CINEMA-BENCH: BENCHMARKING CINEMATIC EXPRESSIVENESS CAPABILITIES OF MULTI-SHOT VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in video generation have moved toward multi-shot creation, approaching professional video production. However, the field still lacks a clear definition of cinematic expression capabilities and quantitative evaluation protocols to benchmark the progress. To fill this gap, we present **CINEMA-Bench** (**CIN**ematic **E**xpressiveness in **M**ulti-shot Video **A**ssessment), a reference-grounded benchmark for evaluating the cinematic expression of multi-shot video generation models. Inspired by established film theory, we propose a hierarchical "Vocabulary–Grammar–Rhythm" taxonomy spanning three levels and 21 dimensions, covering shot-level camera language, inter-shot continuity grammar, and sequence-level temporal perception. We construct a multi-shot hierarchical prompt suite derived from real-world films and TV series across Mikels' eight emotions. We further design automatic metrics validated by human alignment to assess cinematic composition from the Director's Perspective, and complement them with user studies from the Audience's Perspective. Systematic evaluation of six state-of-the-art multi-shot video generation models reveals that, *despite reasonable use of cinematic elements, current methods still exhibit notable gaps in inter-shot continuity, rhythm and emotional delivery, shot-count controllability, and audience perception*, offering actionable insights for cinematically expressive video generation. This work presents an early exploration toward a comprehensive definition and quantification of cinematic expressiveness, bridging artistic expressiveness and computational evaluation.

1 INTRODUCTION

Recent advances in video generation have shifted the focus from single-shot synthesis to multi-shot video creation. Beyond the visual fidelity emphasized in single-shot generation (Yang et al., 2025; Wang et al., 2023; Blattmann et al., 2023; OpenAI, 2024), multi-shot generation introduces new aspects such as cinematic transitions and editing patterns (Kling et al., 2025; Seedance et al., 2026; Guo et al., 2025; Wu et al., 2026a;b; Wang et al., 2026). However, the field lacks both a clear definition of cinematic expression capabilities and quantitative evaluation protocols for benchmarking progress. This absence makes it difficult to assess the cinematic quality of generated multi-shot videos and to guide future advancement toward cinematic multi-shot video generation.

In professional filmmaking, directors consider not only the narrative content but also how cinematographic elements serve storytelling and emotional conveyance (Bowen, 2023; Pearlman, 2019; Smith et al., 2005). Different shot language (e.g., shot scale, camera angle, color) can convey varying semantic content and emotions, for example establishing shots introduce environments or close-ups reveal details, low-angle shots convey power or high-angle shots suggest vulnerability. Visual rhythm further shapes emotional tone, for instance, mild pacing creates calm, while intense pacing builds excitement. Therefore, given the essential role of these cinematographic elements in shaping audience perception, a critical question emerges: *can current multi-shot video generation models achieve such cinematic expression, and how should this capability be quantitatively measured?*

Existing benchmarks on video generation primarily focus on visual fidelity of single-shot videos, such as visual quality, prompt alignment (Huang et al., 2024b; Liu et al., 2024; Sun et al., 2025).

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

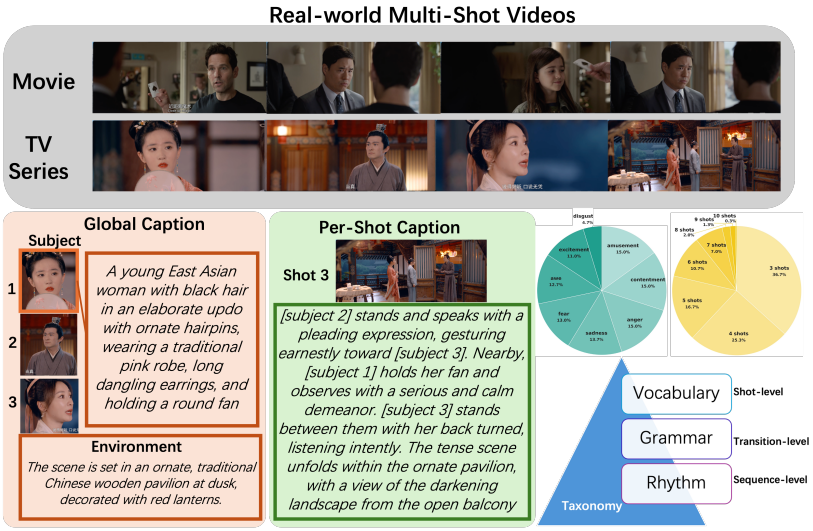


Figure 1: Overview of **CINEMA-Bench**. We collect multi-shot prompts derived from real professional videos, and evaluate the cinematic expression capability of multi-shot video generation models via a hierarchical Vocabulary-Grammar-Rhythm taxonomy spanning 21 dimensions, from both the Director's Perspective and the Audience's Perspective. The original professional videos span eight Mikels' emotion categories and cover length from 3 to 10 shots.

Recently, several multi-shot video benchmarks such as MSVBench (Shi et al., 2026) and LoCoT2V-Bench (Zheng et al., 2025), extend evaluation to temporal aspects like character/background consistency, motion plausibility etc. However, these efforts still center on technical fidelity across temporal axis rather than fine-grained cinematic criteria in professional video creation. Moreover, they concern textual prompt alignment without reference to professional cinematic standards. This makes it infeasible to assess whether a model's cinematic choices reach the level of real filmmaking practice.

To fill this gap, we present **CINEMA-Bench** (**CIN**ematic **E**xpressiveness in **M**ulti-shot Video **A**ssessment), a novel benchmark for evaluating the cinematic expression capabilities of multi-shot video generation models. Drawing on established film theories (Bowen, 2023; Pearlman, 2019; Smith et al., 2005), we propose a hierarchical cinematographic-aware taxonomy organized as ***Vocabulary, Grammar, and Rhythm***, which systematically characterizes cinematic expression at three levels, 21 dimensions: ***shot-level camera language*** (e.g., scale, angle, movement, lighting), ***inter-shot continuity grammar*** (e.g., space continuity, motion continuity, visual style continuity), and ***sequence-level temporal perception*** encompassing both compositional assessment (e.g., narrative coherence, shot language use, rhythm-emotion alignment) and audience preception (i.e., narrative comprehension, cognitive engagement, emotional resonance). Moreover, we collect 300 real-world multi-shot videos spanning eight Mikels' emotion categories (Mikels et al., 2005) and construct multi-shot prompts with hierarchical metadata, including global caption, shot-level descriptions and cinmeatic labels. This enables structured, emotion-aware evaluation grounded in professional filmmaking practice. Based on these real-film-derived prompts, we investigate the cinematic expression capabilities of six latest multi-shot video generation models. We design automatic metrics to quantitively assess cinematic composition in Director's Perspective, and we further conduct human study to investigate the impact on viewer's perception from Audience's Perspective, together providing comprehensive analysis.

Experiment results reveal that 1) *current methods achieve reasonable cinematic elements use but still struggle with inter-shot continuity and the delivery of rhythm and emotional atmosphere*; 2) *most methods lost the shot-count controllability and faces performance collapse when generating videos with more than **six-shots** prompts*; 3) *generated videos still fall far short of professional productions on viewers' perception such as cognitive engagment and emotional resonance*. These findings highlight concrete directions for future cinematically expressive generation. The reliability of proposed taxonomy framework is validated by comprehensive human alignment analysis in section 4.2.3. Our contributions are summarized as follows:

- We construct a reference-grounded benchmark of 300 multi-shot prompts derived from real-world films and TV series across Mikels' eight emotions, enriched with hierarchical metadata (global caption, shot-level captions, and cinematic labels), enabling structured evaluation against professional filmmaking standards.
- Grounded in film theories, we introduce, to the best of our knowledge, the first comprehensive benchmark for cinematic expressiveness, organized by a hierarchical Vocabulary-Grammar-Rhythm taxonomy over 3 levels and 21 dimensions, each with a human-aligned automatic metric. We further complement it with Audience-Perspective human studies, together enabling holistic evaluation.
- We systematically evaluate six state-of-the-art multi-shot video generation models and reveal the insufficeny in rhythm conveyance and multi-shot count control, providing actionable insights for future development of cinematically expressive video generation.

2 Related Work

Multi-Shot Video Generation. Early video generation mostly focus on single-shot synthesis (Yang et al., 2025; Wang et al., 2023), such as Stable Video Diffusion (Blattmann et al., 2023), Sora (OpenAI, 2024). With the recent advent of multi-shot capable models such as Kling (Kling et al., 2025) and Seedance 2.0 (Seedance et al., 2026), attention has gradually shifted toward multi-shot generation, emphasizing cinematic transitions and editing patterns—the very essence of cinematic storytelling. One two-stage paradigm is the hierarchical pipeline, where an LLM decomposes a story into sequential prompts and a text-to-video model generates each shot independently (Zhao et al., 2025; Zheng et al., 2024; Long et al., 2024; Huang et al., 2025). Another two-stage strategy is keyframe-based generation, such as StoryDiffusion (Zhou et al., 2024), IC-LoRA (Huang et al., 2024a), and Captain Cinema (Xiao et al., 2026), which first creates consistent keyframes and then generates the clips conditioned on them. More recently, holistic generation emerges as a distinct paradigm that jointly models all shots in a unified process to inherently enforce global consistency (Guo et al., 2025; Wu et al., 2026a). such as Holocine (Meng et al., 2026),CineTrans (Wu et al., 2026b), MultiShotMaster (Wang et al., 2026). However, the field still lacks both a clear definition and quantitative evaluation protocols to benchmark the cinematic expression capabilities.

