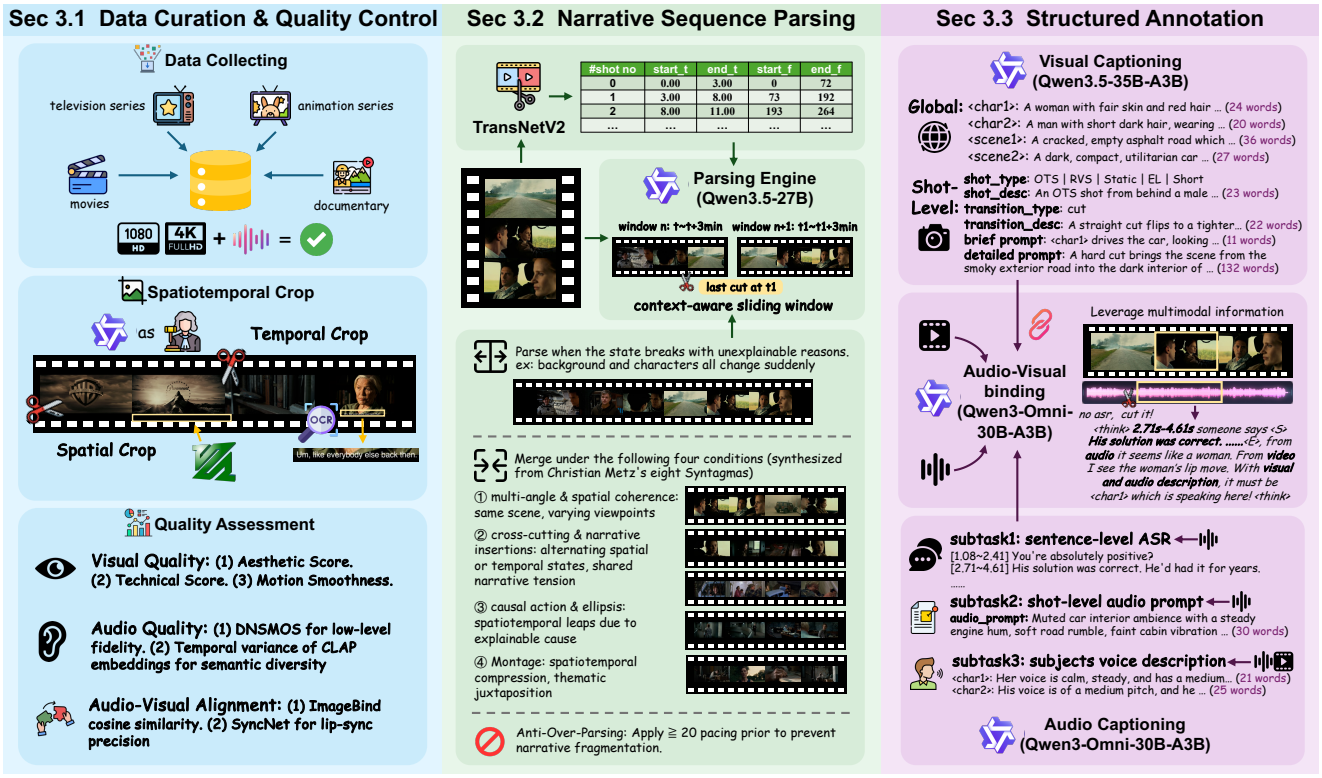


Fig. 4: CineDance-1M curation pipeline consists of three main stages: ***1)*** Data Preparation and Quality Assessment (Sec. 3.1), which encompasses data sourcing, spatiotemporal cropping, and quality assessment; ***2)*** Multi-shot Narrative Parsing (Sec. 3.2); and ***3)*** Hierarchical and Configurable Dual-modal Annotation (Sec. 3.3).

Table 2: Manual visual artifact audit on 500 random short clips per corpus. A clip is counted as non-compliant if it contains at least one residual artifact targeted by our filtering stage.

Corpus	#Clips	#Artifacts	Rate
Koala-36M [68]	500	187	37.4%
CineDance-1M (ours)	500	14	2.8%

pipeline's efficacy and optimality, and to ensure the exceptional quality of our dataset, we conduct comprehensive benchmarking across various implementations with in-depth discussions.

3.1 Data Preparation and Quality Assessment

Source data collection. The raw video corpus of CineDance-1M is constructed from two principal sources: ***1)*** we extract source data from widely adopted public datasets, including MiraData [30], LVD-2M [81], and Koala36M [68], applying stringent duration constraints and multi-shot filtering; and ***2)*** we augment this corpus by adhering to the established data collection pipelines of SkyReels-V2 [5] and OpenHumanVid [34]. Subsequently, we conduct a thorough manual curation process to exclude low-quality samples and potentially sensitive content. The finalized source set consists of **45,181 raw videos**, each with a minimum spatial resolution of 1080p and a cumulative duration exceeding **31,530 hours**. The inherent narrative coherence, sophisticated visual composition, and high-fidelity acoustic characteristics of these videos render this corpus particularly well-suited for advancing research on joint audio–video generative models.

Coarse-to-fine spatiotemporal filtering. To remove black borders and overlaid text, we apply a coarse-to-fine spatiotemporal filtering and cropping pipeline to the raw videos. The first two steps operate at the raw-video level in a coarse manner and serve as preprocessing for the narrative parsing stage in Sec. 3.2: they reduce the interference of letterboxing, subtitles, title cards, openings, and end credits on shot boundary detection and semantic parsing. After raw-video parsing (Sec. 3.2), we further perform fine clip-level verification to further improve the quality of the final clips. The pipeline is divided into three steps:

1) Coarse Spatial Cropping. To remove global

Table 3: Curation funnel of CineDance-1M. Duration denotes the retained temporal coverage after each stage. Quality assessment computes and stores metadata only, without filtering sequences by default.

Stage	Unit	Count	Duration	Operation
Raw collection (Sec. 3.1)	Videos	45,181	32.8K hr	Collect 1080p source videos with provenance records
Spatiotemporal pre-filter (Sec. 3.1)	Videos	44,579	32.5K hr	Remove subtitles, borders, title cards, openings, and credits
Shot detection (Sec. 3.2)	Shots	25,899,474	32.5K hr	Detect atomic shot boundaries using TransNetV2
Narrative parsing (Sec. 3.2)	Seq.	1,201,912	32.5K hr	Group shots into state-consistent narrative sequences
Sequence pruning (Sec. 3.1)	Seq.	1,079,382	28.6K hr	Remove sequences that are single-shot or shorter than 10s
Post-verification (Sec. 3.1)	Seq.	1,021,657	26.3K hr	Reject temporally invalid or artifact-contaminated sequences
Quality assessment (Sec. 3.1)	Seq.			Metric-based quality assessment without actual pruning

subtitles and letterboxing, we sample segments at the 25%, 50%, and 75% milestones at 2 FPS. Using `EasyOCR` [27] for text detection and `FFmpeg` for black border detection, we compute the optimal bounding boxes and perform a global spatial crop.

2) MLLM-Guided Temporal Truncation. We feed the initial and final segments of duration $t = \max(5\ \text{min}, 0.1L)$, where L denotes the total video length, to MLLMs to accurately detect and remove non-narrative introductory and concluding content. If the model judges an entire queried segment as introductory or concluding, we iteratively move inward and query the next segment of the same duration until a narrative boundary is identified.

3) Fine Clip-Level Verification. To prevent under- or over-cropping caused by dynamic aspect ratios (e.g., IMAX transitions) and avoid missing intermittent text, we additionally perform clip-level OCR and black border detection after the raw video parsing detailed in Sec. 3.2. We discard clips whose average frame-level text-area ratio exceeds a predefined threshold.

Metric-based quality assessment. To systematically ensure the dataset's fidelity and enable versatile downstream filtering, we evaluate each final video clip across three core dimensions following raw video parsing (detailed in Sec. 3.2): ***1) Video Quality.*** We systematically evaluate *Aesthetic Quality* and *Technical Score* (DOVER [74]), along with *Motion Smoothness* (AMT [35]) at both the shot and video levels. ***2) Audio Quality.*** Apart from low-level signal fidelity (DNS-MOS [56]), we assess acoustic richness by measuring the temporal variance of CLAP [78] embeddings, both evaluated at the video level. ***3) Audio-Video Alignment.*** We quantify cross-modal consistency via ImageBind [14] for global audio-video alignment, and employ SyncNet [10] to measure lip-synchronization.
Instead of applying hard pruning with fixed thresholds, we store all quality scores as metadata, enabling users to flexibly construct task-specific subsets and control the quantity–quality trade-off.
Visual artifact audit. To verify that the coarse-to-fine filtering stage removes the visual artifacts it targets, we conduct a manual audit against Koala-36M, a recent large-scale and high-quality T2V dataset. We randomly sample 500 clips from CineDance-1M and Koala-36M, respectively, and ask three trained annotators to independently inspect residual artifacts, including burnt subtitles or logos, letterboxing, watermarks, network overlays, title-card or end-credit frames, screen recordings, transition effects, still-frame holds, and near-uniform fill frames. A clip is counted as non-compliant with our artifact-removal criteria if any of these artifacts appears; annotator disagreements are resolved by joint review. Although CineDance-1M clips are substantially longer on average than Koala-36M clips, CineDance-1M reduces the non-compliance rate from 37.4% on Koala-36M to 2.8%, a 13.4× reduction, as shown in Tab. 2. The remaining failures are mainly intermittent watermark frames, which appear only briefly and are difficult to remove without overly aggressive cropping.

3.2 Bottom-Up Narrative Sequence Parsing via State-Based Shot Grouping

Narrative sequences definition. Existing long-video datasets [30, 75, 77, 81] typically rely on low-level visual cues (e.g., pixel similarity) for video segmentation. However, this rigid strategy disrupts inherent narrative continuity, as physical scene transitions do not necessarily signal narrative breaks (e.g., a continuous conversation that moves from indoors to outdoors). To preserve cinematic cohesion, we formally define a ***Narrative Sequence*** as a continuous flow of diegetic time and causality that aligns with real-world chronological progression, regardless of spatial shifts. Within this unit, all character and environmental state changes are strictly driven by continuous logical events and clear causal explanations (exceptions are detailed in the following sections).
State-based parsing rule. Based on our previous definition, we conceptually model the cinematic flow as a transition of semantic states, $S = (\mathcal{T}, \mathcal{P}, \mathcal{C}, \mathcal{E})$, repre-

Table 4: Narrative parsing ablation on a 12-film, 25.2-hour human reference set. "Align." measures boundary alignment with TransNetV2 shot cuts. **Bold** indicates the best machine result, * denotes the selected implementation, and † marks the empirical basis of the ***20s anti-fragmentation prior***.

