The Script is All You Need: An Agentic Framework for Long-Horizon Dialogue-to-Cinematic Video Generation

Anonymous ACL submission

Abstract

Recent advances in video generation have enabled high-fidelity short clips from text prompts, but generating long-horizon, dialogue-driven cinematic sequences remains challenging. A key bottleneck is a "semantic gap" between sparse conversational intent and the fine-grained, executable cinematic plan required for shot design, camera control, and continuity. We propose an end-to-end, script-centric agentic framework for dialogue-to-cinematic video generation. Our framework first uses `ScripterAgent` to translate coarse dialogue into a structured, shot-level cinematic script. To support this step, we construct `ScriptBench`, a benchmark of 1,750 instances annotated via an expert-guided pipeline with multi-round verification. `ScripterAgent` is trained via SFT for structural competence, followed by RL for cinematic alignment. The generated scripts are then executed by `DirectorAgent`, which orchestrates state-of-the-art video models using shot-aware segmentation and a frame-anchored cross-scene continuous generation strategy to improve long-horizon coherence. Comprehensive evaluation with both `CriticAgent` and human experts shows that conditioning video models on our scripts consistently increases human-rated script faithfulness alongside character consistency and narrative coherence. Our results also reveal a practical trade-off in current video generators between visual spectacle and strict script adherence.

1 Introduction

"To make a great film you need three things:
***the script**, **the script** and **the script**."*
— Alfred Hitchcock

The emergence of powerful video generation models like Sora2-Pro (Brooks et al., 2024), Veo-3.1, and Wan2.6 (Wan et al., 2025) has marked a new era in artificial intelligence, demonstrating a remarkable ability to synthesize high-fidelity video clips from simple text prompts. However, a critical yet underexplored question remains: **Can these models generate long-form, narratively coherent cinematic videos from high-level creative concepts?** Dialogue is inherently underspecified: it conveys intent and emotion but rarely specifies what must be shown, how the camera should move, or how shots should be paced. This mismatch exposes a "semantic gap" between a high-level creative concept (dialogue) and the low-level, executable plan required to realize it as a coherent cinematic video.

This paper inverts the conventional video-language relationship from passive description (video-to-text) to active planning and execution. We tackle a new, challenging task: given only coarse-grained dialogue, the model must *anticipate* and *generate* an executable filmmaking plan. This introduces three fundamental challenges: **(1) fine-grained contextual understanding** to resolve ambiguities in sparse dialogue; **(2) domain knowledge of filmmaking** to produce valid camera specifications; and **(3) creative reasoning** to bridge what is said with what must be shown. To address these challenges, we introduce a complete, agentic framework for dialogue-to-cinematic-video generation, composed of three core components: `ScripterAgent`, `DirectorAgent`, and `CriticAgent` (Figure 1). Here, we use the term agent to refer to an LLM-driven module that makes explicit intermediate decisions (e.g., shot-aware segmentation, conditioning selection, and iterative verification) and invokes external tools/models to complete each stage.

To facilitate our research, we first construct `ScriptBench`, a large-scale benchmark with 1,750 instances for this task. Each instance is annotated using a novel, expert-guided pipeline that ensures cinematic plausibility through multi-round error correction. Building on this, we develop `ScripterAgent`, a model trained to transform di-

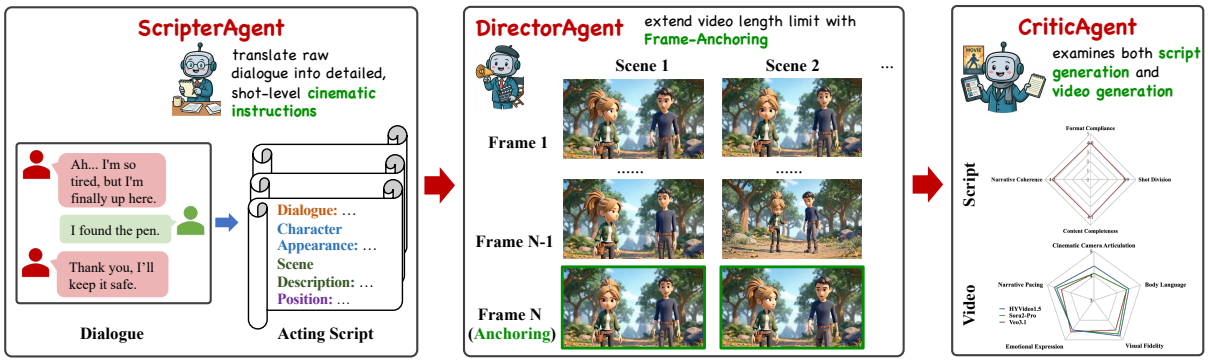

Figure 1: Our pipeline consists of three key components: (1) a `ScripterAgent`, trained to align its outputs with professional directorial standards; (2) a `DirectorAgent`, which ensures seamless visual continuity across scenes, thereby overcoming the temporal incoherence caused by the fixed-duration constraints of video generation models; and (3) a `CriticAgent`, which evaluates the generated film from both technical and cinematic perspectives.