Video Evaluation. Existing video generation benchmarks primarily focus on the visual fidelity of single-shot videos like prompt alignment, and motion quality (Huang et al., 2024b; Liu et al., 2024; Sun et al., 2025). Recently, there emerge several benchmarks for multi-shot video generation (Zheng et al., 2025). For example, MSVBench (Shi et al., 2026) provides hierarchical scripts and reference images, measuring story-video alignment, character/background consistency, etc. Cine250K (Wu et al., 2026b) and ShotWeaver40K (Wu et al., 2026a) contribute multi-shot datasets with hierarchical captions and transition annotations. Despite these, existing multi-shot benchmarks largely emphasize temporal fidelity rather than cinematic expressiveness—*whether shot language, transitions, and rhythm effectively serve narrative and emotional conveyance*. Cine250K and ShotWeaver40K address only partial elements such as transition types or inter-shot consistency, lacking a systematic definition and reliable metrics of cinematic expressiveness. Moreover, most benchmarks are limited to fewer than five shots, insufficient for evaluation on longer cinematic video generation.

3 CINEMA-Bench

3.1 Multi-Shot Prompt Suite

Existing video generation benchmarks often rely on LLM-generated prompts that lack grounding in real cinematic content, making it difficult to measure the gap between generated videos and professional filmmaking. To address this, we construct a prompt suite derived from real multi-shot videos, providing multi-level prompt and metadata that serve as the foundation for systematic evaluation.

3.1.1 Multi-Shot Video Collection

Definition. We define a *multi-shot sequence* as a series of consecutive shots that depict a coherent sub-event involving one or more subjects within a consistent scene. It lies between a single shot

and a full dramatic scene in film theory, representing a self-contained fragment of continuous action or event. Under this definition, a valid multi-shot sequence must satisfy two criteria: *scene consistency* (shared spatial setting across shots) and *semantic coherence* (logical continuity of action or narrative).

Data Sources. We collect raw videos from films and TV series that exhibit professional-grade cinematography, lighting, and editing. From these, we segment and curate **300 multi-shot sequences**, each containing 3-10 shots. The collection is relatively balanced across eight Mikels emotion categories (Mikels et al., 2005; Qiu et al., 2026) (i.e., amusement, awe, contentment, excitement, anger, disgust, fear, sadness; disgust appears less often, as it is seldom a dominant tone in film) to ensure diversity in mood and tone (see Figure 1).**Segmentation Pipeline.** We first apply TransNetV2 (Soucek & Lokoc, 2024) to detect shot boundaries in the raw videos. Starting from these individual shots, we progressively merge adjacent shots into multi-shot sequences based on inter-shot relevance. Specifically, we compute CLIP feature cosine similarity between consecutive shots: pairs with similarity above 0.85 are directly merged, and those below 0.5 are directly split. For intermediate cases (similarity between 0.5 and 0.85), we invoke Qwen3-VL-XB to reason about semantic continuity and make the final merge-or-split decision.

3.1.2 Multi-Shot Prompt Construction

Based on the collected multi-shot videos, we extract multi-level textual annotations that capture both narrative content and cinematic features. The annotation hierarchy consists of three components:

(1) **Global Caption.** Global Caption captures the subject captions and the enviroment description, aligning with most latest multi-shot video generation models' input format. For subject caption, we identify up to five prominent subjects and provide a detailed appearance description for each. For human subjects, the description covers demographics, hair, facial features, clothing, and build. For non-human subjects (vehicles, animals, objects), it covers type, color, material, and distinctive visual features. The scene setting and overall visual style are also described. (2) **Shot-Level Captions.** Each individual shot is annotated with a concise description covering subjects' expressions and actions, and scene background. Subject references use consistent identifiers linked to the subject captions, enabling cross-shot entity tracking. (3) **Cinematic Labels.** We further automaticly annotate each shot with structured cinematic attributes including shot scale, camera angle, camera movement, and lighting (detailed definitions are provided in Appendix). **Annotation and Validation.** All captions and cinematic labels are extracted using Gemini-2.5-Pro (Comanici et al., 2025). To verify annotation reliability, we invite professional filmmakers to manually annotate a randomly sampled subset of 13.7% of the data. The details of human-model alignment can be found in Table 4, confirming the trustworthiness of the automatic annotations.

3.2 Taxonomy

Just as a sentence is composed of words, governed by grammar, and shaped into rhythm and tone, a multi-shot video follows an analogous structure. Its "words" are the cinematic elements of individual shots such as semantics, shot scale, camera angle, movement, lighting,etc. Its "grammar" is the set of continuity editing principles that bind shots together including spatial continuity, motion continuity, and visual style consistency. Its "rhythm" emerges at the sequence level, orchestrating pacing and transitions to deliver a unified cognitive and emotional experience to the viewer. Guided by this analogy and film theories (Bowen, 2023; Pearlman, 2019; Smith et al., 2005), we propose a three-level "Vocabulary-Grammar-Rhythm" taxonomy comprising **21 dimensions** to systematically characterize cinematic expression in multi-shot videos:

3.2.1 Vocabulary: Shot-Level Camera Language

Each shot in a multi-shot video communicates meaning not only through what it depicts, but also through *how* it is filmed. As (Bordwell et al., 2008) note, every camera decision such as scale, angle, movement, lighting, shapes how the audience perceives plot, characters, space, and emotion. We characterize this ***shot-level vocabulary along 9 dimensions***: (1) **Semantics.** We decompose shot content into three sub-dimensions: *subjects* (the narrative or visual focal points such as persons, animals, vehicles, buildings, landscapes, etc.), *environment* (location type, time of day, weather), and *events/actions* (key interactions or state changes). Together they form the narrative foundation of

each shot and determine the storyline being conveyed. (2) **Shot Scale.** The framing distance from the subject, ranging from wide shot to extreme close-up (5 categories). Shot scale regulates psychological distance of viewers, for exmple, wide shots situate subjects within the environment to establish context or convey a sense of isolation, while close-ups could reveal facial emotion and create intimacy (Balazs, 1970). (3) **Camera Angle.** The vertical and lateral orientation of the camera relative to the subject's center line. Angle encodes power relations between the subjects and viewers: a low angle lends dominance to the subject, a high angle diminishes it, and a Dutch angle signals psychological instability (Bordwell et al., 2008). (4) **Camera Movement.** This captures the movement of camera relative to the subject. Movement directs viewer attention and creates spatial energy, for instance, a slow push-in builds tension, while a static frame invites contemplation (Brown, 2021). (5) **Motion Speed.** The perceived kinetic energy of the shot, assessed as the *combined* effect of camera movement and subject movement. Slow motion can heighten dramatic weight, while fast motion conveys urgency or chaos. (6) **Composition.** The spatial arrangement of elements within the frame. Composition guides the eye and expresses narrative tone. A symmetry can suggest order and formality, while asymmetry introduces tension or dynamism. (7) **Color Tone.** The overall chromatic quality of the shot. Color tone functions as an emotional cue: warm palettes evoke intimacy or comfort, while cool palettes suggest detachment or melancholy. (8) **Lighting Direction.** The primary direction of light on the subject. Lighting shapes mood and character perception. Side light creates mystery and depth, back light produces silhouettes that suggest isolation or concealment (Brown, 2021). (9) **Shot Aesthetics.** A holistic score assessing how well the above elements work together as a unified visual design. This reflects the overall craft quality that determines whether a shot feels intentional and expressive.

We leverage automatic labeling method and query Gemini-2.5-Pro (Comanici et al., 2025) with tailored prompts to infer each cinematic labels and the full categories and prompts are provided in the Appendix.

3.2.2 GRAMMAR: TRANSITION-LEVEL CONTINUITY LOGIC

Continuity editing is the essential grammar that binds adjacent shots into a seamless visual flow (Bordwell et al., 2008). Classical continuity editing theory prescribes a set of rules, such as match on action, the 180-degree rule, eyeline match, consistent lighting and color across cuts. They are designed to maintain spatiotemporal coherence and prevent viewer disorientation. From these established principles (Smith et al., 2005) and prior works (Wang et al., 2019; Zhang et al., 2025), we distill **6 binary dimensions** that capture the most common visual continuity errors between adjacent shots. These dimensions are grouped into three continuity categories, plus an overall **transition score**:

Spatial Continuity. We assess two aspects: (1) ***Jump cut***: A jump cut occurs when two adjacent shots frame the same subject/scenario from nearly identical angles with only a slight change in scale or position, violating the traditional 30-degree rule. It results in an abrupt, jarring skip that breaks the illusion of continuous action/event. This dimension evaluates *whether slight changes in shot scale or angle produce unnatural edits*. We calculate it by the composition similarity between the last k frames of the preceding shot s_k and the first k frames of the subsequent shot s_l (k=3). For each frame f, we extract three color-invariant descriptors: a locally-normalized grayscale layout $\phi_s(f)$, a Canny edge map $\phi_e(f)$, and a perceptual hash $\phi_p(f)$. Averaging each descriptor over the sampled frames of s_k and s_l gives $\bar{\phi}_*^k$ and $\bar{\phi}_*^l$, and the jump-cut score is $G_{jc}(s_k, s_l) = \sum_{*\in\{s,e,p\}} w_* \,\mathrm{sim}\big(\bar{\phi}_*^k,\, \bar{\phi}_*^l\big)$, where the weights (w_s, w_e, w_p)=$(0.45, 0.35, 0.20)$ are assigned in decreasing order of each descriptor's discriminability for framing continuity. (2) ***Shot scale transition rationality***: A scale gap larger than three levels between adjacent shots (e.g., a wide shot followed by an extreme close-up) is too abrupt for the audience to preserve spatial context. This dimension therefore measures *whether the scale difference between adjacent shots exceeds three levels*. Each shot scale label is mapped to an ordinal level $\ell(\cdot) \in \{1.0,\, 2.0,\, 3.0,\, 4.0,\, 5.0\}$ (from extreme close-up to wide shot), and the transition between adjacent shots s_k and s_l is scored as $G_{\text{scale}}(s_k, s_l) \;=\; \big|\ell(s_k) - \ell(s_l)\big|$.