Source / Backbone	Strategy	#Seq.	Mean (s)	Min (s)	$<$ 20s (%) ↓	Align. (%) ↑	F1 (%) ↑
Human GT	Reconciled annotation	842	108.3	***18.4***†	0.6	–	–
Qwen3.5-35B-A3B	Direct timestamp parsing	1,748	51.6	2.1	36.2	53.2	51.7
	Bottom-up shot grouping	1,347	67.6	5.8	22.3	**100.0**	66.9
Qwen3.5-122B-A10B	Direct timestamp parsing	1,342	67.4	3.7	24.3	64.4	62.7
	Bottom-up shot grouping	1,082	84.1	9.5	11.7	**100.0**	77.4
Qwen3.5-27B	Direct timestamp parsing	1,078	84.2	6.4	16.2	70.6	73.4
	Bottom-up shot grouping*	873	104.4	17.2	**3.1**	**100.0**	**88.4**

senting time, place, characters, and events. We establish the ***first-principle parsing rule***: a narrative sequence terminates when an *unexplainable state break* occurs (*e.g.*, abrupt shifts in character identity, total replacement of the character ensemble, or complete change of the environmental layout).

Cinematic-theory-guided merging rules. However, while a state break is a necessary condition for a narrative boundary, it is not a sufficient one. As illustrated in Fig. 1, cross-cutting shots and montage techniques [12] are frequently employed in cinematic storytelling. To address this, inspired by Christian Metz's *Grande Syntagmatique* [48], we adapt and filter classical film syntax into ***four merging rules*** to instruct the MLLM not to cut during these cinematic exceptions:

1) Multi-Angle and Spatial Coherence: Shots exhibiting $\Delta\mathcal{P}$ = camera-angle-only while maintaining $\Delta\mathcal{T} \approx 0$ and a unified ongoing event ($\mathcal{E}_{continuous}$) are merged, representing continuous action within the same physical environment explored from different viewpoints.

2) Cross-Cutting and Narrative Insertions: Sequences rapidly alternating between two distinct spatial states ($\mathcal{P}_A \neq \mathcal{P}_B$) or temporal states ($\mathcal{T}_A \neq \mathcal{T}_B$) are merged if they are bound by a unified causal tension ($\mathcal{E}_{shared}$). This encapsulates simultaneous alternating (e.g., a phone call) and sandwich insertions (e.g., a brief flashback).

3) Causal Action and Ellipsis: Shots exhibiting significant spatial leaps ($\Delta\mathcal{P} \neq 0$) and temporal gaps ($\Delta\mathcal{T} > 0$) are merged if the event $\mathcal{E}_{i+1}$ is a direct, explainable causal consequence of $\mathcal{E}_i$ (e.g., a gun fired indoors causing a window to shatter outdoors).

4) Montage: Sequences exhibiting disjointed shifts in time and space without explicit micro-causal links are merged if they are unified by a macro-level thematic or emotional arc ($\mathcal{E}_{theme}$), such as a montage of a city waking up or a training sequence.

Empirical anti-fragmentation rule. Classical continuity editing emphasizes preserving coherent spatial and temporal relations across shots [4]. Motivated by this principle, we introduce an explicit *anti-fragmentation rule* to prevent MLLMs from splitting a cohesive narrative segment into meaningless short clips. As shown in Tab. 4, our reconciled human parsing results give a minimum sequence duration of 18.4 seconds. We therefore round this empirical lower-tail value to **20** seconds and use it as a soft minimum-duration prior during MLLM parsing. When a candidate boundary would produce a segment shorter than this threshold, the parser is instructed to keep expanding the temporal window unless a clear character, scene, or event-state break is observed.

Bottom-up sliding inference. We first formulate our parsing, merging, and anti-fragmentation rules as structured prompts. Based on comprehensive benchmarks of the Qwen3.5 family [85], we then adopt the Qwen3.5-27B model as our core parsing engine, equipped with two core mechanisms:

1) Bottom-up shot indexing. Directly prompting an MLLM to localize narrative boundaries across long videos causes severe temporal hallucination. Therefore, we establish shots (extracted via `TransNetV2` [59]) as the atomic units of our dataset. Given the start and end timestamps of each shot, the MLLM is then tasked solely with outputting discrete shot indices as parsing boundaries based on our prompts, significantly reducing timestamp hallucinations and improving parsing accuracy.

2) Context-aware sliding window inference. Given the massive duration of raw videos and MLLM context window constraints, processing entire movies simultaneously is intractable. To resolve this, we implement a context-aware sliding window strategy, as summarized in Alg. 1. Specifically, starting from the narrative onset determined by temporal truncation in Sec. 3.1, we sequentially feed approximately $\Delta = 3$ minute windows into the MLLM. Since parsing windows must align with atomic shot boundaries, each window endpoint is selected by GetWindowEnd, which returns

Algorithm 1: Bottom-Up Narrative Sequence Parsing

Input: Raw video V; shot detector D; parsing engine M; reference window size Δ; rule set $R = \{R_p, R_m, R_a\}$
Output: Boundary list B for cropping narrative sequences

Function `GetWindowEnd(S, l, Δ)`:
  Find the shot index r such that the boundary-aligned window $\{s_l, \ldots, s_r\}$ has duration closest to Δ;
  return r;
Function `ExtendWindowEnd(S, l, r, Δ)`:
  $r \leftarrow$ `GetWindowEnd(S, l,` $\mathrm{Dur}(S, l, r) + \Delta$`)`;
  return r;
Detect atomic shots $S = \{s_i\}_{i=1}^{n}$ using $D(V)$;
Initialize boundary list $B \leftarrow \emptyset$ and window start $l \leftarrow 1$;
while $l \leq n$ **do**
  $r \leftarrow$ `GetWindowEnd(S, l, Δ)`;
  $C \leftarrow \emptyset$;
  while $C = \emptyset$ **and** $r < n$ **do**
    Build window $W = \{s_l, \ldots, s_r\}$ with shot indices, timestamps and sampled frames;
    Query M with W and R to obtain candidate boundaries C;
    if $C = \emptyset$ **then**
      $r \leftarrow$ `ExtendWindowEnd(S, l, r, Δ)`;
  if $C = \emptyset$ **then**
    $B \leftarrow B \cup \{n + 1\}$;
    break;
  $B \leftarrow B \cup C$;
  $l \leftarrow \max(C)$;
return B

the shot boundary whose temporal span is closest to Δ. The last detected boundary in the current window serves as the starting point of the next window. If the MLLM outputs no boundary, indicating that the current window is likely a continuous narrative sequence, we call ExtendWindowEnd to expand the current window by another reference duration Δ, again snapping the endpoint to the nearest shot boundary. This process is repeated until a valid boundary is identified or the end of the video is reached. This dynamic windowing strategy preserves global narrative coherence while maintaining high computational efficiency.

3.3 Configurable Structured Dual-Modal Annotation

Configurable anchored design. Given the diverse prompt formats and input conditions in current multi-shot long-form generation, we introduce an anchor token mechanism to ensure our annotations are highly adaptable for future explorations. Specifically, we define a global character list ($[\langle \texttt{char}_1 \rangle, \ldots, \langle \texttt{char}_N \rangle]$) and a scene list ($[\langle \texttt{scene}_1 \rangle, \ldots, \langle \texttt{scene}_M \rangle]$) for the entire narrative sequence. Subsequently, the dual-modal shot-level prompts (detailed in the following sections) explicitly refer back to these anchor tokens. This mechanism significantly enhances prompt configurability while establishing robust connections across global-to-shot and shot-to-shot levels.

Table 5: Audio-only speaker diarization accuracy on our 100-clip cinematic-dialogue benchmark, measured by permutation-invariant segment-level matching averaged across clips. The results motivate decoupling ASR from global character-level speaker binding.

Backbone / Family	Implementation	Acc. (%)
Specialized diarization	Pyannote-3.1	62.7
	DiariZen	63.1
Closed-source MLLM	Gemini-2.5-Pro	82.8
	Gemini-3.1-Pro	87.4
Qwen3-Omni 30B-A3B	Whole-clip prompting	56.4
	Sliding-window prompting	83.1

Shot-wise structural visual captioning. For each shot, we extract a five-dimensional set of shot attributes (scale, angle, movement, narrative function, duration category) alongside specific shot transition types [4]. Furthermore, we supplement these categorical labels with a dedicated shot description and a transition description to facilitate advanced training control. Crucially, we define a localized character list and an active scene for each shot, indicating the specific characters present and the environment where the shot takes place. Both the brief and detailed visual prompts then explicitly refer to these anchor tokens. To ensure global consistency and reduce model hallucinations, the global character and scene lists alongside all shot-level visual annotations are generated in a single pass of Qwen3.5-35B-A3B [85].

Granular audio diarization and captioning. Based on comprehensive benchmarks of various implementations, we decompose the audio annotation process into three manageable sub-tasks via Qwen3-Omni-30B-A3B [82]: ***1)*** extracting sentence-level Automatic Speech Recognition (ASR) segments, ***2)*** generating shot-level audio prompts, and ***3)*** creating character voice descriptions. This design is motivated by **three observations**. ***First***, speech is a relatively special component of audio annotations. Unlike ambient sound, music, or sound effects, dialogue is often organized in different prompt formats by different joint audio-video generation foundation models [16, 43, 63]. We therefore store ASR as an independent annotation term, which makes the

Table 6: Comparison of representative video datasets. CineDance-1M pioneers *multi-shot (24.2 shots)*, *long-form (92.8 s)* T2AV generation with unprecedented structural complexity at 1080p. As the only 1M-scale dataset providing dense shot-level audio-video annotations (averaging $>$ 6,400 words), it uniquely enables granular cross-modal control.

Dataset	Visual	Clip Structure			Audio		Text	Scale		Time
	Res.	Avg. dur.	Avg. shots	Shot caps.	Audio	Audio ann.	Cap. len.	Total dur.	Clips	Year
HowTo100M [49]	240p	3.6s	1	None	None	None	4	134.5Khr	136M	2019
HD-VILA-100M [83]	720p	13.4s	1	None	None	None	32.5	371.5Khr	103M	2022
Koala-36M [68]	720p	13.6s	1	None	None	None	202.3	137Khr	36M	2024
VIDGEN-1M [62]	720p	10.6s	1	None	Partial	None	89.3	2.9Khr	1M	2024
MiraData [30]	720p	72.1s	7.15	None	None	None	319	6.6Khr	330K	2024
LVD-2M [81]	720p	20.2s	1.86	None	None	None	88.8	14.6Khr	2.1M	2024
OpenHumanVid [34]	720p	4.6s	1	None	All	None	99.7	12Khr	16M	2025
OpenS2V-5M [88]	720p	5.6s	1	None	Partial	None	312.06	5.8Khr	3.75M	2025
UltraVideo [84]	4K/8K	5.3s	1.17	None	No	None	824.3	62hr	42K	2025
OpenVid-1M [51]	720p	7.2s	1	None	Partial	None	126.5	2.1Khr	1M	2025
CineTrans [77]	720p	10.7s	2.53	2	None	None	250.78	752hr	252K	2025
SpeakerVid-5M [96]	1080p	8.3s	1.27	None	All	ASR	20.69	11.6Khr	5.07M	2025
Ours	**1080p**	**92.8 s**	**24.2**	**22**	**All**	**Structured**	**6496.3**	**26.3Khr**	**1M**	**2026**

Table 7: Multimodal ASR segment-to-character binding accuracy on the 100-clip human-labeled benchmark. Bold indicates our production configuration, and italic indicates the unaffordable closed-source ceiling.