alogue into a structured cinematic script. We employ a two-stage training paradigm: supervised fine-tuning (SFT) to learn the script's structure, followed by Group Relative Policy Optimization (GRPO) (Shao et al., 2024) with a hybrid reward function that combines rule-based structural correctness with learned expert aesthetics. The resulting script is then passed to `DirectorAgent`, which orchestrates state-of-the-art video models using a novel Cross-Scene Continuous Generation strategy with frame-anchoring to produce long-horizon, coherent videos that overcome the temporal limitations of current generators.

Our comprehensive experiments demonstrate the effectiveness of this script-centric approach. The full `ScripterAgent` model significantly outperforms existing methods, with human experts rating its outputs higher in both *Dramatic Tension* (4.1 vs. 3.7) and *Visual Imagery* (4.3 vs. 3.8). Furthermore, using our generated scripts as input universally improves the performance of all tested video models, boosting metrics like *Script Faithfulness* by up to +0.8 points. Our analysis also uncovers a critical trade-off in these models between visual spectacle and script faithfulness: models like `Sora2-Pro` excel in visual appeal, while others like `HYVideo1.5` prioritize narrative integrity. This work provides an end-to-end solution for dialogue-driven cinematic video generation and offers a new paradigm for automated storytelling.

Our contributions are summarized as follows:

1. We formulate long-horizon cinematic video generation as a planning-and-execution problem and propose a script-centric agentic framework.

2. We introduce `ScriptBench`, a benchmark of 1,750 instances annotated via an expert-guided, multi-stage pipeline with adaptive error correction, and we develop `ScripterAgent` trained by a two-stage SFT + RL paradigm to improve both structural validity and cinematic quality.

3. We propose a Cross-Scene Continuous Generation strategy with frame anchoring to improve long-horizon coherence when using fixed-duration video generators.

2 `ScripterAgent`: Dataset and Model

Prevailing video-language research has largely focused on the paradigm of passive description, where models learn to generate textual descriptions for existing video content (e.g., captioning or question answering). In contrast, our task inverts this relationship: given only coarse-grained dialogue, the model must *anticipate* and *generate* an executable filmmaking plan. To bridge this gap and foster research in automated cinematic planning, we introduce `ScriptBench`, a new benchmark designed specifically for this task, along with `ScripterAgent`, a dedicated model trained to transform dialogue into professional-quality cinematic scripts.

2.1 `ScriptBench`

To facilitate our study, we constructed a large-scale, high-quality dataset of cinematic scripts. We curated raw instances from high-fidelity cinematic cutscenes. These sources were selected for their rich dialogue, professional cinematography, and high visual consistency, which closely approximate real-world film production.

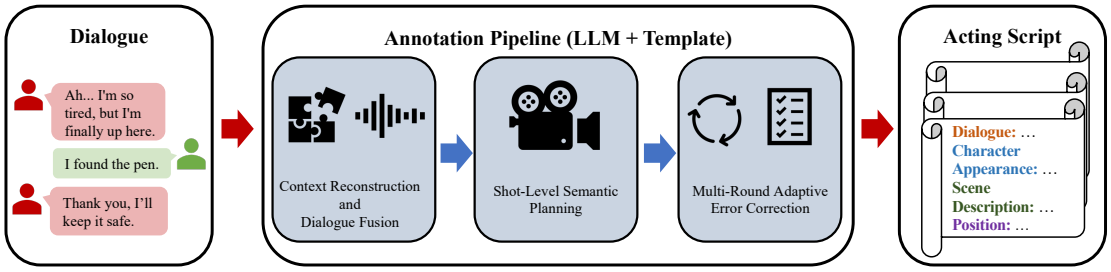

Figure 2: The three-stage, expert-guided pipeline for creating the `ScriptBench`.