Motion Continuity. We assess: (1) ***Camera movement direction consistency***: whether opposing motion directions across a cut cause disorientation. For each shot s_k, we track sparse feature points with optical flow and compute the mean motion vector $\bar{\mathbf{v}}_k \;=\; \frac{1}{N_k}\sum_{i=1}^{N_k} \mathbf{v}_k^{(i)}$. The direction difference is measured by the angle between the mean vectors of two adjacent shots:

$G_{move}(s_k, s_l) = \arccos\Big(\frac{\bar{\mathbf{v}}_k \cdot \bar{\mathbf{v}}_l}{\|\bar{\mathbf{v}}_k\| \, \|\bar{\mathbf{v}}_l\|}\Big)$. (2) ***Motion intensity consistency***: evaluates whether the movement speeds of two shots are mismatched. It is computed as the normalized difference between the magnitudes of the mean motion vectors: $G_{\text{intensity}}(s_k, s_l) = \big|\, \|\bar{\mathbf{v}}_k\| - \|\bar{\mathbf{v}}_l\| \,\big| / \big|\max(\|\bar{\mathbf{v}}_k\|,\ \|\bar{\mathbf{v}}_l\|)$,

Visual Style Consistency. We assess ***Color tone consistency*** as *whether adjacent shots maintain coherent hue, and saturation*. For each pair of adjacent shots, we extract their middle frames f^k_{mid} and f^l_{mid}, compute normalized 2D Hue-Saturation histograms $H(\cdot)$ in HSV space, and score color consistency by their correlation $G_{\text{color}}(s_k, s_l) = \rho\big(H(f^k_{mid}),\, H(f^l_{mid})\big)$ where $\rho(\cdot, \cdot)$ denotes the histogram correlation coefficient.

Finally, we leverage Gemini-2.5-Pro to provide a holistic **Transition Score** of each transition $G_{\text{trans}}(s_k, s_l)$, offering an overall judgment based on the three criteria above.

3.2.3 RHYTHM: SEQUENCE-LEVEL TEMPORAL PERCEPTION

Rhythm arises from how visual elements such as shot scale, camera angle, camera movement, color tone, are arranged and varied across a sequence of shots (Bowen, 2023; Pearlman, 2019; Adams et al., 2000). These patterns guide the viewer's attention and emotion through alternating *cycles of tension and release*, effectively aligning the audience's internal experience with the narrative and intended emotions. To assess rhythm quality, we adopt two complementary perspectives. The ***director's perspective*** evaluates *how* cinematic tools are deployed, while the ***audience's perspective*** measures *whether* the intended effect is achieved—together providing a complete picture of sequence-level cinematic expression.

Director's Perspective. These dimensions assess the craft behind the sequence:

Narrative Coherence: evaluates whether a sequence maintains a coherent storyline in terms of visual continuity. Building on the holistic Transition Score $G_{\text{trans}}(s_k, s_l)$ defined in Sec. 3.2.2, we model the narrative coherence of a sequence $\mathcal{S} = \{s_1, \ldots, s_T\}$ as the average transition score over all adjacent shot pairs: $G_{\text{narr}}(\mathcal{S}) = \frac{1}{T-1} \sum_{k=1}^{T-1} G_{\text{trans}}(s_k, s_{k+1})$.

Cinematographic Diversity: evaluates the ***richness*** of cinematic "vocabulary" (such as shot scale, angle, composition) and the and ***evenness*** (how uniformly they are distributed). Much as a skilled speaker is distinguished not by vocabulary size alone but by the ability to deploy it flexibly, a model that commands a broad palette of cinematic elements without over-relying on a narrow subset can be regarded as having strong cinematographic expression. Specifically, let C_d denote the number of pre-defined categories in cinematic dimension d. (1) ***Richness (per video).*** For a single generated video v containing N_v shots, we count the number of distinct categories U^v_d that appear in dimension d and compute

$$R^v_d \;=\; \frac{U^v_d}{C_d}\,, \tag{1}$$

The method-level richness is the average over all V videos: $\bar{R}_d = \frac{1}{V} \sum_{v=1}^{V} R^v_d$. (2) ***Evenness (across all videos).*** We pool the per-shot labels of dimension d from every generated sequence and compute the Pielou evenness (Pielou, 1966):

$$E_d \;=\; \frac{H_d}{\ln C_d}\,, \qquad H_d = -\sum_{k=1}^{C_d} p_k \ln p_k\,, \tag{2}$$

where $\{p_k\}$ is the category frequency distribution obtained by aggregating all shots across videos. E_d=1 when categories are used uniformly; $E_d \to 0$ when the model collapses to a single category. ***CineDiversity*** is calculated by fusing these two axes, and α=0.5 balances richness and evenness: $\text{CineDiversity} \;=\; \frac{1}{D} \sum_{d=1}^{D} \Big[\, (1-\alpha)\ \bar{R}_d \;+\; \alpha\ E_d \,\Big]$.

Visual Rhythm Rationality: chacterizes the machness between the pacing of visual changes and the intended emotional tone of the narrative. Rapid shot-scale change paired with accelerating motion can convey *excitement*, whereas sustained framing with warm tones and stable lighting evokes *contentment*. Viusal Rhythm Rationality evaluates *whether the visual rhythm of a generated sequence plausibly embodies its target emotion*. We first quantify rhythm into interpretable features, then ask an LLM judge to score the match between those features and the intended emotion. ***Stage 1: Rhythm quantification.*** Each shot s_t in a T-shot sequence is encoded into $\mathbf{x}_t \in \mathbb{R}^7$ by mapping

its seven cinematic dimensions (here we do not consider the shot duration, since current video generation models usually take the duration as parameter rather than flexible self-decision) to numeric scales that preserve perceptual distance.[1] At each cut $t \to t{+}1$ the normalised change per dimension is

$$\delta_d^{(t)} = \begin{cases} |\hat{x}_d^{(t+1)} - \hat{x}_d^{(t)}| \,/\, \Delta_d^{\max} & d \text{ is ordinal,} \\ \mathbb{1}[c_d^{(t+1)} \neq c_d^{(t)}] & d \text{ is nominal,} \end{cases} \tag{3}$$

Ordinal dimensions use the absolute difference divided by the dimension's range; nominal dimensions use a binary switched-or-not indicator, so that all values lie in $[0, 1]$. From these we derive three sequence-level descriptors: the *per-dimension change rate* $\bar{\delta}_d = \frac{1}{T-1}\sum_t \delta_d^{(t)}$, capturing how briskly dimension d evolves; the *overall rhythm intensity* $\rho = \frac{1}{T-1}\sum_t \big(\frac{1}{D}\sum_d \delta_d^{(t)}\big)$, the mean cross-dimension change per cut; and the *rhythm variability* $\sigma_\rho = \mathrm{std}\big(\{\frac{1}{D}\sum_d \delta_d^{(t)}\}_t\big)$, which measures variability of pacing. These form the *rhythm profile* of a sequence. ***Stage 2: LLM-based Scoring on Emotion-Rhythm Alignment.*** The rhythm profile and the target narrative emotion are presented to an LLM judge, which assesses whether the rhythm plausibly conveys intended emotion and rates a score $s_v \in \{1, 2, 3, 4\}$ (*mismatch / weak / good / strong*) with structured evidence.

Audience's Perspective. Grounded in the Narrative Engagement Measure (Busselle & Bilandzic, 2009), Transportation Theory (Green & Brock, 2000), and Cognitive Film Theory (Bordwell, 1985), we summarize three dimensions to investigate the viewer's perception for the overall sequence: ***Narrative Comprehension***: reflects the viewer's success in constructing a coherent mental model of characters, events, and their causal relations, following the situation model account of comprehension (Busselle & Bilandzic, 2009; Zwaan & Radvansky, 1998). ***Cognitive Engagement*:** measures the degree to which the video captures the viewer's cognitive resources, driven by information gaps that arouse curiosity (Loewenstein, 1994; Bordwell, 1985; Green & Brock, 2000; Busselle & Bilandzic, 2009). ***Emotional Resonance*:** reflects the extent to which the shot sequence evokes emotions aligned with the story situation, such as empathy toward characters and the arousal of mood-consistent affect (Busselle & Bilandzic, 2009; Tan, 2013; Smith, 2003).

4 Experiments

4.1 Setup

Models. We evaluate six multi-shot video generation methods across three paradigms. 1) Pretrained Video Diffusion. We test Wan2.2-14B (Wan et al., 2025) using a hierarchical prompt. 2) Two-Stage Keyframe-to-Video. We implement a pipeline of StoryDiffusion (Zhou et al., 2024) + Wan2.2-14B, where StoryDiffusion generates a keyframe and Wan2.2-14B synthesizes the video clip conditioned on keyframes. 3) Holistic Multi-Shot Generation. We reproduce four recent models, including CineTrans (Wu et al., 2026b), HoloCine (Meng et al., 2026), and MultiShotMaster (Wang et al., 2026).and also leverage Seedance 2.0 (Seedance et al., 2026) to investigate cinematic expressiveness. More implemtation details see in Appendix.

4.2 Results and Analysis

4.2.1 Main Results

Aesthetic. We assess overall aesthetic quality by averaging the aesthetic scores across all shots. In Table 1, Seedance 2.0 achieves the highest aesthetic score of 3.92, clearly surpassing all open-source models. Among open-source methods, MultiShotMaster ranks first with a score of 3.32, followed by Wan2.2-14B and HoloCine. In contrast, the two-stage pipeline StoryDiffusion + Wan2.2-14B and CineTrans score much lower at 1.93 and 1.59, indicating that decoupled keyframe-to-video generation and earlier pretrained backbones compromise per-shot visual fidelity.