Method	Regime	Acc. (%)
Open-source omni-MLLMs		
Qwen2.5-Omni-7B-Instruct	Whole-clip	58.4
Qwen2.5-Omni-7B-Instruct	Windowed	78.6
Qwen3-Omni-30B-A3B-Instruct	Whole-clip	67.2
Qwen3-Omni-30B-A3B-Instruct	**Windowed**	**95.4**
Closed-source frontier omni systems		
Gemini-2.5-Flash	Whole-clip	64.7
Gemini-2.5-Flash	Windowed	88.9
Gemini-2.5-Pro	Windowed	92.8
Gemini-3.1-Pro	*Windowed*	*96.3*

dataset more flexible for different downstream training and prompting protocols. ***Second***, we do not perform audio-level speaker-to-character binding during ASR. As shown in Tab. 5, existing systems struggle to maintain usable audio-only speaker diarization accuracy (i.e., above 90%) on long-form cinematic dialogue. Thus, the ASR stage only extracts spoken text, while global speaker-to-character assignment is handled by the multimodal binding stage described in the next section. ***Third***, even strong open-source omni-modal models such as Qwen3-Omni-30B-A3B can exhibit a high hallucination rate when asked to jointly perform multiple audio-related tasks in a single pass. By minimizing individual task complexity, this decomposed strategy mitigates hallucinations and improves annotation quality.

Windowed audio-video identity binding. Given the video clip, audio track, global character anchors, shot-level visual annotations, character voice descriptions, and sentence-level ASR texts, we perform a dedicated cross-modal identity binding step. It serves two purposes. ***1)*** First, it attaches each generated voice description to the correct character anchor ⟨`charX`⟩. ***2)*** Second, it assigns each ASR sentence to the character who speaks it. The first task is relatively easier because explicit character-level visual and acoustic cues are already available. The main challenge lies in resolving the speaker of each sentence in long-form cinematic dialogue, where off-screen speech, shot changes, and overlapping characters frequently introduce ambiguity. To reduce model hallucinations, we filter out non-speech intervals based on ASR outputs and partition the remaining sequence into localized windows that preserve intact shots and complete spoken sentences. Given these localized audio-video windows, together with prior visual annotations and ASR texts, the model performs cross-modal binding to assign each sentence to the corresponding character anchor ⟨`charX`⟩. As shown in Tab. 7, this windowed strategy substantially improves ASR segment-to-character binding accuracy. Compared with Tab. 5, Qwen3-Omni-30B-A3B improves from 83.1% to 95.4%, indicating that visual grounding provides complementary evidence for binding. By discarding non-speech intervals and restricting reasoning to localized audio-video contexts, it also reduces computational overhead, making corpus-scale deployment feasible.

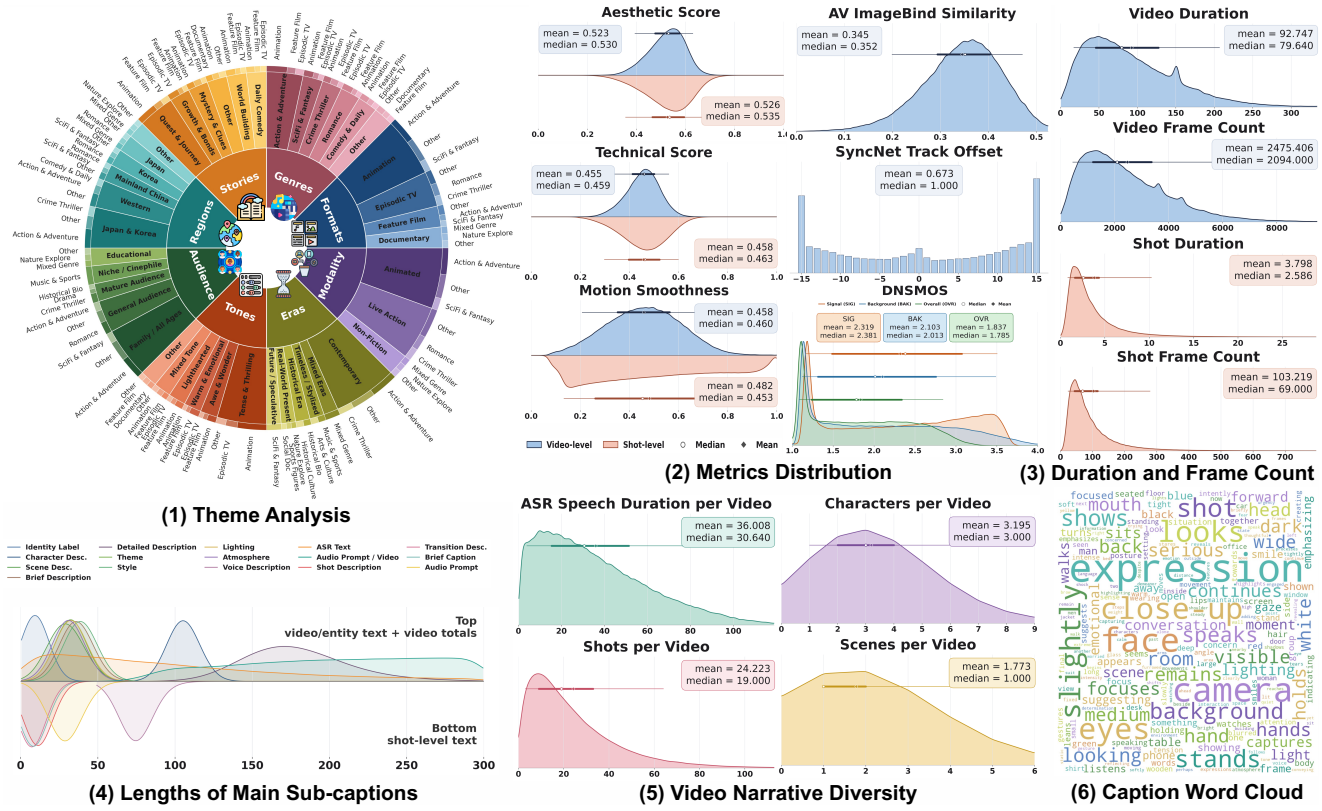


Fig. 5: Statistical overview of the CineDance-1M dataset across multiple dimensions.

3.4 Statistical Comparison and Analysis

Comparison with prominent video datasets. Tab. 6 comprehensively compares CineDance-1M with other prominent datasets. Notably, CineDance-1M is a pioneering audio-video dataset that extends the average sequence duration to an unprecedented 92.8 seconds and encompasses over 24.2 physical shots per sequence. Furthermore, it features highly structured, configurable text prompts for both modalities, providing optimal fine-tuning signals for joint audio-video generation. While recent long-video datasets like MiraData [30] and LVD-2M [81] exist, their low shot counts indicate predominantly single-shot or static single-scene recordings.

Numerical statistics. Fig. 5 details the statistical distributions of CineDance-1M across six dimensions: ***1)*** An eight-dimensional taxonomy (Genre, Format, Region, Modality, Story Logic, Era, Tone, and Audience) ensures high categorical diversity for generalizable generation, constructed with advanced LLM assistance. ***2)*** Consistently high scores in video/audio quality and audio-video alignment validate the dataset's overall fidelity, while enabling researchers to set customizable filtering thresholds. ***3)*** Duration and frame count distributions at both video and shot levels emphasize the dataset's inherently long-form nature, addressing the critical scarcity of high-quality long-duration priors. ***4)*** Prompt length distributions reveal a remarkably dense annotation volume (averaging 6,400 words per video), offering exceptionally fine-grained guidance for controllable synthesis. ***5)*** Video-level distributions of shots, characters, unique scenes, and ASR duration strictly quantify the high structural and intricate narrative complexity of our sequences. ***6)*** A semantic word cloud highlights the core descriptive and cinematic vocabulary embedded within our annotations.

4 CineBench: Advancing Cinematic Evaluation via Hierarchical Benchmark

To systematically assess model capabilities in complex narrative synthesis, we introduce CineBench, a comprehensive evaluation suite designed for multi-shot, long-form joint audio-video generation.

4.1 Task Definition

CineBench evaluates whether a generative model can synthesize a temporally ordered multi-shot sequence from the structured conditions defined in Sec. 3.3. Each

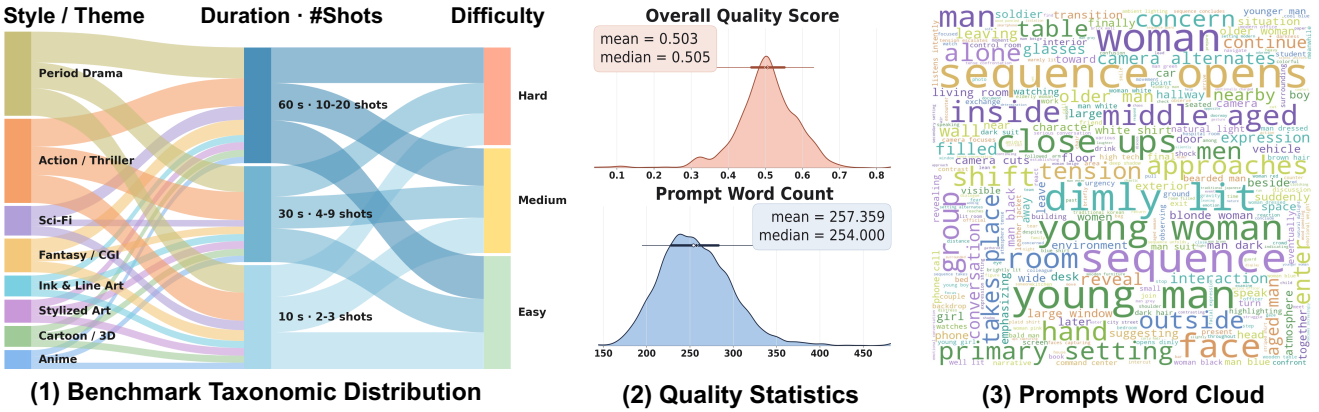


Fig. 6: Comprehensive statistical overview of CineBench, illustrating its diverse taxonomic flow, rigorous quality distributions, and rich semantic vocabulary.

instance follows the same anchor-centric schema as CineDance-1M, but is rendered into model-specific inputs before evaluation. The target output is a coherent long-form sequence that preserves the intended characters, scenes, events, and audio-visual content across shots. Video-only models are evaluated on visual and narrative dimensions, while native audio-video models are additionally evaluated on audio quality and audio-video synchronization.

4.2 Benchmark Construction

Taxonomy and difficulty stratification. To comprehensively evaluate multidimensional generative capabilities, we stratify benchmark instances across three dimensions: ***Theme/Style***, ***Duration/Shot Count***, and ***Generation Difficulty***. Here, *Generation Difficulty* targets the quality challenge within the model's intrinsic capacity boundary, while *Duration/Shot Count* characterizes the capacity boundary itself, i.e., how long a video and how many shots a model can support. Although current methods are still limited in achievable duration and shot count, this design makes CineBench a forward-looking benchmark that remains meaningful as long-form generation models continue to improve.