A central contribution of our work is a scalable yet high-fidelity annotation pipeline that expands sparse dialogue into rich, shot-level cinematic scripts. Although we leverage the SOTA LLM (`gemini-2.5-pro`), the entire process is tightly governed by expert-defined templates, domain constraints, and validation rules to ensure both cinematic plausibility and physical consistency. The pipeline operates in three stages:

Stage 1: Context Reconstruction and Dialogue Fusion. The model first parses the multimodal inputs to reconstruct a comprehensive understanding of the scene. It jointly analyzes the textual script and dialogue audio to infer character relationships, scene settings, plot developments, emotional tendencies, and speaking intent. This process fuses the disparate signals into a coherent narrative context that makes implicit causal relations explicit.

Stage 2: Shot-Level Semantic Planning. Utilizing the reconstructed context, the model plans shots under four constraints to ensure visual and narrative continuity. *Shot integrity* enforces self-contained units, introducing cuts only upon clear camera or scene changes. *Duration adaptation* caps shots at 10 seconds to align with generation limits. *Semantic coherence* aligns boundaries with narrative transitions (e.g., emotional shifts), while *Technical feasibility* prevents segmentation during complex camera motions. These principles jointly ensure the shot units are narratively meaningful and technically viable for downstream generation.

Stage 3: Multi-Round Adaptive Error Correction. In the final stage, the system executes a multi-round, adaptive error-correction loop (Figure 2, right) to ensure both structural validity and semantic fidelity of the generated scripts. We design four verification modules: (a) *Dialogue Completeness*, which ensures that all spoken content is either explicitly transcribed or marked as `[No Dialogue]`; (b) *Character Appearance Consistency*, which enforces strict adherence to predefined character descriptions; (c) *Scene Coherence*, which tracks environmental elements and validates narratively justified transitions; and (d) *Positional and Physical Rationality*, which verifies spatial relations against plausible blocking and camera geometry. An automated detector iteratively scans the scripts, feeds corrective signals back to the generator, and repeats the loop until all constraints are satisfied. To further validate practical reliability, professional script consultants conducted a random audit on 60% of the generated instances, revealing that while the automated pass rate reached 94%, expert review exposed subtle semantic errors such as character teleportation, dialogue–action conflicts, and inconsistent prop states. These findings were incorporated into refined prompt constraints and verification logic, resulting in a controlled refinement process that constitutes a key novelty of our pipeline and yields cinematic scripts that are structured, internally consistent, and grounded in long-horizon narrative and physical continuity.

Dataset Statistics and Usage This pipeline yielded 1,750 finalized script instances, each in one-to-one correspondence with a raw multimodal input. The average duration of each video clip is approximately 15.4 seconds, providing sufficient temporal scope for multi-shot sequences while remaining tractable for current generative models. The dataset is partitioned into a training set (1700 instances) and a test set (50 instances). This partitioning scheme intentionally challenges the model to infer complete cinematic elements from conversational content alone, emulating the real-world process where directors visualize a story from a dialogue-driven script.

2.2 ScripterAgent

Building upon ScriptBench, we develop ScripterAgent to automatically transform coarse-grained dialogue into a fine-grained, structured cinematic script. While large-scale foundation models demonstrate strong general capabilities, we hypothesize that this specialized task that requires domain knowledge of shot composition, pacing, and visual continuity, can benefit from targeted training on curated data. To this end, we employ a two-stage training paradigm: supervised fine-tuning (SFT) to learn the script format and narrative structure, followed by reinforcement learning (RL) to align the model's outputs with professional directorial aesthetics.

2.2.1 Stage I: SFT for Structural Competence

The initial stage focuses on teaching the model the fundamental syntax and structure of cinematic scripts. We formulate this as a sequence-to-sequence task, where the input x is a multi-turn dialogue from our dataset, and the output y is the target script in a structured JSON format. We fine-tune *Qwen-Omni-7B* as our base model, π_{base}, chosen for its strong capabilities in long-context processing and instruction following. The training objective is to maximize the conditional log-likelihood of the ground-truth script:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{t=1}^{|y|}\log \pi_\theta(y_t|y_{<t}, x)\right]$$

where θ denotes the model parameters. We train for 20 epochs using the AdamW optimizer with a learning rate of $\eta = 1\times10^{-5}$, a batch size of 4, and a maximum sequence length of 8,192 tokens. This SFT stage equips the model to generate scripts that are structurally correct and content-complete, forming a solid foundation for the subsequent creative refinement.