Narrative Coherence and Visual Coherence Error. In Table 1, Seedance 2.0 attains the highest narrative coherence score of 3.52, followed by MultiShotMaster at 3.35, both clearly surpassing Wan2.2-14B and HoloCine. The two-stage pipelines StoryDiffusion + Wan2.2-14B and CineTrans

[1] Full quantisation tables are provided in the appendix.

lag far behind at 1.52 and 1.20, suggesting that two-stage generation struggles to maintain narrative continuity across shots. For visual coherence errors, MultiShotMaster achieves the lowest 180-degree rule violation rate of 0.05 and a competitive intensity shift of 0.08, indicating strong control over inter-shot consistency. Notably, no model exhibits jump cuts, one critical type of visual error.

Table 1: Main results across six multi-shot generation models. Aesthetic, Coherence, and Rhythm are rated on a 1-4 scale; Cinematic metrics and Visual Errors are normalized to [0,1] (lower is better for Visual Errors). The best and second-best results are **bolded** and underlined, respectively.

Method	Aesthetic	Coherence	Cinematic			Rhythm	Visual Error (↓)				
			Richness	Evenness	Diversity		Jump	Scale	Oppos	Intens	Color
Wan2.2-14B	3.17	3.03	**0.82**	0.64	**0.73**	2.76	0.00	0.02	0.17	0.21	0.59
StoryDiffusion + Wan2.2-14B	1.93	1.52	0.51	0.62	0.57	1.70	0.00	0.01	0.22	0.30	0.69
CineTrans	1.59	1.20	0.43	0.57	0.50	1.87	0.00	**0.00**	0.13	0.10	**0.29**
HoloCine	3.17	2.83	0.55	0.64	0.60	2.70	0.00	0.02	0.14	0.23	0.34
MultiShotMaster	3.32	3.35	0.47	0.58	0.52	2.94	0.00	0.03	**0.05**	**0.08**	0.61
Seedance 2.0	**3.92**	**3.52**	0.53	**0.70**	0.62	**3.04**	0.00	0.02	0.16	0.25	0.51

Cinematic Language Use. As shown in Table 1, Wan2.2-14B unexpectedly scores high on both metrics. A plausible explanation is that its generated videos, though often short, still contain diverse cinematic elements, much like a concise speaker with a rich lexicon. Seedance achieves the highest evenness with moderate richness, suggesting that it varies cinematic vocabulary little within a single video but draws from a broad vocabulary pool across videos. This resembles a person who uses simple words in each conversation yet switches to different ones across topics. StoryDiffusion also attains competitive richness and evenness, indicating that two-stage pipelines tend to favor diverse shot language, though their weak visual coherence limits overall cinematic expressiveness.

Rhythm Rationality. From Table 1, Seedance 2.0 achieves the highest Rhythm score of 3.04, indicating that it can effectively organize visual rhythm to convey the intended emotional atmosphere. MultiShotMaster follows closely with a score of 2.94, also showing strong rhythmic control. In contrast, StoryDiffusion + Wan2.2-14B obtains the lowest score of only 1.70. Results reveal that emotion-aware visual rhythm decision-making remains an underexplored capability for most existing methods, posing a critical open problem for future multi-shot video generation.

4.2.2 Long Video Generation Capabilities

Table 2: Performance change when extending to longer videos (7–10 shots) compared with shorter ones (3-6 shots). Each cell reports the difference (long − short).

Method	Aesthetic	Coherence	Cinematic			Rhythm	Visual Error
			Richness	Evenness	Div		
Wan2.2-14B	-0.16	-0.18	-0.03	-0.05	-0.04	-0.08	0.05
Storydiff	-0.19	-0.07	-0.04	0.02	-0.01	-0.15	-0.02
CineTrans	-0.03	-0.14	0.02	-0.00	0.01	0.18	0.04
Holocine	-0.74	-0.67	0.16	-0.00	0.08	-0.38	-0.01
MultiShotMaster	-0.02	-0.08	-0.05	-0.00	-0.03	-0.10	0.02
Seedance 2.0	0.04	-0.13	-0.02	0.02	-0.00	0.04	0.01

Table 3: Shot-count control ability. Each row shows the number of shots generated by every method given a prompt specifying the target shot count.

Prompt shots	Generated shots					
	Wan 2.2-14B	Storydiff	Cinetrans	Holocine	MultiShotMaster	Seedance 2.0
3	1.76	3.00	2.98	2.93	2.96	3.06
4	1.65	4.00	3.79	3.58	3.97	4.05
5	2.16	5.00	4.48	4.44	4.90	5.12
6	2.13	6.00	5.44	4.38	5.87	6.19
7	2.00	7.00	5.81	2.76	6.91	7.19
8	1.83	8.00	4.50	2.33	6.83	8.00
9	2.00	9.00	6.50	1.25	8.75	10.50
10	3.00	10.00	2.00	7.00	8.00	10.00

Performance Degradation on Long-Video Generation. As shown in Table 2, almost all methods degrade when extended to longer videos of 7-10 shots, especially in narrative coherence and shot aesthetics. This shows that keeping visual quality and consistency across long sequences remains difficult. HoloCine drops the most, with its aesthetic score falling by 0.74 and coherence by 0.67, while Seedance 2.0 is the most robust on these dimensions and even gains slightly in aesthetics. Visual-coherence errors remains stable for most methods. These results indicate that future multi-shot generation research should explicitly address long-range consistency and potential error propagation to enable scalable long-video generation.

Multi-Shot Count Controllability. Table 3 shows that shot-count controllability varies widely across methods. StoryDiffusion achieves precise shot-count control, showing the advantage of Two-Stage Generation for strict controllability. Seedance 2.0 and MultiShotMaster closely track the shot counts with small deviations. Wan 2.2-14B collapses to fewer than 3 shots regardless of the target and HoloCine degrades severely when exceeding 6 shots. These reveal that most models can handle up to ***seven*** shots reliably, while generating longer videos remains a key challenge.

4.2.3 PROFESSIONAL EVALUATION

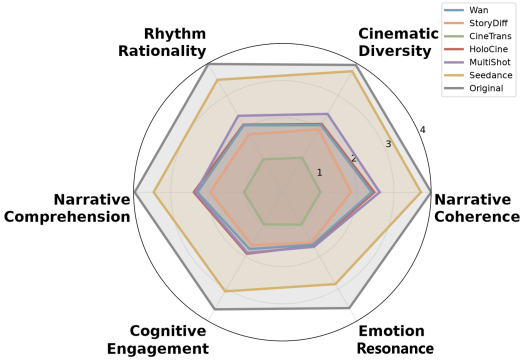


Figure 2: Human study from director's perspective(narrative coherence, cinematic diversity,rhythm rationality) and audience perspective (narrative comprehension, cognitive engagement, emotional resonance) (all in 1-4 scale).

Human Study. In Figure 2, it shows that the Original videos serve as the upper bound, with Seedance ranking second and clearly outperforming the other baselines. Among the rest, MultiShot leads especially on director-oriented dimensions and CineTrans ranks lowest overall. The results reveal a consistent ranking: Original $>$ Seedance $>$ MultiShot $>$ Wan $\approx$ HoloCine $>$ StoryDiff $>$ CineTrans, and also highlights a clear gap between current generation methods and real cinematic videos, particularly on higher-level dimensions such as emotional resonance and cognitive engagement. To ensure reliability, we obtain strong inter-annotator agreement all dimensions (Spearman's $\rho \in [0.67, 0.79]$, Kendall's $\tau \in [0.60, 0.71]$; see Table 6 in Appendix).

Human Alignment Analysis. To verify the reliability of our taxonomy, we conduct a human alignment study. We randomly sample 13.6% of the generated and original videos and ask experienced annotators to rate them on 18 predefined dimensions (excluding objective audience-side scores). Table 4 reports the Accuracy, Spearman's ρ, and Kendall's τ between automatic metrics and human annotations. The automatic labels align well with human judgments on nominal dimensions (e.g., 96.5% on semantic check and 97.1% on shot scale transition), and show consistent correlations on ordinal dimensions such as shot aesthetic score, transition score, and narrative coherence (up to $\rho = 0.666$). These results confirm that our taxonomy produces human-aligned assessments at the shot, transition, and sequence levels.

Table 4: Agreement between automatic and human annotations. Nominal dimensions are reported as per-class *accuracy*; ordinal ones are reported as Spearman's ρ and Kendall's τ. All values are shown as percentages (%).

Level	Nominal		Ordinal		
	Dimension	Acc. (%)	Dimension	ρ (%)	τ (%)
shot-level	shot size	65.60	shot aesthetic score	49.10	43.30
	camera angle	72.80			
	camera movement	70.10			
	motion speed	58.70			
	composition	62.20			
	color tone	56.80			
	light position	49.50			
	semantic check	96.50			
transition-level	jump cut	68.60	transition score	46.40	40.00
	shot scale trans.	97.10			
	motion direction	69.70			
	motion intensity	71.90			
	tone consistency	50.80			
sequence-level	–	–	narrative coherence	66.60	58.00
			cinematic diversity	37.40	28.40
			rhythm rationality	34.90	30.40

5 CONCLUSION AND FUTURE DIRECTION

This work provides the very first attempts to introduce a systematic definition and a evaluation protocal for cinematic experssiveness of multi-shot video generation. We evaluate representative multi-shot video generation methods along six aspects: aesthetic quality, narrative coherence, cinematic language use, rhythm-emotion conveyance, shot-count controllability, and audience perception. Our study leads to three main findings: First, most holistic multi-shot methods perform reasonably well on visual quality and coherence. Seedance 2.0 achieves the best overall performance, including shot-count controllability, followed closely by MultiShotMaster. Two-stage pipelines allow strict shot-count control, but unnatural inter-shot transitions clearly hurt coherence and rhythm. Unexpectedly, Wan2.2-14B fails to control shot count yet still produces competent cinematic language and rhythm, suggesting that cinematic patterns can be learned implicitly during pretraining. This may point to data-driven modeling of directorial techniques as a promising direction. Second, existing methods show clear performance drops and even collapse on longer videos. Modeling long-range consistency and reducing cross-shot error accumulation thus remain key challenges for scalable long-video generation. Finally, human study shows that generated videos still fall behind professional productions on higher-level perceptual dimensions such as rhythm-emotion conveyance, cognitive engagement, and emotional resonance. How to make a model direct like a professional filmmaker, by proactively selecting cinematic language, and shaping rhythm and emotional arcs, remains an open problem for future research.