The difficulty level is deterministically defined by rule-based features: ***1)*** entity complexity, measured by the number of distinct character anchors; ***2)*** scene complexity, measured by the number of distinct scene anchors; and ***3)*** dialogue/audio complexity, measured by speaker-adjusted ASR length. Specifically, for each candidate window, we compute

$$D = n_{\text{char}} + 1.5\, n_{\text{scene}} + 0.4 \log\big(1 + n_{\text{spk}} L_{\text{ASR}}\big),$$

where n_{char}, n_{scene}, and n_{spk} denote the numbers of distinct character, scene, and speaker anchors within the window, and L_{ASR} denotes the cumulative ASR character count. We consider ***three*** *Duration/Shot Count* tiers: 10s with 2–3 shots, 30s with 4–9 shots, and 60s with 10–20 shots. Within each duration tier, Easy, Medium, and Hard labels are assigned by empirical tertile cuts of D, preventing longer windows from being trivially classified as harder. To ensure a balanced evaluation, we uniformly sample high-quality instances across the resulting *Theme/Style* × *Duration/Shot Count* × *Difficulty* grid, ultimately curating a diverse set of **1000** testing cases.

Prompt and condition construction. To serve as a universal evaluation suite, CineBench retains the fine-grained, structured representations of the global and shot-level dual-modal annotations, facilitating seamless adaptation to the diverse input formats required by various methods. Benchmark test cases are carefully sampled from the curated dataset across the aforementioned taxonomy tiers, prioritizing high-quality sequences. Crucially, all sampled textual annotations undergo rigorous manual verification to guarantee absolute correctness and premium visual quality. ***Finally, to strictly prevent data leakage and ensure fair evaluation, all these test cases are permanently removed from CineDance-1M***. A comprehensive statistical overview of CineBench is provided in Fig. 6.

4.3 Comprehensive Evaluation Suite

We assess performance across six dimensions using a robust automated suite: ***1) Video Quality*** measures shot-level *Aesthetic Quality* (aesthetic-predictor-v2-5) [25], *Imaging Quality* (MUSIQ) [32], and *Motion Smoothness* (AMT) [35]. ***2) Audio Quality*** reports global *AudioBox-Aesthetics* (PQ, CE, CU) [64] and speech intelligibility measured by WER/CER using Whisper-

large-v3. ***3) AV Sync*** combines *Sync-C* and *Sync-D* [10] for overall lip-sync with ImageBind *IB-Score* [14]. ***4) Prompt Alignment*** captures dual-granularity video alignment via *ViCLIP* [71] (shot-level) and *VideoScore-v1.1* [19] (video-level), while audio alignment utilizes ImageBind *IB-A Score* [14]. ***5) Narrative Continuity*** introduces annotation-grounded alternatives to address the critical failure of standard frame-level metrics that inappropriately penalize legitimate camera cuts. Specifically, *Identity Continuity* leverages ArcFace [11] clustering matched against declared $\langle \text{char}_k \rangle$ tokens to score cross-shot character consistency, and *Scene Continuity* computes the mean pairwise DINOv2 [52] cosine across shots sharing the same $\langle \text{scene}_k \rangle$ token. ***6) Shot Structure Response*** evaluates whether the generated video responds to the target multi-shot structure by comparing the ground-truth shot partition $\{I_i\}_{i=1}^{N}$ with the detected generated shot partition $\{\hat{I}_j\}_{j=1}^{M}$. We compute

$$S_{\text{seg}} = \frac{1}{2}\left(\frac{1}{N}\sum_i \max_j \text{IoU}(I_i, \hat{I}_j) + \frac{1}{M}\sum_j \max_i \text{IoU}(I_i, \hat{I}_j)\right).$$

$$S_{\text{cnt}} = \frac{\min(N,M)}{\max(N,M)}$$

Here, S_{cnt} measures shot-count agreement, while S_{seg} measures bidirectional temporal overlap between the target and generated shot partitions. The final Shot Structure Response is $\text{SSR} = S_{\text{cnt}}^{0.35} S_{\text{seg}}^{0.65}$.

4.4 Human Validation Protocol

To assess whether CineBench automatic scores reflect human perception, we conduct a randomized and double-blind human study on sampled generated sequences. Each generated video is rated by 10 independent evaluators using a 5-point Likert scale (1 = unusable, 2 = poor, 3 = acceptable, 4 = good, 5 = excellent), with all source models anonymized and presentation order randomized to reduce subjective bias. The scoring rubric follows the same six dimensions as our automatic suite: ***1)*** *Video Quality* evaluates per-frame fidelity, visual artifacts, and motion smoothness; ***2)*** *Audio Quality* measures speech clarity, background sound naturalness, and the absence of audible artifacts; ***3)*** *Audio-Video Synchronization* assesses lip-sync precision for speech segments and semantic alignment between sound events and visual content; ***4)*** *Prompt Alignment* measures faithfulness to the rendered benchmark condition, including characters, scenes, events, dialogue, and sound descriptions; ***5)*** *Narrative Continuity* evaluates cross-shot identity preservation, scene recurrence, object persistence, and ordered event progression; and ***6)*** *Shot Structure Response* evaluates whether the generated video exhibits the intended number of shots, clear shot transitions, and a temporal shot layout consistent with the target structure. We use inter-rater agreement to measure panel consistency and Spearman rank correlation to quantify the alignment between automatic CineBench scores and human judgments. Detailed results analyses are reported in Sec. 6.

5 CineDance: Multi-Shot Long-Form Audio-Video Generation

5.1 Backbone and Joint Audio-Video Formulation

Backbone. We select the state-of-the-art LTX-2.3 [16] as our backbone strictly for its native capability in joint audio-video generation, which naturally aligns with the dual-modal nature of CineDance. Its core design is an asymmetric dual-stream DiT, where a 13B-parameter video branch and a 3B-parameter audio branch process modality-specific latent tokens. The two streams are bridged by a 3B-parameter audio-video cross-attention branch, allowing temporally aligned information exchange between visual and acoustic representations.

Joint audio-video formulation. Given a structured textual condition c, video latent z_v, and audio latent z_a, the backbone learns a conditional joint generative model $p_\theta(z_v, z_a \mid c)$, through a modality-aware latent denoising objective: $\mathcal{L}_{\text{AV}} = \mathbb{E}_t\left[\lambda_v \mathcal{L}_v(z_{v,t}, t, c) + \lambda_a \mathcal{L}_a(z_{a,t}, t, c)\right]$, where $\mathcal{L}_v$ and $\mathcal{L}_a$ denote the video and audio denoising losses, respectively. The coefficients λ_v and λ_a control the relative contribution of the two modality-specific objectives.

5.2 Rethinking Multi-Shot Long-Form Generation: A Conditioning and Output Perspective

As surveyed in Sec. 2.3, the multi-shot generation landscape has been predominantly described through the lens of methodological mechanisms, such as masked attention, autoregressive frameworks, and sparse attention patterns. While this mechanism-centric taxonomy reflects the methodological evolution of existing approaches, it remains a descriptive categorization of how methods operate internally. In this section, we step back and propose a more fundamental taxonomy at a higher level of abstraction, organized by ***input conditions*** and ***output formats***. We group existing methods into the following ***three*** paradigms: ***1) Per-shot generation with memory propagation.*** This paradigm decomposes multi-shot long-form generation

into separate short-video generation tasks and stitches the resulting clips afterward [2, 76, 92, 93, 99]. Cross-shot consistency is mainly maintained through memory propagation, such as carrying latent frames from previous shots into later generations. Keyframe-based methods also fall into this category, since keyframes usually serve as the start and end frames of individual short clips, and the coherence of these keyframes largely determines the coherence of the assembled multi-shot video. This per-shot decomposition tends to produce shots with similar or fixed durations, which may conflict with the desired narrative rhythm and temporal logic. Although memory propagation can mitigate drift, cross-shot consistency remains limited because different shots are still generated separately. ***2) Structure-aware single-forward generation.*** This paradigm requires detailed multi-shot structure as input, including the start and end times of each shot, shot-level prompts, and sometimes finer control signals [47, 55, 67, 77]. The model generates the full video in a single forward process, using shot-aware attention masks, segment-wise prompt assignment, or RoPE-based temporal indexing to separate and align shots. While this avoids post-hoc stitching and provides explicit layout control, it relies heavily on user-specified structure. Shot durations cannot adapt freely to narrative rhythm, imperfect conditions may introduce artifacts, and attention manipulation can still leave visible cross-shot consistency errors. ***3) Reference-to-multi-shot generation.*** Recent closed or commercial systems, such as Seedance 2.0 [57] and SkyReels-V4 [6], introduce an emerging reference-to-multi-shot setting, where the reference image is arranged as a grid of shot-level visual guidance. Importantly, the sub-images in the grid are not start or end frames to be reproduced exactly; they provide high-level visual-semantic guidance for scene, style, and composition. The model interprets these visual cues and automatically organizes the output into multiple shots, without requiring user-specified shot boundaries. Compared with the previous paradigms, this setting allows shot durations to adapt more naturally to the generated narrative and often yields stronger visual consistency in our benchmarking observations. While this paradigm is flexible and user-friendly, it remains less explored in open-source methods. ***Overall***, these paradigms still rely on explicit structural aids, ranging from user-specified shot boundaries to reference-based visual guidance.

5.3 Reference Frames as Training-Time Scaffolds

Observation and Motivation. A direct way to adapt a native T2AV backbone to multi-shot generation is to train it with structured shot-level prompts. However, this strategy provides only sparse textual supervision for a highly structured generation problem. The model must simultaneously infer shot boundaries and durations, preserve cross-shot visual continuity, and maintain audio-video consistency from text alone, making direct text-conditioned adaptation unstable and inefficient. This difficulty is consistent with the gap observed in Sec. 1, where existing foundation models respond weakly to shot-structured prompts and often fail to form clear multi-shot organization. We further diagnose this limitation through temporal attention visualization. As shown in Fig. 7, the direct T2AV baseline exhibits a monotonically decaying ***temporal locality pattern*** that is agnostic to shot boundaries [72]. Its attention mass falls off smoothly with frame distance and shows no discontinuity at the prompt-specified shot boundaries. This suggests that the shot organization specified in the prompt is not effectively reflected in the backbone's temporal attention. Motivated by the adaptive shot organization enabled by *reference-to-multi-shot generation*, CineDance uses reference frames as ***training-time visual-temporal scaffolds*** rather than mandatory inference-time inputs. We first make multi-shot organization easier to learn with dense visual and temporal anchors, and then gradually weaken these anchors so that the model can retain the learned structure under reduced inference conditions.

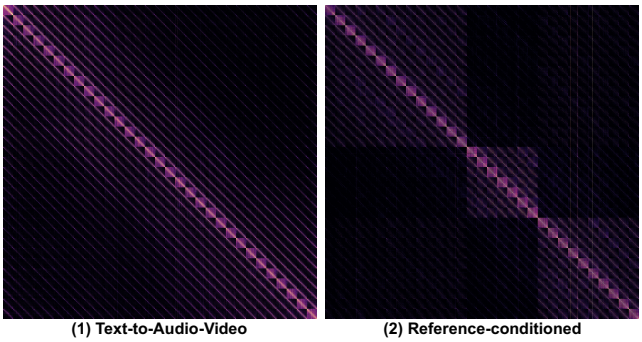


Fig. 7: Self-attention map visualization under the same generation condition. Both maps are extracted from the same transformer block at the same intermediate denoising step, averaged over attention heads. **(1)** Direct T2AV generation mainly exhibits shot-agnostic temporal locality, **(2)** whereas reference conditioning produces clearer segment-level attention structure.