2.2.2 Stage II: RL for Cinematic Alignment

While SFT ensures structural validity, it is insufficient for capturing the subjective artistry of professional filmmaking. As suggested in our evaluation (Table 1), effective scriptwriting transcends logical correctness, involving aesthetic judgments about shot composition, pacing, and emotional impact. To bridge this gap, we introduce a reinforcement learning stage to align ScripterAgent with expert directorial preferences. We employ *Group Relative Policy Optimization (GRPO)* (Shao et al., 2024), an advanced preference alignment method whose group-based relative scoring is well-suited for creative tasks that have a subjective, one-to-many nature of valid outputs.

Hybrid Reward Function. A key novelty of our RL stage is a hybrid reward function, R_{total}, that balances objective correctness with subjective quality. It is a weighted sum of two complementary signals, with $\alpha = 0.4$:

$$R_{\text{total}}(y) = \alpha \cdot R_{\text{structure}}(y) + (1-\alpha)\cdot R_{\text{human}}(y)$$

- *Rule-Based Structural Reward ($R_{structure}$):* This component provides an objective signal for technical correctness, mirroring the verification modules from our data annotation pipeline. It evaluates and aggregates normalized scores from four automated checks: *Format Compliance* (correct JSON structure), *Dialogue Completeness* (all spoken lines accounted for), *Scene and Character Consistency*, and *Physical Rationality* (plausible character positions and camera geometry).
- *Human Preference Reward (R_{human}):* To capture cinematic aesthetics, we model expert human judgment. A team of three senior art directors scored SFT model outputs on a 1–5 scale across four creative dimensions: shot division rationality, character acting and emotion, visual aesthetics, and directorial intent. Using 500 such annotated samples ($\mathcal{D}_{\text{pref}}$), we trained a BERT-based regression model to predict a normalized preference score in [0, 1], serving as a scalable proxy for expert cinematic taste.

GRPO Optimization. During optimization, for each input x, we generate $K = 8$ candidate scripts $\{y^{(k)}\}_{k=1}^{K}$ from the current policy $\pi_\theta(\cdot|x)$ and calculate their rewards $R_k = R_{\text{total}}(y^{(k)})$. These rewards are then used to compute a normalized advantage within the group:

$$A_k = \frac{R_k - \bar{R}}{\sigma_R + \epsilon}, \quad \text{where} \quad \bar{R} = \frac{1}{K}\sum_{k=1}^{K} R_k$$

The policy is updated by maximizing the advantage-weighted log-likelihood, constrained by a KL-divergence penalty to prevent large deviations from the SFT initialization:

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{K}\sum_{k=1}^{K} A_k \cdot \log \pi_\theta(y^{(k)}|x)\right] - \beta\cdot\mathbb{E}_{x\sim\mathcal{D}}\left[\text{KL}\left(\pi_\theta(\cdot|x)\,\|\,\pi_{\text{SFT}}(\cdot|x)\right)\right]$$

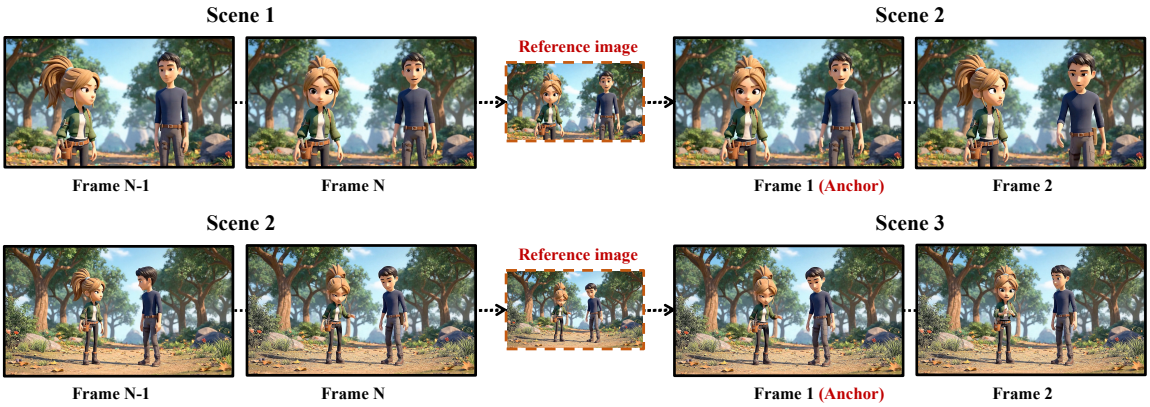

Figure 3: *Illustration of the Frame-Anchoring Mechanism.* The `DirectorAgent` utilizes the last frame of the preceding scene (N) to visually condition the generation of the current scene's first frame (1). This frame-anchoring strategy preserves visual consistency in character appearance, attire, and scene layout across multiple generation cycles, creating a seamless visual relay.