REFERENCES

Brett Adams, Chitra Dorai, and Svetha Venkatesh. Study of shot length and motion as contributing factors to movie tempo. In *Proceedings of the eighth ACM international conference on Multimedia*, pp. 353–355, 2000.

Bela Balazs. Theory of the film; character and growth of a new art. 1970.

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

David Bordwell. *Narration in the fiction film*. Univ of Wisconsin Press, 1985.

David Bordwell, Kristin Thompson, and Jeff Smith. *Film art: An introduction*, volume 7. McGraw-Hill New York, 2008.

Christopher Bowen. *Grammar of the Edit*. Routledge, 2023.

Blain Brown. *Cinematography: Theory and practice: For cinematographers and directors*. Taylor & Francis, 2021.

Rick Busselle and Helena Bilandzic. Measuring narrative engagement. *Media psychology*, 12(4): 321–347, 2009.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Melanie C Green and Timothy C Brock. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology*, 79(5):701, 2000.

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17281–17291, 2025.

Kaiyi Huang, Yukun Huang, Xintao Wang, Zinan Lin, Xuefei Ning, Pengfei Wan, Di Zhang, Yu Wang, and Xihui Liu. Filmaster: Bridging cinematic principles and generative ai for automated film generation. *arXiv preprint arXiv:2506.18899*, 2025.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024a.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024b.

Team Kling, Jialu Chen, Yuanzheng Ci, et al. Kling-omni technical report, 2025. URL `https://arxiv.org/abs/2512.16776`.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149. IEEE, 2024.

George Loewenstein. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75, 1994.

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

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

Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior research methods*, 37(4):626–630, 2005.

OpenAI. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, 2024. OpenAI Technical Report.

Karen Pearlman. On rhythm in film editing. In *The Palgrave handbook of the philosophy of film and motion pictures*, pp. 143–163. Springer, 2019.

Evelyn C Pielou. The measurement of diversity in different types of biological collections. *Journal of theoretical biology*, 13:131–144, 1966.

Zongyang Qiu, Bingyuan Wang, Xingbei Chen, Yingqing He, and Zeyu Wang. Emovid: A multimodal emotion video dataset for emotion-centric video understanding and generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8612–8620, 2026.

Team Seedance, Liyang Chen De Chen, et al. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

Haoyuan Shi, Yunxin Li, Nanhao Deng, Zhenran Xu, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. Msvbench: Towards human-level evaluation of multi-shot video generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24034–24058, 2026.

Greg M Smith. *Film structure and the emotion system*. Cambridge University Press, 2003.

Tim Smith et al. *An attentional theory of continuity editing*. PhD thesis, University of Edinburgh, 2005.

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

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8406–8416. IEEE, 2025.

Ed S Tan. *Emotion and the structure of narrative film: Film as an emotion machine*. Routledge, 2013.

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

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

Miao Wang, Guo-Wei Yang, Shi-Min Hu, Shing-Tung Yau, Ariel Shamir, et al. Write-a-video: computational video montage from themed text. *ACM Trans. Graph.*, 38(6):177–1, 2019.

Qinghe Wang, Xiaoyu Shi, Baolu Li, Weikang Bian, Quande Liu, Huchuan Lu, Xintao Wang, Pengfei Wan, Kun Gai, and Xu Jia. Multishotmaster: A controllable multi-shot video generation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16268–16278, 2026.

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

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

Junfei Xiao, Ceyuan Yang, Lvmin Zhang, Shengqu Cai, Yang Zhao, Yuwei Guo, Gordon Wetzstein, Maneesh Agrawala, Alan Yuille, and Lu Jiang. Captain cinema: Towards short movie generation. In *International Conference on Learning Representations*, volume 2026, pp. 83349–83375, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Yuqi Zhang, Bin Guo, Nuo Li, Ying Zhang, Shijie Wang, Zhiwen Yu, and Qing Li. Tree-of-adeditor: heuristic tree reasoning for automated video advertisement editing with large language model. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 8705–8713, 2025.

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequences. In *International Conference on Learning Representations*, volume 2025, pp. 50060–50090, 2025.

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

Xiangqing Zheng, Chengyue Wu, Kehai Chen, and Min Zhang. Locot2v-bench: Benchmarking long-form and complex text-to-video generation. *arXiv preprint arXiv:2510.26412*, 2025.

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

Rolf A Zwaan and Gabriel A Radvansky. Situation models in language comprehension and memory. *Psychological bulletin*, 123(2):162, 1998.

APPENDIX

A DETAILS OF TAXONOMY

Automatic Labeling. Table 7 presents the detailed categories and definition for each dimension of shot-level camera language. Below are the algorithmic details for each dimension and transition score.

Camera Movement. We leverage automatic method to label the ***Camera Movement*** since MLLMs struggle to handle this dimension. We classify camera movement by decomposing background optical flow into translation, zoom, and rotation components, and assigning a label based on the dominant motion component. Specifically, we compute dense optical flow between sampled frame pairs and, on a background mask $\mathcal{B}$ formed by pixels with the smallest flow magnitudes, aggregate the mean translation (T_x, T_y), divergence S (zoom), and curl Θ (rotation) across the shot. The final label is then obtained by a normalized decision rule:

$$\text{Label} = \Phi\left(\underbrace{\frac{|T_x|}{\alpha W}, \frac{|T_y|}{\alpha H}}_{\text{pan/tilt}}, \underbrace{\frac{|S|}{\beta}}_{\text{push/pull}}, \underbrace{\frac{|\Theta|}{\gamma}}_{\text{arc}} \,\middle|\, f_{\text{still}}\right), \tag{4}$$

where α, β, γ are perceptual normalization constants, empirically set so that a normalized score of one corresponds to a just-noticeable camera motion (roughly 1% frame-width translation, 2.5% scale change, or 3° rotation), and f_{still} denotes the fraction of near-static pixels. $\Phi(\cdot)$ returns *Static* when the background is dominantly still or all normalized scores fall below one, and otherwise selects the movement type with the largest score.

Labeling for other Shot-level Dimensions and Transition Score. We leverage Gemini-2.5-Pro with tailored prompts to score other eight shot-level labels. Consideing the accuracy and token usage, we group them into four categories based on their visual nature: (1) ***Semantics***, (2) ***Camera Composition*** (shot scale, camera angle, motion speed, and composition), (3) ***Lighting and Color*** (color tone and lighting direction), and (4) ***Shot Aesthetics***. The detailed prompts are shown in xx.

Illustration of Rhythm Quantification. In Table 8, we provide the mapping numeric scales of each camera language, which preserve the perceptual distance. There are two types of mapping way: for labels like shot scale that have strong semantic order, we use a linear mapping; for labels like composition that have weak semantic order, we use a switch-indictor that counts the number of switches. and for camera angle that have categories with explicit order and categores without strong order, we combine these two mapping ways. We explain the process of rhythm quantification with an example in Figure 3 (for ease of illustration, the change intensity is reported here without normalization.).

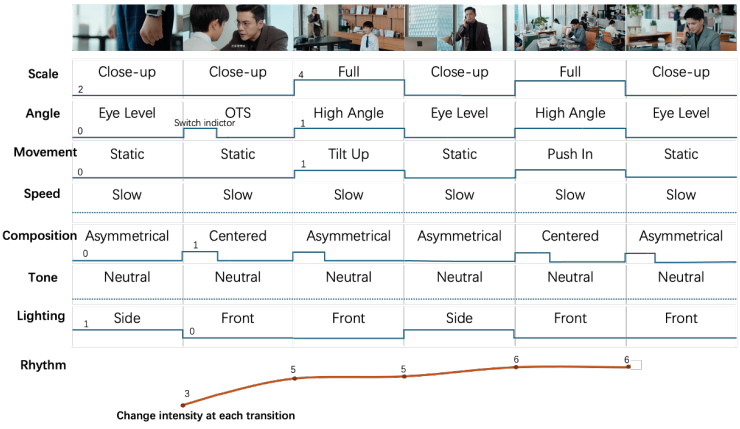


Figure 3: Illustration of the Rhythm Quantification Process. We can see that rhythm is impacted by diverse camera language changes.

B Implementation Details

To ensure a fair comparison, every model receives identical information (global caption and per-shot captions), formatted to match its input interface via either prompt concatenation or explicit structured inputs. Since most models require an explicit duration as parameter, we fix the length of each shot to 3 seconds, thereby eliminating shot duration as a confounding factor in video pacing. Due to model and GPU memory constraints, Wan2.2-14B is limited to min(3×N, 9) seconds, where N is the number of shots from referenced original videos. Seedance 2.0 imposes a maximum generation length of 15 seconds, so its duration is set to min(3×N, 15) seconds. Other methods use 3×N seconds. All multi-shot video generation experiments are performed on 8× NVIDIA H100 (80GB) GPUs. Visual coherence errors are binary dimensions whether the calculated score fulfills the criteria, or not. For each, $\mathbf{1}[G_{jc}(s_k, s_l) \geq \tau_{jc}]$, $\tau_{jc} = 0.55$; $\mathbf{1}[G_{\text{scale}}(s_k, s_l) > \tau_{scale}]$, $\tau_{scale} = 2$; $\mathbf{1}[G_{\text{move}}(s_k, s_l) > \tau_{\text{move}}]$, $\tau_{\text{move}} = 120°$; $\mathbf{1}[G_{\text{intensity}}(s_k, s_l) > \tau]$ with $\tau = \frac{2}{3}$, which means that the slower shot moves at less than one third of the faster shot's speed.