Training-time visual-temporal scaffolds. Since the internal mechanisms of recent reference-to-multi-shot systems are not publicly specified, we implement a lightweight yet effective reference-conditioning interface following common practice in open reference-conditioned video generation. Specifically, each reference frame is encoded by the video VAE into a latent representation

and appended to the video latent tokens as additional reference tokens, serving as the ***visual scaffold*** that provides dense cues for scene layout, style, composition, and recurring visual content.

A key design choice is to assign proper RoPE [60] indices to these reference tokens. Each reference latent uses the same spatial RoPE indices as a normal video latent frame, so that the reference content, together with its spatial RoPE assignment, can be aligned with the spatial layout of generated video tokens. LTX-2 uses the timestamp of each conditioning frame, computed from its frame index and the target frame rate, as its temporal RoPE index. To make the initial training well-conditioned, we assign each reference frame the ground-truth temporal index of its corresponding frame in the training video, forming a simple but effective ***temporal scaffold***.

We formalize the reference scaffold as $s_{\text{ref}}(\eta_v, \eta_t) = \{(\rho(r_k;\eta_v), \psi(\tau_k;\eta_t))\}_{k=1}^{K}$, where $r_k = \mathcal{E}_{\text{VAE}}(x_k^{\text{ref}})$ is the VAE latent of the k-th reference frame, and τ_k is its ground-truth frame-level timestamp. We further introduce two strength parameters, η_v and η_t, to control the guidance strength of the visual and temporal scaffolds, respectively.

Progressive scaffold removal. The training-time scaffold should make optimization easier, but it should not remain a mandatory inference-time input. We therefore progressively weaken and remove the reference scaffold through the ***D***ual-***A***xis ***R***eference ***C***urriculum (***DARC***), a two-step reference-strength annealing process.

For visual scaffold, We first weaken the visual scaffold through continuous noising of the reference latent. Given the clean reference latent r_k, we define

$$\rho(r_k;\eta_v(u)) = \eta_v(u) r_k + \big(1-\eta_v(u)\big)\epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, I),$$

where u denotes the training step and $\eta_v(u) \in [0,1]$ is a ***monotonically decreasing*** visual-reference strength schedule. As $\eta_v(u)$ decreases, the reference latent is gradually interpolated from a clean visual anchor to pure Gaussian noise.

Even when the reference latent is fully noised, the model may still rely on the temporal RoPE index of the reference token to organize shot placement and duration. ***For the temporal scaffold***, we therefore relax the reference token's temporal index through a stochastic switch:

$$\psi(\tau_k;\eta_t) = \begin{cases} \tau_k, & \text{with probability } 1-q(\eta_t), \\ \hat{\tau}_k, & \text{with probability } q(\eta_t), \end{cases}$$

where τ_k is the ground-truth timestamp of the corresponding reference frame, and $\hat{\tau}_k = \lfloor kN/(K+1) \rfloor$ is a placement-independent ordinal index. The switching probability $q(\eta_t)$ increases as the temporal strength decreases, so the reference tokens retain only their relative temporal order while losing their exact timestamps.

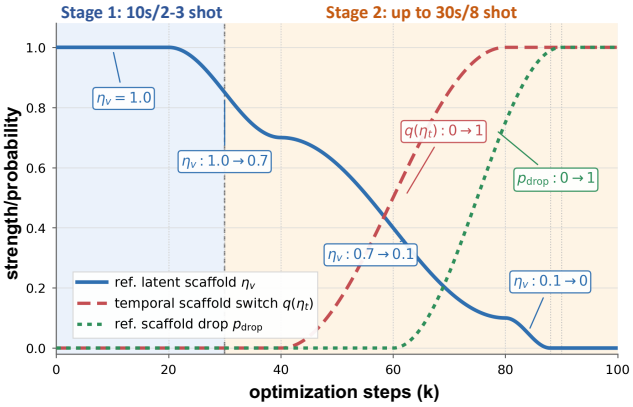


Fig. 8: Overall training schedule. ***1)*** The background color shows the data-driven curriculum. ***2)*** The three curves summarize DARC: the visual scaffold strength η_v decreases, the temporal-switch probability $q(\eta_t)$ increases, and the reference-dropping probability p_{drop} is activated in the late stage.

Although the reference latent and its RoPE assignment no longer provide reliable visual details or exact timestamp information, the reference token itself remains in the sequence and may still act as an extra implicit conditioning slot, hindering the training condition from matching the true T2AV regime. We therefore complete the scaffold removal with a ***final reference-dropping stage***, where the reference latents and their RoPE indices are removed together:

$$\bar{s}_{\text{ref}} = \begin{cases} s_{\text{ref}}, & \text{with probability } 1-p_{\text{drop}}(u), \\ \varnothing, & \text{with probability } p_{\text{drop}}(u), \end{cases}$$

where $\varnothing$ removes all reference tokens and their positional assignments from the input sequence. The dropping probability $p_{\text{drop}}(u)$ increases in the late stage of training, gradually turning weakened reference conditioning into genuine text-only T2AV training.

5.4 Long-Form Adaptation and Implementation Details

Data-driven curriculum. The previously introduced DARC mainly addresses the latent-domain gap when adapting the backbone from single-shot generation to multi-shot generation. Since CineDance further targets long-form audio-video generation, we additionally introduce a data-driven curriculum to expand the model's effective temporal capacity: ***1)*** Stage 1 trains on short

Fig. 9: Qualitative comparison of CineDance with baseline models. **Red and yellow boxes** indicate character identity inconsistencies across shots, while **blue boxes** denote failed shot transitions.

windows of 2–3 adjacent shots with durations of 10–12 seconds, so that the model first learns local shot switching. ***2)*** Stage 2 extends the training window to up to 8 shots and approximately 30 seconds, encouraging longer-range cross-shot consistency. Both stages sample temporally contiguous windowed video clips from the original long-form sequences within CineDance-1M.

Structured prompt organization. For each windowed video clip containing M selected shots, we render a structured textual condition consisting of character and scene headers followed by re-indexed shot-level blocks:

$$c_{\text{text}} = \big[\underbrace{\mathcal{H}_{\text{char}}, \mathcal{H}_{\text{scene}}}_{\text{global headers}}, \underbrace{\mathcal{B}_1, \ldots, \mathcal{B}_M}_{\text{shot blocks}}\big].$$

The headers are defined as

$$\mathcal{H}_{\text{char}} = \{\, \langle \text{char}c \rangle = e_c \,\}_{c=1}^{C},$$
$$\mathcal{H}_{\text{scene}} = \{\, \langle \text{scene}s \rangle = g_s \,\}_{s=1}^{S},$$

where e_c and g_s denote the character descriptions and scene description, respectively. *Only characters and scenes appearing in the selected shots are included.* Each shot block is rendered as

$$\mathcal{B}_i = \big[\text{SHOT } i \mid \text{scene } s_i \mid \text{camera } \kappa_i\big] \oplus d_i^v \oplus d_i^a \oplus \{(\text{spk}_{i,\ell}, \text{speech}_{i,\ell})\}_{\ell=1}^{L_i},$$

where s_i is the scene anchor, κ_i is the camera descriptor, d_i^v and d_i^a are the visual and audio descriptions, and $(\text{spk}_{i,\ell}, \text{speech}_{i,\ell})$ denotes the ℓ-th ASR segment in shot i, with a speaker anchor and the corresponding transcript. Regardless of the original shot indices in the source sequence, the selected shots are re-indexed from 1 to M to provide a compact local timeline.

Training details. Fig. 8 summarizes the overall training schedule, including all curriculum and scaffold-removal stages introduced above. To ensure smooth transitions across training phases, all curriculum control variables are updated with cosine schedules. We initialize the model from LTX-2.3 and fully fine-tune it using AdamW [42] for 100K optimization steps, with a global batch size of 32, a constant learning rate of 5×10^{-5}, and a weight decay of 0.01. All videos are trained at 480p resolution, resized to 480×832, with a frame rate of 24 FPS; audio is resampled to 16 kHz. We train in bfloat16 precision and use a balanced audio-video loss weighting $\lambda_v : \lambda_a = 1 : 1$.

6 Experiments

6.1 Experimental Setup

We evaluate all methods on CineBench using the automatic metric suite defined in Sec. 4.3. Although CineBench includes three *Duration/Shot Count* tiers, existing methods still have limited support for stable 60s audio-video generation. Therefore, we conduct the main benchmark on ***1)*** 10s with 2–3 shots, ***2)*** 30s with 4–9 shots two tiers, which cover both short local shot

Table 8: Quantitative comparison of CineDance with SOTA baselines on CineBench. Best and second-best results are **bolded** and underlined. The "–" symbol denotes no audio generation capability. O indicates original implementations, while L utilizes LTX-2.3 as the video generator.

Method	Video Quality			Audio Quality				AV Sync		Prompt Alignment			Continuity	
	AQ ↑	IQ ↑	MS ↑	PQ ↑	CE ↑	CU ↑	WER ↓	Sync-C↑/D↓	IB-S ↑	ViC ↑	VSc ↑	IB-A ↑	ID ↑	Scene ↑
CineTrans [77]	0.47	0.48	0.97	–	–	–	–	–	–	0.14	2.92	–	0.21	0.63
Mask2DiT [55]	0.56	0.67	**0.99**	–	–	–	–	–	–	0.14	2.76	–	0.23	0.55
MultiShotMaster [67]	0.51	0.51	0.96	–	–	–	–	–	–	0.15	3.19	–	0.13	0.37
HoloCine [47]	0.50	0.36	**0.99**	–	–	–	–	–	–	0.17	3.21	–	0.12	0.58
StoryMem [92]	0.57	0.66	0.98	–	–	–	–	–	–	0.10	2.68	–	0.26	0.53
VGoT [99] O	0.42	0.53	0.91	–	–	–	–	–	–	0.19	0.72	–	0.18	0.52
VGoT [99] L	0.45	0.55	0.93	6.45	4.05	5.75	0.73	1.42 / 9.45	0.18	0.14	0.75	0.11	0.16	0.49
MovieAgent [76] O	0.54	0.62	0.95	–	–	–	–	–	–	0.20	0.71	–	0.15	0.54
MovieAgent [76] L	0.58	0.65	0.97	6.58	4.12	5.82	0.69	1.35 / 9.12	0.21	0.15	0.78	0.13	0.21	0.57
STAGE [93] O	0.49	0.50	0.98	–	–	–	–	–	–	**0.23**	2.82	–	0.03	0.35
STAGE [93] L	0.42	0.44	0.92	6.32	3.65	5.50	0.81	1.28 / 9.88	0.15	0.16	0.65	0.10	0.12	0.39
LTX-2.3 (Base) [16]	0.52	0.56	0.98	6.66	**4.20**	**5.93**	0.66	1.11 / **8.42**	0.26	0.05	3.04	0.17	0.23	0.63
CineDance (Ours)	**0.59**	**0.68**	**0.99**	**6.72**	**4.20**	5.91	**0.52**	**1.53** / 8.46	**0.27**	0.21	**3.22**	**0.19**	**0.30**	**0.69**

transitions and longer multi-shot consistency while remaining feasible for all compared methods. The 60s with 10–20 shots tier is retained in CineBench as a forward-looking evaluation split and is reported separately when a method supports such generation.