where the KL coefficient $\beta = 0.04$. The model is trained for 5,000 steps using the Adam optimizer with a learning rate of $\eta = 10^{-6}$ and a batch size of 4. This two-stage paradigm successfully elevates `ScripterAgent`'s capabilities from generating structurally correct scripts to producing cinematically compelling plans aligned with professional standards.

3 `DirectorAgent`: Long-Horizon Script-to-Video Execution

Terminology. We use *shot* to denote the atomic cinematographic unit with a single continuous camera take in the script. Due to the fixed temporal limit of current video generators, we further partition the full script into *segments* (also referred to as *scenes* in the execution module), where each segment corresponds to a single model call and may contain one or multiple shots. We use *clip* to refer to the final output video obtained by concatenating all segments in temporal order.

While `ScripterAgent` provides a structured, shot-by-shot blueprint, translating this plan into a continuous video presents its own formidable challenge. The `DirectorAgent` is designed to bridge this execution gap, acting as an automated orchestrator that transforms the generated script into a coherent, high-fidelity video sequence. Its primary function is to overcome a fundamental limitation of current video generation models: restricted temporal capacity. State-of-the-art models are typically limited to generating short clips (e.g., 8–12 seconds), far short of the 1–3 minute duration of a complete narrative scene. Naively segmenting a script and generating clips independently leads to severe artifacts, such as identity drift, inconsistent styling, and a loss of narrative continuity.

The core novelty of the `DirectorAgent` is a *Cross-Scene Continuous Generation Strategy*, which ensures both semantic coherence and visual consistency across multiple generated segments. This strategy combines (1) intelligent, shot-aware segmentation that respects cinematographic boundaries with (2) a frame-anchoring mechanism that conditions each new segment on the final state of the previous one.

Intelligent Shot-Based Segmentation. Instead of making arbitrary temporal cuts, the agent partitions the full script into a sequence of "scenes" that align with the natural cinematographic boundaries defined by `ScripterAgent`. This process adheres to four key principles:

1. *Shot Integrity:* Each scene must contain one or more complete shot units, preventing cuts in the middle of a continuous camera take.

2. *Duration Adaptation:* The duration of a scene is constrained to fit within the target model's generation window, with a 10% safety buffer.

3. *Semantic Coherence:* Divisions are prioritized at natural narrative breakpoints, such as the end of a character's line or a shift in emotional tone.

4. *Technical Feasibility:* Segmentation is favored at fixed camera positions, avoiding cuts during complex camera movements which are harder to transition between seamlessly.

Method	AI Rating (0-5)				Human Rating (0-5)		
	Format Comp.	*Shot Division*	*Content Comp.*	*Narrative Coher.*	*Character Consist.*	*Dramatic Tension*	*Visual Imagery*
CHAE (Wang et al., 2022)	3.3	3.2	3.4	3.5	3.1	3.3	3.4
MoPS (Ma et al., 2024a)	3.2	3.1	3.3	3.4	3.0	3.2	3.3
SEED-Story (Yang et al., 2025)	3.6	3.5	3.7	3.8	3.6	3.7	3.8
`ScripterAgent` (SFT only)	3.9	3.6	3.8	3.9	3.7	3.6	3.8
`ScripterAgent` (SFT+RL)	**4.0**	**3.9**	**4.1**	**4.2**	**4.0**	**4.1**	**4.3**

Table 1: Script generation evaluation on the `ScriptBench` test set.

Frame-Anchored Continuity. To ensure seamless visual transitions between scenes, the `DirectorAgent` employs a *First-Last Frame Connection Mechanism*. As illustrated in Figure 3, the final frame of a generated scene i is extracted and used as a visual anchor or conditioning image for the generation of the subsequent scene $i + 1$. This technique provides a strong visual prior for the video model, explicitly instructing it to maintain consistency in character identity, clothing, facial details, and spatial layout. By forming a visual relay from one segment to the next, this mechanism substantially reduces the identity drift and jarring scene changes that plague naive segmentation approaches. To further enhance transition quality, we also inject explicit text like "Continuing from the previous scene" into subsequent prompts.