C Additional Results

Figure 4 shows the distribution of cinematic language across the eight emotions, and we also analyze the distributions. of rhythm intensity and variability across eight emotions in Figure 5.

D Human Annotation

We recruit two experts from Beijing Film Academy with professional annotation experience and over five years of video production expertise to conduct the human annotations for human alignment and expert study. The interface of human annotation is shown in Figure 6.

Annotations for Human Alignment. We randomly sample 13.6% of the generated and original videos for human annoations following specific guidelines (results shown in Table 4). The human check results for eight Mikels' emotions labeling is shown in Table 5.

Table 5: Overall Emotion Top-K Accuracy

	Top-1	**Top-3**	**Top-5**
Accuracy	0.2762	0.6503	0.8601

Details of Human Study. Before human study, we conduct an hour training on the scoring criteria shown in Table. We randomly sample 51 groups of videos, each containing outputs from the six methods along with the original reference, and ask the annotators to watch them in randomized order and rate independently, yielding scores across 51 videos × 7 methods × 6 dimensions. Table 6 reports the inter-annotator consistency, where all dimensions achieve strong and highly significant agreement (Spearman's $\rho \in [0.673, 0.789]$, Kendall's $\tau \in [0.603, 0.714]$), with an aggregated $\rho = 0.726$ and $\tau = 0.646$ over all 2,142 paired ratings, confirming the reliability of our evaluation.

Table 6: Inter-annotator agreement. We report consistency over 51 videos × 7 methods. All correlations are highly significant ($p < 0.001$).

Dimension	**Spearman ρ**	**Kendall τ**
Narrative coherence	0.76	0.68
Cinematic diversity	0.75	0.68
Rhythm appropriateness	0.79	0.71
Narrative comprehension	0.72	0.65
Cognitive engagement	0.70	0.62
Emotion resonance	0.67	0.60
All 6 dims (combined)	**0.73**	**0.65**

E THE USE OF LARGE LANGUAGE MODELS

We use large language models for polishing the writing, and we have conducted a careful check, taking full responsibility for all content in this work.

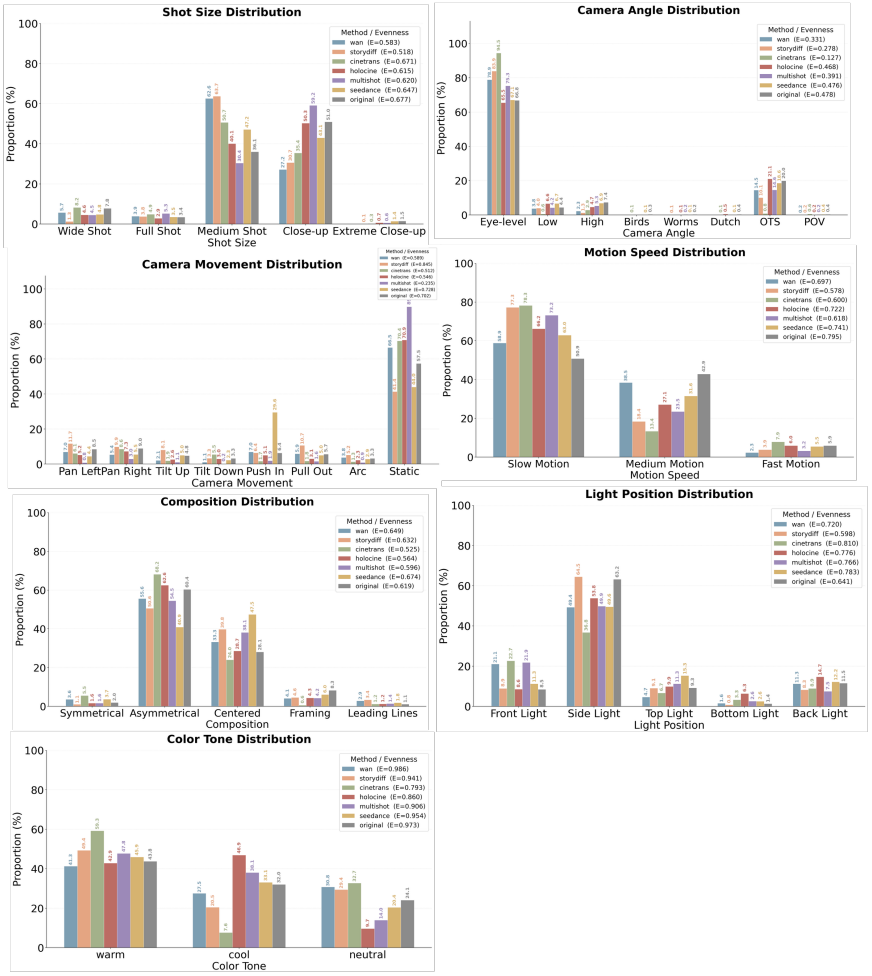


Figure 4: Cinematic Language Distribution. We observe that Seedance 2.0 achieves a balanced use of cinematic vocabulary comparable to that of professional videos. More interestingly, existing methods exhibit uneven strengths across individual dimensions: for instance, MultiShotMaster performs competitively on shot size (scale) yet falls short on camera angle and camera movement.

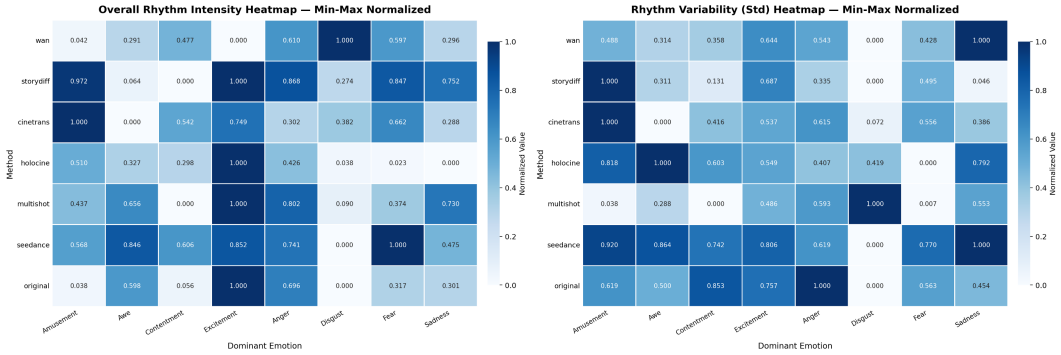


Figure 5: Rhythm Intensity and Variability across Eight Emotions. Intensity captures the change intensity across transitions and Variability presents how rhythm varies along the sequence. The rows of original videos show that high-energy emotions such as excitement and anger tend to produce higher intensity and greater variability, while more subdued emotions such as sadness and disgust yield lower and more consistent values. Seedance 2.0 yields the distribution most closely aligned with the original, despite a still noticeable gap, indicating that it best captures how pacing should be modulated to serve emotional expression among the evaluated models.

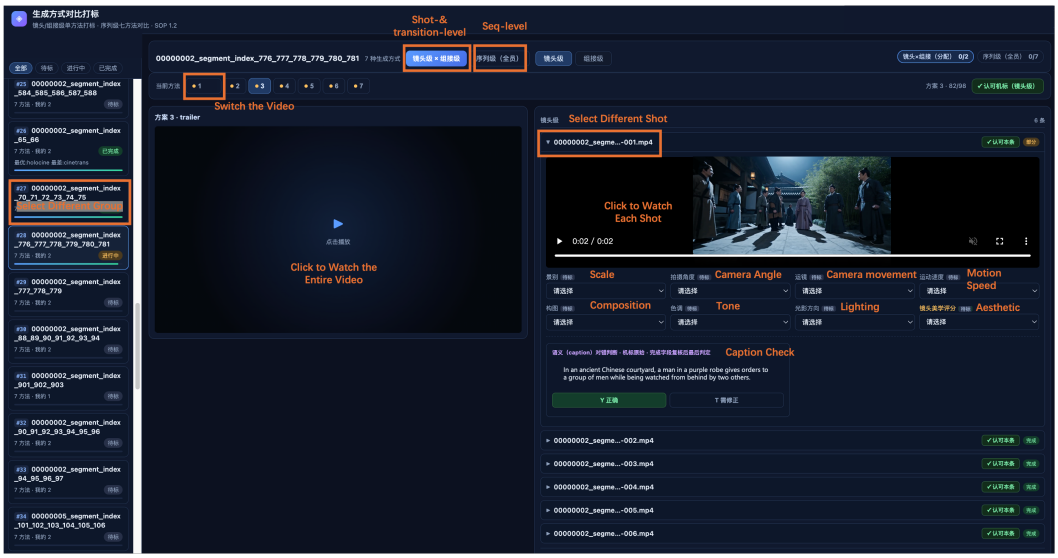


Figure 6: Interface of Human Annotation.