6.2 Main Comparison

Baselines. Following the taxonomy in Sec. 5.2, we compare CineDance with representative methods from two existing multi-shot generation paradigms. ***1)*** For per-shot generation with memory propagation, we include STAGE [93], VGoT [99], MovieAgent [76], and StoryMem [92]. To ensure fair comparison under a unified generator capacity, we additionally evaluate STAGE, VGoT, and MovieAgent after standardizing their underlying video generator to LTX-2.3. StoryMem involves its own fine-tuning stage, and therefore we do not construct an additional LTX-2.3-based variant for it. ***2)*** For structure-aware single-forward generation, we include CineTrans [77], Mask2DiT [55], MultiShotMaster [67], and HoloCine [47]. We also include the base LTX-2.3 model [16] as the backbone baseline to measure the effect of CineDance adaptation. For all methods, we convert the structured annotations in CineBench into the *most faithful input conditions* supported by each method, including shot-level prompts or shot structures. All methods are evaluated at 480p resolution, which falls within the supported resolution range of all baselines; for frame rate, we follow the recommended settings in each official implementation.

Qualitative Comparison. As shown in Fig. 9, different baseline paradigms exhibit distinct failure modes under complex multi-shot narratives. Specifically, CineTrans and STAGE suffer from severe identity inconsistency and video quality degradation. HoloCine maintains moderate background consistency, but occasionally shows weak shot response and fails to produce clear transitions aligned with the target structure. Meanwhile, the base LTX-2.3 fails to follow multi-shot textual conditions and exhibits spatial degradation over extended durations. In contrast, CineDance intrinsically acquires shot-transition capability without explicit condition injection, preserving the base model's strong visual fidelity while aligning more faithfully with structural prompts.

Table 9: Shot Structure Response on methods without oracle shot-layout control.

Method	S_{cnt} ↑	S_{seg} ↑	SSR ↑
STAGE StoryMem VGoT MovieAgent LTX-2.3 (Base)	1.0000	0.5220	0.6553
	0.6682	0.3821	0.4646
CineDance (ours)	0.9657	0.5866	0.6985

Quantitative Comparison. ***1)*** For basic video and audio quality, CineDance achieves the best or competitive scores across most metrics. These low-level quality metrics are closely related to the capability of the underlying backbone, and CineDance remains competitive with, or even surpasses, pipelines that generate shots separately. Compared with the base LTX-2.3, CineDance obtains clear improvements, indicating that long-form adapta-

Table 10: Ablation studies on CineBench. The upper block analyzes the effect of training data, while the lower block compares training strategies under the same CineDance-1M setting. CD-1M denotes CineDance-1M. Best results are **bolded**.

Ablated Variants	AV Sync		Prompt Alignment			Continuity		SSR ↑
	Sync-C ↑/D ↓	IB-S ↑	ViC ↑	VSc ↑	IB-A ↑	ID ↑	Scene ↑	
Effect of training data								
OpenhumanVid (1M clips)	1.08 / 8.56	0.24	0.06	3.02	0.15	0.22	0.60	0.4412
OpenhumanVid (full)	1.15 / 8.38	0.26	0.08	3.06	0.17	0.24	0.63	0.4785
CD-1M shot-only clips	1.28 / 8.92	0.23	0.14	3.12	0.16	0.19	0.55	0.5246
CD-1M w/o structured shot-level ann.	1.41 / 8.62	0.25	0.17	3.15	0.18	0.27	0.66	0.6215
CD-1M full (Ours)	**1.53** / **8.46**	**0.27**	**0.21**	**3.22**	**0.19**	**0.30**	**0.69**	**0.6985**
Effect of training strategy								
Direct T2AV	1.21 / 9.05	0.20	0.16	2.98	0.13	0.25	0.64	0.5630
Full DARC (Ours)	**1.53** / **8.46**	**0.27**	**0.21**	**3.22**	**0.19**	**0.30**	**0.69**	**0.6985**

tion on CineDance-1M does not degrade the pretrained visual-audio prior and further demonstrates the effectiveness of our training strategy and the high quality of the dataset. ***2)*** For audio-video synchronization and prompt alignment, CineDance also achieves the best or competitive performance. The relatively low Sync-C scores across methods are mainly caused by challenging benchmark cases with off-screen speech, large facial motion, or non-frontal faces, where lip-centric synchronization models become less reliable. Nevertheless, CineDance improves both shot-level and video-level prompt following, showing that structured CineDance training helps the model better understand multi-shot narrative conditions. ***3)*** For narrative continuity, CineDance obtains stronger identity and scene consistency than existing baselines, demonstrating its ability to preserve recurring characters and environments across shot transitions, which is consistent with the qualitative observations in Fig. 9. Against the base LTX-2.3, the improvements further confirm that the gains come from CineDance adaptation rather than the pretrained backbone alone.

Shot-structure response analysis. For fair comparison, we report Shot Structure Response (SSR) separately in Tab. 9, since Paradigm-2 baselines receive explicit ground-truth shot-layout conditions. Paradigm-1 methods obtain identical SSR scores because their normalized shot partitions are the same after stitching. Their S_{cnt} reaches 1.0 because these methods are given the target number of shots and generate each shot separately, indicating that they are not fully free from shot-structure information either. However, their lower S_{seg} shows that matching the shot count alone does not guarantee a faithful temporal layout, as uniform or suboptimal average shot durations may still deviate from the ground-truth shot distribution. The base LTX-2.3 obtains substantially weaker SSR, confirming that the pretrained backbone has limited response to shot-level structure, consistent with our observations in Sec. 1 and Sec. 5.3. Although CineDance does not receive explicit shot-count or shot-layout control, its S_{cnt} remains close to 1.0, and its S_{seg} and final SSR surpass Paradigm-1 methods, demonstrating stronger intrinsic shot-structure response and validating SSR as a meaningful and robust metric for shot-structure benchmarking.

6.3 Ablation Study

Ablation on datasets. We first ablate the role of training data while keeping the backbone, training strategy, and optimization budget unchanged. Specifically, we compare five variants: ***1)*** training on 1M randomly sampled clips from OpenHumanVid [34], ***2)*** training on the full OpenHumanVid, ***3)*** training on isolated shot-only clips from CineDance-1M, ***4)*** training on CineDance-1M without structured shot-level annotations, and ***5)*** training on the full CineDance-1M. OpenHumanVid [34] serves as an external human-centric audio-video baseline. The shot-only variant uses the shot-level annotation of each isolated shot as its prompt, without modeling multi-shot context. As shown in Tab. 10, short-clip training mainly preserves the basic capability of the LTX-2.3 backbone, but brings limited improvement on narrative-oriented metrics such as multi-shot consistency and SSR. Training on CineDance-1M without structured shot-level annotations yields moderate gains on these narrative metrics, suggesting that multi-shot videos themselves provide useful supervision for multi-shot long-form generation. However, this variant remains clearly below the full setting, showing that structured shot-level annotations are necessary for learning fine-grained prompt following and controllable shot organization. These results demonstrate the value of CineDance-1M from two

Table 11: Human Evaluation Results. We report the Mean Opinion Score (1-5 scale, higher is better) across the six proposed dimensions. Best results are **bolded**.

Method	VQ	AQ	Sync	Align.	Cont.	SSR
CineTrans [77]	2.92	–	–	3.08	2.71	3.15
Mask2DiT [55]	3.91	–	–	2.72	2.85	3.18
MultiShotMaster [67]	3.32	–	–	3.12	2.31	3.62
HoloCine [47]	3.18	–	–	3.75	2.58	3.61
StoryMem [92]	3.72	–	–	2.85	3.31	3.42
VGoT [99] O	3.42	–	–	3.55	3.35	3.41
VGoT [99] L	3.48	3.72	3.65	3.41	3.22	3.31
MovieAgent [76] O	3.55	–	–	3.51	3.21	3.58
MovieAgent [76] L	3.61	3.85	3.71	3.54	3.25	3.45
STAGE [93] O	3.28	–	–	3.68	2.52	3.58
STAGE [93] L	3.32	3.61	3.41	3.38	2.81	3.42
LTX-2.3 (Base) [16]	3.75	**4.05**	3.12	3.32	2.88	2.61
CineDance (Ours)	**4.12**	3.98	**4.11**	**4.18**	**4.25**	**4.38**

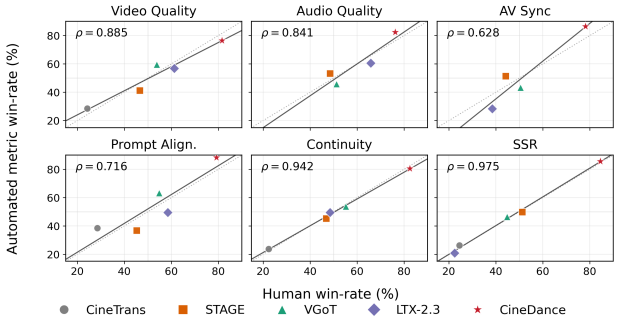


Fig. 10: Human alignment of the CineBench automatic metrics. Each subplot compares model-level human win rates with the corresponding automated metric win rates across six evaluation dimensions. The reported Spearman rank correlation ρ indicates strong human alignment for most dimensions.

complementary aspects: ***1)*** long-form multi-shot video data and ***2)*** structured shot-wise annotations.

Ablation on methods. We further ablate the training strategy under the same CineDance-1M data setting. The Direct T2AV variant uses the same structured textual prompts as Full DARC, but directly optimizes the text-only audio-video generation objective without any reference-scaffolded training. Full DARC instead starts from reference-scaffolded multi-shot learning and progressively removes the visual and temporal scaffolds until the model is trained under the same text-only condition used at inference. As shown in Tab. 10, Direct T2AV preserves the basic generation ability of the backbone, but yields limited gains in shot-level prompt following, cross-shot continuity, and shot-structure response. Full DARC consistently improves all evaluated dimensions, indicating that progressive scaffold removal provides a more effective easy-to-hard training path for learning multi-shot organization from CineDance-1M.

6.4 Human Study

Human preference study. As shown in Tab. 11, CineDance achieves the best Mean Opinion Score (MOS) on five out of six dimensions, including video quality, audio-video synchronization, prompt alignment, narrative continuity, and shot-structure response, while remaining competitive in audio quality. Compared with the base LTX-2.3 model, CineDance shows substantial improvements on structure-related dimensions, demonstrating that CineDance adaptation effectively enhances multi-shot controllability, cross-shot coherence, and narrative organization. Compared with existing multi-shot baselines, the gains are especially pronounced in continuity and SSR, indicating that human evaluators perceive more stable recurring characters, more coherent scenes, and clearer shot transitions. These results show that CineDance improves not only automatic scores but also human-perceived multi-shot generation quality.

Human Alignment Analysis. Fig. 10 further examines whether CineBench automatic metrics are aligned with human judgments. For each evaluation dimension, we compute the model-level win rate induced by automatic scores and compare it with the corresponding human win rate. The two rankings show positive correlations across all six dimensions, indicating that the proposed automatic suite generally reflects human preferences. In particular, Continuity and SSR achieve the strongest correlations, with Spearman coefficients of $\rho = 0.942$ and $\rho = 0.975$, respectively, suggesting that our annotation-grounded continuity metrics and shot-structure response metric are well aligned with human perception of multi-shot coherence. Video Quality and Audio Quality also show high correlations, while AV Sync exhibits a relatively weaker but still positive correlation. Overall, the alignment analysis supports CineBench as a meaningful automatic evaluation suite for multi-shot, narrative-driven audio-video generation.

7 Conclusion

In this paper, we introduced CineDance-1M, a pioneering million-scale 1080p dataset designed to overcome the critical bottleneck in multi-shot, long-form joint audio-video generation. Driven by film-theory-inspired narrative parsing and hierarchical dual-modal annotation, our curation pipeline guarantees exceptional semantic alignment. Alongside the dataset, we established CineBench for human-aligned evaluation and developed a robust

baseline, CineDance, which demonstrates precise cross-modal synchronization and spatio-temporal consistency, successfully bridging the divide between isolated short clips and cohesive narrative synthesis.

Limitations, broader impacts and ethical considerations. While CineBench is currently bounded at 1080p resolution, we intend to advance ultra-high-definition applications in future work by releasing an additional 4K multi-shot audio-visual dataset. Furthermore, we acknowledge that highly realistic synthesis inherently exacerbates deepfake risks, underscoring the urgent need for robust deepfake detection frameworks to safeguard information authenticity.

Declarations

Data Availability. CineBench prompts, evaluation metadata, and metric implementations will be made publicly available on the project page upon acceptance or publication. For CineDance-1M, we will publicly release the structured annotations and metadata associated with the dataset, together with filtering scripts and reconstruction instructions. For video sources derived from public platforms such as YouTube, we will provide metadata-based access rather than redistributing raw video files. For self-collected sources for which redistribution is permitted, access to the corresponding data will be provided through a gated application process for research use. Raw source videos subject to third-party copyright, platform terms, or redistribution restrictions will not be directly redistributed. We will maintain a takedown-supported access protocol and promptly remove or restrict access to any data upon valid copyright, privacy, or source-platform requests.

Code Availability. The code for data processing, benchmark construction, evaluation, and model adaptation will be released on the project page upon acceptance or publication.

References

1. Afouras, T., Chung, J.S., Zisserman, A.: Lrs3-ted: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496 (2018) 4
2. An, Z., Jia, M., Qiu, H., Zhou, Z., Huang, X., Liu, Z., Ren, W., Kahatapitiya, K., Liu, D., He, S., et al.: Onestory: Coherent multi-shot video generation with adaptive memory. arXiv preprint arXiv:2512.07802 (2025) 5, 14
3. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 1728–1738 (2021) 4
4. Bordwell, D., Thompson, K., Smith, J.: Film art: An introduction, vol. 7. McGraw-Hill New York (2008) 8, 9
5. Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., et al.: Skyreels-v2: Infinite-length film generative model. arXiv preprint arXiv:2504.13074 (2025) 4, 6
6. Chen, G., Lin, D., Yang, J., Zhang, Y., Fei, Z., Li, D., Chen, S., Ao, C., Pang, N., Wang, Y., et al.: Skyreels-v4: Multi-modal video-audio generation, inpainting and editing model. arXiv preprint arXiv:2602.21818 (2026) 14
7. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 721–725. IEEE (2020) 4
8. Chen, T.S., Siarohin, A., Menapace, W., Deyneka, E., Chao, H.w., Jeon, B.E., Fang, Y., Lee, H.Y., Ren, J., Yang, M.H., et al.: Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13320–13331 (2024) 4
9. Cheng, H.K., Ishii, M., Hayakawa, A., Shibuya, T., Schwing, A., Mitsufuji, Y.: Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 28901–28911 (2025) 5
10. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Asian conference on computer vision, pp. 251–263. Springer (2016) 7, 13
11. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4690–4699 (2019) 13
12. Eisenstein, S.M., Bois, Y.A., Glenny, M.: Montage and architecture. Assemblage (10), 111–131 (1989) 8
13. Fan, F., Luo, C., Gao, W., Zhan, J.: Aigcbench: Comprehensive evaluation of image-to-video content generated by ai. BenchCouncil Transactions on Benchmarks, Standards and Evaluations **3**(4), 100152 (2023) 5
14. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 15180–15190 (2023) 7, 13
15. Guo, X., Huo, J., Shi, Z., Song, Z., Zhang, J., Zhao, J.: T2vtextbench: A human evaluation benchmark for textual control in video generation models. arXiv preprint arXiv:2505.04946 (2025) 5
16. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., et al.: Ltx-2: Efficient joint audio-visual foundation model. arXiv preprint arXiv:2601.03233 (2026) 4, 5, 9, 13, 17, 19
17. He, R., Wei, M., Yang, Z., Ordonez, V.: Entitybench: Towards entity-consistent long-range multi-shot video generation. arXiv preprint arXiv:2605.15199 (2026) 5
18. He, X., Jiang, D., Nie, P., Liu, M., Jiang, Z., Su, M., Ma, W., Lin, J., Ye, C., Lu, Y., et al.: Videoscore2: Think before you score in generative video evaluation. arXiv preprint arXiv:2509.22799 (2025) 5
19. He, X., Jiang, D., Zhang, G., Ku, M., Soni, A., Siu, S., Chen, H., Chandra, A., Jiang, Z., Arulraj, A., et al.: Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 2105–2123 (2024) 5, 13

20. Hu, T., Yu, Z., Zhang, G., Su, Z., Zhou, Z., Zhang, Y., Zhou, Y., Lu, Q., Yi, R.: Harmony: Harmonizing audio and video generation through cross-task synergy. arXiv preprint arXiv:2511.21579 (2025) 5
21. Hu, T., Yu, Z., Zhou, Z., Liang, S., Zhou, Y., Lin, Q., Lu, Q.: Hunyuancustom: A multimodal-driven architecture for customized video generation. arXiv preprint arXiv:2505.04512 (2025) 1
22. Hu, T., Zhang, J., Huang, H., Yi, R., Su, Z., Weng, J., Xue, Z., Ma, L., Yang, M.H., Tao, D.: Evolution of video generative foundations. arXiv preprint arXiv:2604.06339 (2026) 1
23. Hu, T., Zhang, J., Su, Z., Yi, R.: Ultragen: High-resolution video generation with hierarchical attention. arXiv preprint arXiv:2510.18775 (2025) 1
24. Hua, D., Wang, X., Zeng, B., Huang, X., Liang, H., Niu, J., Chen, X., Xu, Q., Zhang, W.: Vabench: A comprehensive benchmark for audio-video generation. arXiv preprint arXiv:2512.09299 (2025) 5
25. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21807–21818 (2024) 4, 5, 12
26. Huang, Z., Zhang, F., Xu, X., He, Y., Yu, J., Dong, Z., Ma, Q., Chanpaisit, N., Si, C., Jiang, Y., et al.: Vbench++: Comprehensive and versatile benchmark suite for video generative models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 4, 5
27. Jaided, A.: Easyocr: Ready-to-use ocr with 80+ supported languages. GitHub repository (2020) 7
28. Ji, S., Chen, X., Yang, S., Tao, X., Wan, P., Zhao, H.: Memflow: Flowing adaptive memory for consistent and efficient long video narratives. arXiv preprint arXiv:2512.14699 (2025) 5
29. Jia, W., Lu, Y., Huang, M., Wang, H., Huang, B., Chen, N., Liu, M., Jiang, J., Mao, Z.: Moga: Mixture-of-groups attention for end-to-end long video generation. arXiv preprint arXiv:2510.18692 (2025) 5
30. Ju, X., Gao, Y., Zhang, Z., Yuan, Z., Wang, X., Zeng, A., Xiong, Y., Xu, Q., Shan, Y.: Miradata: A large-scale video dataset with long durations and structured captions. Advances in Neural Information Processing Systems **37**, 48955–48970 (2024) 2, 4, 6, 7, 10, 11
31. Kara, O., Singh, K.K., Liu, F., Ceylan, D., Rehg, J.M., Hinz, T.: Shotadapter: Text-to-multi-shot video generation with diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 28405–28415 (2025) 5
32. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 5148–5157 (2021) 12
33. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024) 1, 4
34. Li, H., Xu, M., Zhan, Y., Mu, S., Li, J., Cheng, K., Chen, Y., Chen, T., Ye, M., Wang, J., et al.: Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 7752–7762 (2025) 4, 6, 10, 18
35. Li, Z., Zhu, Z.L., Han, L.H., Hou, Q., Guo, C.L., Cheng, M.M.: Amt: All-pairs multi-field transforms for efficient frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9801–9810 (2023) 7, 12
36. Liao, M., Lu, H., Zhang, X., Wan, F., Wang, T., Zhao, Y., Zuo, W., Ye, Q., Wang, J.: Evaluation of text-to-video generation models: A dynamics perspective. Advances in Neural Information Processing Systems **37**, 109790–109816 (2024) 5
37. Lin, B., Ge, Y., Cheng, X., Li, Z., Zhu, B., Wang, S., He, X., Ye, Y., Yuan, S., Chen, L., et al.: Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131 (2024) 4
38. Ling, X., Zhu, C., Wu, M., Li, H., Feng, X., Yang, C., Hao, A., Zhu, J., Wu, J., Chu, X.: Vmbench: A benchmark for perception-aligned video motion generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13087–13098 (2025) 5
39. Liu, K., Li, W., Chen, L., Wu, S., Zheng, Y., Ji, J., Zhou, F., Luo, J., Liu, Z., Fei, H., et al.: Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. arXiv preprint arXiv:2503.23377 (2025) 5
40. Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., Shan, Y.: Evalcrafter: Benchmarking and evaluating large video generation models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 22139–22149 (2024) 5
41. Liu, Y., Li, L., Ren, S., Gao, R., Li, S., Chen, S., Sun, X., Hou, L.: Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. Advances in Neural Information Processing Systems **36**, 62352–62387 (2023) 4, 5
42. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 16
43. Low, C., Wang, W., Katyal, C.: Ovi: Twin backbone cross-modal fusion for audio-video generation. arXiv preprint arXiv:2510.01284 (2025) 5, 9
44. Luo, X., Li, Q., Liu, X., Qin, W., Yang, M., Wang, M., Wan, P., Zhang, D., Gai, K., Huang, S.L.: Filmweaver: Weaving consistent multi-shot videos with cache-guided autoregressive diffusion. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, pp. 7689–7697 (2026) 5
45. Luo, Y., Shi, X., Zhuang, J., Chen, Y., Liu, Q., Wang, X., Wan, P., Xue, T.: Shotstream: Streaming multi-shot video generation for interactive storytelling. arXiv preprint arXiv:2603.25746 (2026) 5
46. Mao, Y., Shen, X., Zhang, J., Qin, Z., Zhou, J., Xiang, M., Zhong, Y., Dai, Y.: Tavgbench: Benchmarking text to audible-video generation. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 6607–6616 (2024) 5
47. Meng, Y., Ouyang, H., Yu, Y., Wang, Q., Wang, W., Cheng, K.L., Wang, H., Li, Y., Chen, C., Zeng, Y., et al.: Holocine: Holistic generation of cinematic multi-shot long video narratives. arXiv preprint arXiv:2510.20822 (2025) 5, 14, 17, 19
48. Metz, C.: La grande syntagmatique du film narratif. Communications **8**(1), 120–124 (1966) 8
49. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 2630–2640 (2019) 4, 10
50. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612 (2017) 4

51. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. arXiv preprint arXiv:2407.02371 (2024) 4, 10
52. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 13
53. Pan, K., Tian, Q., Zhang, J., Kong, W., Xiong, J., Long, Y., Zhang, S., Qiu, H., Wang, T., Lv, Z., et al.: Omniweaving: Towards unified video generation with free-form composition and reasoning. arXiv preprint arXiv:2603.24458 (2026) 5
54. Phung, Q., Mai, L., Heilbron, F.D.C., Liu, F., Huang, J.B., Ham, C.: Cineverse: Consistent keyframe synthesis for cinematic scene composition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2626–2636 (2026) 5
55. Qi, T., Yuan, J., Feng, W., Fang, S., Liu, J., Zhou, S., He, Q., Xie, H., Zhang, Y.: Mask^ 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) 5, 14, 17, 19
56. Reddy, C.K., Gopal, V., Cutler, R.: Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6493–6497. IEEE (2021) 7
57. Seedance, T., Chen, D., Chen, L., Chen, X., Chen, Y., Chen, Z., Chen, Z., Cheng, F., Cheng, T., Cheng, Y., et al.: Seedance 2.0: Advancing video generation for world complexity. arXiv preprint arXiv:2604.14148 (2026) 1, 14
58. Shi, H., Li, Y., Deng, N., Xu, Z., Chen, X., Wang, L., Hu, B., Zhang, M.: Msvbench: Towards human-level evaluation of multi-shot video generation. arXiv preprint arXiv:2602.23969 (2026) 3, 4, 5
59. Soucek, T., Lokoc, J.: 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) 3, 8
60. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024) 5, 15
61. Sun, K., Huang, K., Liu, X., Wu, Y., Xu, Z., Li, Z., Liu, X.: T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 8406–8416 (2025) 4, 5
62. Tan, Z., Yang, X., Qin, L., Li, H.: Vidgen-1m: A large-scale dataset for text-to-video generation. arXiv preprint arXiv:2408.02629 (2024) 10
63. Team, O., Yu, D., Chen, M., Chen, Q., Luo, Q., Wu, Q., Cheng, Q., Li, R., Liang, T., Zhang, W., et al.: Mova: Towards scalable and synchronized video-audio generation. arXiv preprint arXiv:2602.08794 (2026) 9
64. Tjandra, A., Wu, Y.C., Guo, B., Hoffman, J., Ellis, B., Vyas, A., Shi, B., Chen, S., Le, M., Zacharov, N., et al.: Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. arXiv preprint arXiv:2502.05139 (2025) 12
65. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 1, 4
66. Wang, D., Zuo, W., Li, A., Chen, L.H., Liao, X., Zhou, D., Yin, Z., Dai, X., Jiang, D., Yu, G.: Universe-1: Unified audio-video generation via stitching of experts. arXiv preprint arXiv:2509.06155 (2025) 5
67. Wang, Q., Shi, X., Li, B., Bian, W., Liu, Q., Lu, H., Wang, X., Wan, P., Gai, K., Jia, X.: Multishotmaster: A controllable multi-shot video generation framework. arXiv preprint arXiv:2512.03041 (2025) 5, 14, 17, 19
68. Wang, Q., Shi, Y., Ou, J., Chen, R., Lin, K., Wang, J., Jiang, B., Yang, H., Zheng, M., Tao, X., et al.: Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 8428–8437 (2025) 4, 6, 10
69. Wang, W., Yang, Y.: Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. Advances in Neural Information Processing Systems **37**, 65618–65642 (2024) 5
70. Wang, W., Yang, Y.: Videoufo: A million-scale user-focused dataset for text-to-video generation. arXiv preprint arXiv:2503.01739 (2025) 4
71. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv preprint arXiv:2307.06942 (2023) 4, 13
72. Wen, Y., Wu, J., Jain, A., Goldstein, T., Panda, A.: Analysis of attention in video diffusion transformers. arXiv preprint arXiv:2504.10317 (2025) 14
73. Wu, B., Zou, C., Li, C., Huang, D., Yang, F., Tan, H., Peng, J., Wu, J., Xiong, J., Jiang, J., et al.: Hunyuanvideo 1.5 technical report. arXiv preprint arXiv:2511.18870 (2025) 4
74. Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 20144–20154 (2023) 7
75. Wu, W., Liu, M., Zhu, Z., Xia, X., Feng, H., Wang, W., Lin, K.Q., Shen, C., Shou, M.Z.: Moviebench: A hierarchical movie level dataset for long video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 28984–28994 (2025) 7
76. Wu, W., Zhu, Z., Shou, M.Z.: Automated movie generation via multi-agent cot planning. arXiv preprint arXiv:2503.07314 (2025) 5, 14, 17, 19
77. Wu, X., Gao, B., Qiao, Y., Wang, Y., Chen, X.: Cinetrans: Learning to generate videos with cinematic transitions via masked diffusion models. arXiv preprint arXiv:2508.11484 (2025) 2, 5, 7, 10, 14, 17, 19
78. Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., Dubnov, S.: Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023) 7
79. Xiao, J., Yang, C., Zhang, L., Cai, S., Zhao, Y., Guo, Y., Wetzstein, G., Agrawala, M., Yuille, A., Jiang, L.: Captain cinema: Towards short movie generation. In: The Fourteenth International Conference on Learning Representations (2025) 5
80. Xie, Z., Tang, D., Tan, D., Klein, J., Bissyand, T.F., Ezzini, S.: Dreamfactory: Pioneering multi-scene long video generation with a multi-agent framework. arXiv preprint arXiv:2408.11788 (2024) 5

81. Xiong, T., Wang, Y., Zhou, D., Lin, Z., Feng, J., Liu, X.: Lvd-2m: A long-take video dataset with temporally dense captions. Advances in Neural Information Processing Systems **37**, 16623–16644 (2024) 2, 4, 6, 7, 10, 11
82. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al.: Qwen3-omni technical report. arXiv preprint arXiv:2509.17765 (2025) 3, 9
83. Xue, H., Hang, T., Zeng, Y., Sun, Y., Liu, B., Yang, H., Fu, J., Guo, B.: Advancing high-resolution video-language representation with large-scale video transcriptions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5036–5045 (2022) 10
84. Xue, Z., Zhang, J., Hu, T., He, H., Chen, Y., Cai, Y., Wang, Y., Wang, C., Liu, Y., Li, X., et al.: Ultravideo: High-quality uhd video dataset with comprehensive captions. arXiv preprint arXiv:2506.13691 (2025) 4, 10
85. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 3, 8, 9
86. Yang, S., Wang, Z., Yang, X., Zhang, S., Kong, X., Wu, T., Zhao, X., Zhang, R., Zhao, A., Rao, A.: Shotverse: Advancing cinematic camera control for text-driven multi-shot video creation. arXiv preprint arXiv:2603.11421 (2026) 5
87. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: The Thirteenth International Conference on Learning Representations (2025) 1, 4
88. Yuan, S., He, X., Deng, Y., Ye, Y., Huang, J., Lin, B., Ma, C., Luo, J., Yuan, L.: Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025) 10
89. Yuan, S., Huang, J., Xu, Y., Liu, Y., Zhang, S., Shi, Y., Zhu, R., Cheng, X., Luo, J., Yuan, L.: Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. Advances in Neural Information Processing Systems **37**, 21236–21270 (2024) 5
90. Yuan, S., Yin, Y., Li, Z., Huang, X., Yang, X., Yuan, L.: Helios: Real real-time long video generation model. arXiv preprint arXiv:2603.04379 (2026) 5
91. Zhang, H., Wu, D., Liu, B., Zhong, L., Wei, Y., Ye, X., Liu, N., Liang, Y.: Muss: A large-scale dataset and cinematic narrative benchmark for multi-shot subject-to-video generation. arXiv preprint arXiv:2604.23789 (2026) 5
92. Zhang, K., Jiang, L., Wang, A., Fang, J.Z., Zhi, T., Yan, Q., Kang, H., Lu, X., Pan, X.: Storymem: Multi-shot long video storytelling with memory. arXiv preprint arXiv:2512.19539 (2025) 5, 14, 17, 19
93. Zhang, P., Jia, Z., Liu, K., Weng, S., Li, S., Shi, B.: Stage: Storyboard-anchored generation for cinematic multi-shot narrative. arXiv preprint arXiv:2512.12372 (2025) 5, 14, 17, 19
94. Zhang, Q., Cao, Y., Gao, Y., Min, X.: Vidaudio-bench: Benchmarking v2a and vt2a generation across four audio categories. arXiv preprint arXiv:2604.10542 (2026) 5
95. Zhang, R., Yu, B., Min, J., Xin, Y., Wei, Z., Shi, J.N., Huang, M., Kong, X., Xin, N.L., Jiang, S., et al.: Generative ai for film creation: A survey of recent advances. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 6267–6279 (2025) 1
96. Zhang, Y., Li, Z., Wang, D., Zhang, J., Zhou, D., Yin, Z., Dai, X., Yu, G., Li, X.: Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. arXiv preprint arXiv:2507.09862 (2025) 4, 10
97. Zhao, C., Liu, M., Wang, W., Chen, W., Wang, F., Chen, H., Zhang, B., Shen, C.: Moviedreamer: Hierarchical generation for coherent long visual sequence. arXiv preprint arXiv:2407.16655 (2024) 5
98. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Gu, L., Zhang, Y., He, J., Zheng, W.S., et al.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025) 5
99. Zheng, M., Xu, Y., Huang, H., Ma, X., Liu, Y., Shu, W., Pang, Y., Tang, F., Chen, Q., Yang, H., et al.: Videogen-of-thought: Step-by-step generating multi-shot video with minimal manual intervention. arXiv preprint arXiv:2412.02259 (2024) 5, 14, 17, 19
100. Zhou, Y.H., Li, H., Lin, R., Huang, H., Zhou, J., Yuan, C., Lan, T., Zhou, Z., Li, Y., Xu, J., et al.: Mtavg-bench: A comprehensive benchmark for evaluating multi-talker dialogue-centric audio-video generation. arXiv preprint arXiv:2602.00607 (2026) 5
101. Zhou, Z., Lai, Z., Wang, R., Yang, Y., Xing, Z., Yang, Y., Dai, Q., Qiu, L., Luo, C.: Avgen-bench: A task-driven benchmark for multi-granular evaluation of text-to-audio-video generation. arXiv preprint arXiv:2604.08540 (2026) 4, 5