Semantic Prompt

```
# Role
You are a video content analyst, skilled at pulling core narrative information out of a shot.

# Task
Watch the input shot. Extract semantic information along three sub-dimensions —
【subjects】, 【environment】, 【events/actions】 — and write an overall semantic description.

# Instructions
1. Describe only what is **visible in the frame**. Do not invent plot, do not guess at directorial intent.
2. subjects:
  Entities that carry the narrative or hold visual focus — people, animals, objects, vehicles, buildings,
natural scenery, etc.
  - subject_type enum:
    person | group_of_people | animal | vehicle | object | building |
    landscape | food | plant | text_graphic | other
  - description: the subject's notable traits
    · person: number of people, gender, age range, hair/clothing, identity cues
    · animal: species, color, size, posture
    · object/vehicle: type, color, material, state
    · building/landscape: type, style, scale, distinguishing features
  - state_or_action: current state, action, or facial expression
  - With several subjects, order them by visual importance, max 3
  - With no clear subject, return an empty array for subjects
3. environment:
  - location_type: indoor | outdoor | mixed | unclear
  - time_of_day: day | night | dawn | dusk | unclear
  - scene / weather / key_objects: brief description in English
4. events/actions: the key actions, interactions, or state changes inside the shot
5. overall_summary: one English sentence (≤30 words) capturing the shot's main content

# Output JSON Schema
{
  "subjects": ["string (English)"],
  "environment": ["string (English)"],
  "events": ["string (English)"],
  "overall_summary": "string"
```

Camera Composition Prompt

```
# Role
You are a professional Director of Photography, expert at identifying cinematic shot language.

# Task
Analyze the input shot and judge the following four dimensions:
【shot_size】, 【camera_angle】, 【motion_speed】, 【composition】.

# Definitions & Visual Cues

【shot_size】
Judge based on the **main subject's** proportion within the frame
(the subject may be a person, animal, object, building, landscape, etc.)
and how much of it is cropped/included.
- Wide Shot: subject fully visible but occupies little of the frame; environment dominates
- Full Shot: subject fully in frame, nearly filling the vertical space
- Medium Shot: about half of the subject visible (for a person: waist up)
- Close-up: a key part of the subject fills most of the frame (for a person: face/shoulders up)
- Extreme Close-up: a small region of the subject fills the frame (e.g. an eye, a detail)

Note: If the subject is not a person, apply the same "proportion logic"
(e.g., a whole car in frame = Full Shot; only a headlight = Close-up / Extreme Close-up).

【camera_angle】
- Eye Level: lens aligned with the subject's eye line / centerline
- Low Angle: lens below the subject, tilted up
- High Angle: lens above the subject, tilted down
- Birds Eye: straight-down overhead view
- Worms Eye: extreme low angle near the ground, looking up
- Dutch Angle: frame visibly tilted
- OTS: shot over someone's shoulder toward another subject
- POV: first-person view from a character

【motion_speed】 (combining camera movement + subject movement)
- Slow Motion: slow, gentle pace, or slow-motion effect
- Medium Motion: normal, natural rhythm
- Fast Motion: fast, intense, sense of speed, or sped-up effect

【composition】
- Centered Composition: subject placed at frame center
- Symmetrical Composition: clear left-right or top-bottom symmetry
- Asymmetrical Composition: not symmetric but visually balanced
- Framing Composition: doors/windows/branches used as an inner frame
- Leading Lines Composition: clear leading lines (roads, rails, light) pointing at the subject


# Output JSON Schema
{
  "main_subject": "string",
  "shot_size":    {"label": "string", "reason": "string"},
  "camera_angle": {"label": "string", "reason": "string"},
  "motion_speed": {"label": "string", "reason": "string"},
  "composition":  {"label": "string", "reason": "string"}
}
```

Lighting and Color Prompt

```
# Role
You are a cinematographer skilled at reading light.

# Task
From 2–3 representative frames, judge 【color_tone】 and 【light_position】.

# color_tone
- Warm: red/orange/yellow/warm brown dominates
- Cool: blue/cyan/teal/cool green/purple dominates
- Neutral: balanced, grayscale, neutral daylight, or B&W

# light_position — Steps

Step 1. Pick the subject to read
- People present → frontmost person
- No people → most prominent object/building/foreground
- Abstract frame (pure color, flares, smoke, heavy blur) → uncertain

Step 2. Observe first
- Which side of the subject is in shadow?
- Which part of the subject is brightest?
- Background brighter / darker / similar to subject?

Step 3. Label by evidence (stop at first match)
- Back Light: subject front dark + rim light, or background clearly brighter
- Side Light: clear left/right contrast with a sharp light–dark edge
- Top Light: top brightest, shadows point down (nose/chin, top highlights, short ground shadow)
- Bottom Light: bottom brightest, shadows point up
- Front Light: front evenly lit, no directional hard shadow, flat look
- uncertain: diffuse/overcast/balanced sources/over- or under-exposed/ambiguous

# Output JSON
{
  "color_tone": {
    "label": "Warm | Cool | Neutral",
    "reason": " string"
  },
  "light_position": {
    "label": "Front Light | Side Light | Back Light | Top Light | Bottom Light | uncertain",
    "reason": " string"
  }
}
```

Shot Aesthetic Prompt

```
# Role
You are a film juror evaluating a single shot's image quality and cinematic language.
Distinguish "live-action" from "AI-generated" footage — their typical flaws differ, but standards
apply equally strictly.

# ========== 4-Tier Aesthetic Baseline ==========
1 · Poor      Obvious flaws, uncomfortable to watch. Over/underexposure, out-of-focus, noise,
color distortion, unbalanced composition, visible mistakes.
2 · Fair      UGC feel. No hard flaws but casual snapshot; no design in composition/light/color.
3 · Good      TV-variety / mid-tier drama feel. Basics met, some designed dimensions, but lacks
cinematic quality; lighting doesn't serve narrative.
4 · Excellent Film-industry level. Light and shape serve narrative, composition and color clearly
designed, unified whole, tells a story.

# ========== Hard-Flaw Checklist (Step 0) ==========
AI-type:
  A Structural distortion: face/features warped, wrong fingers/limbs, objects defying common
sense (garbled text, floating chair legs)
  B Temporal drift: inter-frame flicker, texture boiling, subject identity/clothing drift, background
elements appearing/disappearing
  C Physics violation: shadows mismatching light source, wrong reflections/refractions, motion
defying gravity
Traditional:
  D Technical defects: out-of-focus, severe over/underexposure, heavy noise, drastic
unintentional shake, visible bloopers, severe color distortion

# ========== 5 Sub-Dimension Scores (integer 1–4) ==========
  visual_quality        Sharpness, exposure, color, Step-0 flaws
  composition_quality   Subject, balance, leading, negative space
  lighting_quality      Light shaping, layering, atmosphere
  camera_language       Movement intent, shot size and angle
  aesthetic_impression  Tonal unity, mood, cinematic feel

Tier rules: 4 = nameable technique + concrete evidence / 3 = pleasant but no highlights / 2 =
mediocre or 1 AI-flaw hit / 1 = hard flaw


# Output JSON
{
  "scores": {
    "visual_quality":      {"score": number, "comment": "string, <=10 words"},
    "composition_quality":  {"score": number, "comment": "string, <=10 words"},
    "lighting_quality":     {"score": number, "comment": "string, <=10 words"},
    "camera_language":      {"score": number, "comment": "string, <=10 words"},
    "aesthetic_impression": {"score": number, "comment": "
```

Transition Score Prompt

```
# Role
You are a senior editor + AIGC review expert. Score the cut coherence of a "Shot A → Shot B" video segment.

# Analysis Principles
- First judge `is_aigc` (evidence: artifacts, identity drift, non-physical deformation, over-smooth texture, abnormal lighting).

# Scoring Dimensions (1–4, 4 = best)

## Universal dimensions (always scored)

jump_cut_avoidance  [hard flaw: score ≤ 2]
  Jump cut = content before/after the cut is similar but shot size / angle changes too little — feels like "a jitter."
  4: change sufficient, no jump cut  3: change slightly small but natural  2: obvious jump cut, mild jitter  1: severe jump cut / axis crossing

shot_size_transition
  7 shot-size levels (far → near): WS / FS / MFS / MS / MCU / CU / ECU. Adjacent = 1 step apart.
  4: 0–1 step  3: 2 steps, reasonable  2: 3–4 steps, abrupt  1: 5+ steps, extreme

camera_direction_consistency  [hard flaw: score = 1]
  Not required to be same direction — just not opposite.
  4: same direction / perpendicular / one static  3: different direction but natural transition  2: clearly opposite  1: fully opposite / crosses axis

motion_intensity_consistency
  Combines camera movement + subject motion, compared across the cut.
  4: consistent or reasonable progression  3: mild-to-medium difference, natural  2: large difference, abrupt  1: extreme contrast (still to violent)

color_tone_consistency  (warm/cool, saturation, dominant color)
  4: highly consistent  3: slight difference, unified tone  2: warm/cool flipped or saturation jump  1: fully broken

## AIGC dimensions (score only if is_aigc = true, otherwise score = null)

visual_fidelity  [hard flaw: score ≤ 2]
  Per-frame structure / texture / AIGC artifacts (deformation, mushy faces, extra fingers, etc.).
  4: reasonable structure, no artifacts  3: occasional minor, non-disruptive  2: obvious artifacts or structural errors  1: severe breakdown

temporal_stability  [hard flaw: score ≤ 2]
  Internal stability within each shot around the cut (flicker, shape jumps). Use null if unjudgeable.
  4: fully stable  3: extremely minor jitter  2: obvious flicker/jump  1: severely unstable

# overall Scoring Rule (top-down match, stop at first hit)
Definitions: hard = number of dimensions that triggered a hard flaw; probs = number of scored dimensions with score ≤ 2.

1 broken    : hard ≥ 2  OR  probs ≥ 3
2 poor      : hard = 1  OR  (hard = 0 AND probs = 2)
3 good      : hard = 0  AND  probs ≤ 1
4 excellent : hard = 0  AND  every scored dimension ≥ 3

# Output JSON (compact, no extra spaces)
# Format: each dim item is [score, evidence]
# When not AIGC, fill vf / ts as [null, "not AIGC"]
# overall must strictly follow the rule — no subjective boosting

{
  "aigc": true|false,
  "overall_coherence_score": 1-4,
  "summary": "  string"
}
```

Emotion Category Prompt

Role
You are an expert annotator in affective computing and film analysis.
Your task is to identify the DOMINANT emotion evoked in a viewer by a short, multi-shot video, using Mikels' 8 discrete emotion categories.

The 8 categories (with definitions):
- amusement: light-hearted enjoyment, fun, humor, playfulness; often triggered by funny or cute content.
- awe: wonder, admiration, being moved by something vast, beautiful, magnificent, or sublime (e.g., landscapes, achievements).
- contentment: calm satisfaction, peace, warmth, gentle happiness, coziness, intimacy.
- excitement: high-energy positive arousal, thrill, anticipation, enthusiasm, action.
- anger: irritation, rage, hostility, aggression, frustration directed at someone/something.
- disgust: revulsion, repulsion, aversion to something offensive, gross, or morally repugnant.
- fear: threat, danger, dread, anxiety, suspense, horror.
- sadness: sorrow, grief, loneliness, melancholy, loss, hopelessness.

Please:
1. Briefly summarize what happens across the shots (1–2 sentences).
2. List the key emotional cues you observed (visual, motion, cinematographic language, character behavior, color/lighting, audio if any).
3. Provide a probability distribution over all 8 categories (values sum to 1.0).

Return ONLY the following JSON (no markdown, no comments):

```
{
"evidence": ["<cue 1>", "<cue 2>", "<cue 3>"],
"scores": {
"amusement": 0.00,
"awe": 0.00,
"contentment": 0.00,
"excitement": 0.00,
"anger": 0.00,
"disgust": 0.00,
"fear": 0.00,
"sadness": 0.00
},
"dominant_emotion": "<one of the 8 labels>",
"confidence": 0.00
}
```

Table 7: Shot-level vocabulary: dimensions, categories, and definitions.

Dimension	Category	Definition
Semantics	Subjects / Environment / Events	Agreement between the machine-generated caption and the visual content, covering subjects, environment, and events/actions (accepted when semantic coverage exceeds approximately 80%).
Shot Scale	Wide Shot	The environment dominates the frame; the subject appears small.
	Full Shot	The subject's entire body is visible together with its spatial context.
	Medium Shot	Framing extends approximately from the waist to the head.
	Close-up	The face or a local region of the subject occupies most of the frame.
	Extreme Close-up	A fine-scale detail (e.g., an eye, a hand, or a surface texture) fills the frame.
Camera Angle	Eye Level	The optical axis is aligned with the subject's line of sight.
	Low Angle	The camera is positioned below the subject and tilted upward.
	High Angle	The camera is positioned above the subject and tilted downward.
	Bird's Eye	A near-vertical top-down viewpoint.
	Worm's Eye	A near-vertical bottom-up viewpoint.
	Dutch Angle	The horizon or vertical lines are perceptibly tilted.
	Over-the-Shoulder	A foreground shoulder or head frames a second subject in the background.
	Point-of-View	The frame simulates the subjective viewpoint of a character.
Camera Movement	Pan Left	Horizontal panning or lateral translation to the left.
	Pan Right	Horizontal panning or lateral translation to the right.
	Tilt Up	Vertical tilting or crane movement upward.
	Tilt Down	Vertical tilting or crane movement downward.
	Push In	Dolly-in or zoom-in toward the subject.
	Pull Out	Dolly-out or zoom-out away from the subject.
	Arc Shot	Rotational motion about an axis, either orbiting the subject or rotating on the camera's own axis.
	Static Shot	Camera position and framing remain fixed.
Motion Speed	Slow	Gentle, unhurried movement, typically associated with lyrical or observational framing.
	Medium	Clear but non-intense motion, comparable to a walking pace.
	Fast	Rapid, high-energy motion at a running pace or above, conveying tension or impact.
Composition	Symmetrical	Formal, stable structure exhibiting horizontal or vertical mirroring.
	Asymmetrical	Non-mirrored arrangement in which overall visual weight remains balanced.
	Centered	The subject is placed at the geometric center of the frame for emphasis.
	Framing	Environmental elements (e.g., doors, windows) form a frame-within-a-frame.
	Leading Lines	Converging lines, curves, diagonals, or S-/L-shaped structures direct attention toward the subject.
Color Tone	Warm	Overall chromatic bias toward red, orange, or yellow.
	Cool	Overall chromatic bias toward blue, cyan, or green.
	Neutral	No pronounced warm or cool bias; colors appear natural.
Lighting Direction	Front Light	Illumination originates from the front of the subject.
	Side Light	Illumination originates from the side, accentuating facial or surface texture and enhancing dramatic effect.
	Top Light	Illumination originates from directly above, producing downward cast shadows that reinforce contour and contrast.
	Bottom Light	Illumination originates from below the subject.
	Back Light	The light source is behind the subject, yielding a rim-light or silhouette effect.

Table 8: Quantization scheme for the seven shot-language dimensions. For each dimension we list the categorical labels, their numerical encoding, the pairwise distance rule, and the maximum distance $\Delta_{\max}$ used for normalization ($\Delta/\Delta_{\max} \in [0, 1]$).

Dimension	**Categories → Numerical label**	**Distance rule** $\Delta(a, b)$	$\boldsymbol{\Delta_{\max}}$
Shot Scale	Extreme Close-up → 1 Close-up → 2 Medium Shot → 3 Full Shot → 4 Wide Shot → 5	Absolute ordinal difference: $\Delta = \lvert v_a - v_b \rvert$	4
Camera Angle	*Pitch axis:* Worm's Eye → -2 Low → -1 Eye Level → 0 High → $+1$ Bird's Eye → $+2$ *Special (non-ordinal):* Dutch, OTS, POV	If both a, b are ordinal: $\Delta = \lvert v_a - v_b \rvert$ Otherwise (any side special): $\Delta = \mathbb{1}[\text{label}_a \neq \text{label}_b]$ Already normalized to $[0, 1]$.	4
Camera Movement	*Type:* Static; Pan Tilt ; Push; Arc *Direction:* Left / Up / In → -1 Right / Down / Out → $+1$ Static, Arc → 0	See the distance between different camera movement in Table 9.	3
Motion Speed	Static → 0 Slow → 1 Medium → 2 Fast → 3	Absolute ordinal difference: $\Delta = \lvert v_a - v_b \rvert$	3
Composition	Centered, Symmetrical, Asymmetrical, Framing, Leading Lines (nominal)	Switch indicator (no ordering): $\Delta = \mathbb{1}[c_a \neq c_b]$	1
Color Tone	Cool → -1 Neutral → 0 Warm → $+1$ Mixed warm–cool → $+2$	Absolute ordinal difference on the tension axis: $\Delta = \lvert v_a - v_b \rvert$	3
Light Position	Front → $(0, 0)$ Side → $(1, 0)$ Back → $(2, 0)$ Top → $(1, +1)$ Bottom → $(1, -1)$	Chebyshev (L_∞) distance: $\Delta = \max(\lvert x_a - x_b \rvert, \lvert y_a - y_b \rvert)$	2

Table 9: Distance Matrix between Different Camera Movement Types.

	Sta	PanL	PanR	TiltU	TiltD	PushI	PullO	Arc
Static	0	1	1	1	1	1	1	3
Pan L	1	0	2	1	1	1	1	2
Pan R	1	2	0	1	1	1	1	2
Tilt U	1	1	1	0	2	1	1	2
Tilt D	1	1	1	2	0	1	1	2
Push I	1	1	1	1	1	0	2	2
Pull O	1	1	1	1	1	2	0	2
Arc	3	2	2	2	2	2	2	0

Table 10: Scoring Criteria for Human Study

Perspective	Dimension	Definition	Scoring Criteria (1–4)
Director	Narrative Coherence	Assesses whether the overall sequence flows smoothly in both semantic and visual continuity, without violating continuity editing rules.	**4**=Narrative flows smoothly and visual editing enhances storytelling (e.g., match cuts, motion continuity); **3**=Narrative is smooth; visual editing is error-free but adds no value; **2**=Narrative jumps or visual errors weaken semantics; **1**=Both narrative and visuals fail, close to random assembly.
	Cinematic Diversity	Evaluates the richness of shot scale, angle, composition, camera movement, lighting, and color tone.	**4**=Multi-dimensional variation is natural, layered, and clearly designed; **3**=Noticeable variation but uniform or random, lacking hierarchy; **2**=Minor variation, overall repetitive and monotonous; **1**=Almost no variation, sustained single state throughout.
	Visual Rhythm	Whether the pacing of visual changes aligns with the emotional expression (shot duration excluded).	**4**=Rhythm matches emotion with clearly designed dynamics; **3**=Rhythm broadly aligns with emotion, no clear mismatch; **2**=Partial alignment, but insufficient variation or off-timed beats; **1**=Rhythm clearly clashes with emotional expression.
Viewer	Narrative Comprehension	Whether the viewer can understand the content, character relationships, and event logic.	**4**=Narrative is crystal clear, all details understood; **3**=Main storyline and most details understood; **2**=Main storyline grasped but some details unclear; **1**=Completely incomprehensible.
	Cognitive Engagement	Whether the video effectively captures and sustains viewer attention, evoking curiosity, suspense, and the desire to keep watching.	**4**=Attention is fully engaged throughout, actively prompting "what happens next"; ending leaves something unresolved; **3**=Watchable without losing attention, but passive; no lingering suspense; **2**=Strong opening but stagnant middle/end, attention drifts; **1**=Neither event nor sustained emotion, merely playing images.
	Emotional Resonance	Whether the viewer experiences narrative-related emotions during viewing.	**4**=Strong emotional experience with resonance or lasting aftertaste; **3**=Clear and sustained emotional experience; **2**=Mild or fleeting emotional fluctuation; **1**=No emotional response at all.