Effectiveness and Limitations. This strategy effectively extends the coherence window of any underlying video model. By conditioning each new segment on the visual state of its predecessor, it transforms a long-horizon generation problem into a sequence of locally solvable, continuity-preserving subproblems. While this approach significantly reduces identity drift and layout inconsistencies, challenges such as imperfect lip synchronization and residual misalignment of fine-grained actions remain.

4 `CriticAgent`: Cinematic Evaluation

To assess our system comprehensively, we introduce a multifaceted evaluation framework that examines both stages of our pipeline: script generation (dialogue-to-script) and video generation (script-to-video). This framework is essential because cinematic quality is inherently multidimensional, encompassing technical correctness, narrative fidelity, and subjective artistic merit. The framework combines objective metrics, automated scoring via our AI-powered `CriticAgent`, and qualitative evaluations by human experts. All scores are assigned on a 0–5 scale. Detailed metric definitions are provided in Appendix B.

Script Generation Evaluation. For the dialogue-to-script stage, we evaluate the generated scripts on both their structural correctness, using our `CriticAgent`, and their artistic quality, via a panel of human experts. We employ our `CriticAgent`, powered by `gemini-2.5-pro`, to automatically assess generated scripts based on four criteria: *Format Compliance*, *Shot Division Rationality*, *Content Completeness*, and *Narrative Coherence*. To complement the automated assessment, a panel of professional directors and screenwriters evaluates the artistic quality of the scripts. They evaluate the script's potential for being filmed successfully from three key creative aspects: *Character Portrayal Consistency*, *Dramatic Tension & Rhythm*, and *Visual Imagery & Cinematic Expressiveness*.

Video Generation Evaluation. We evaluate the script-to-video generation on two primary axes: script-video alignment and overall video quality. `CriticAgent` evaluates the cinematic quality and faithfulness of the generated video to the source script and reference audio across five dimensions: *Cinematic Camera Articulation*, *Kinetic Body Language & Blocking*, *Visual Descriptive Fidelity*, *Emotional Arc & Micro-Expressions*, and *Narrative Pacing & Timing*. Human annotators also assess the final generated videos, providing ratings on five dimensions that collectively offer a comprehensive view of video quality: *Visual Appeal*, *Script Faithfulness*, *Narrative Coherence*, *Character Consistency*, and *Physical Law Adherence*.

Model	AI Rating (0-5)					Human Rating (0-5)					Overall Mean	
	Cam. Artic.	*Body Block.*	*Visual Fidelity*	*Emotion Arc*	*Pace Timing*	*Visual Appeal*	*Script Faith*	*Character Consist.*	*Physical Law*	*Narrative Coher.*	*Avg. AI*	*Avg. Human*
Raw Dialogue (w/o `ScripterAgent`)												
`Vidu2`	4.1	4.1	4.5	4.4	4.4	3.7	4.1	3.0	3.3	3.1	4.3	3.4
`Seedance1.5-Pro`	4.0	4.0	4.5	4.2	4.3	3.5	3.7	3.2	3.1	3.5	4.2	3.4
`Kling2.6`	4.1	4.1	4.6	4.4	4.4	3.6	3.5	3.3	3.4	3.7	4.3	3.5
`Wan2.6`	4.2	4.2	4.7	4.4	4.4	3.5	3.2	3.1	3.7	3.4	4.4	3.4
`HYVideo1.5`	4.0	4.0	4.5	4.3	4.3	4.0	4.2	4.1	3.8	4.1	4.2	4.0
`Sora2-Pro`	4.1	4.0	4.6	4.3	4.3	4.2	3.6	3.7	4.1	3.9	4.3	3.9
`Veo3.1`	4.0	3.9	4.4	4.4	4.3	3.9	4.0	4.1	3.9	4.0	4.2	4.0
Average Score	4.1	4.0	4.5	4.3	4.3	3.8	3.8	3.5	3.6	3.7	4.2	3.7
w/ `ScripterAgent`												
`Vidu2`	4.2	4.4	4.7	4.5	4.5	3.9	4.3	3.7	3.9	3.8	4.5	3.9
`Seedance1.5-Pro`	4.5	4.6	4.7	4.6	4.7	4.0	4.1	4.1	3.9	4.1	4.6	4.0
`Kling2.6`	4.3	4.5	4.6	4.5	4.6	3.9	4.1	4.0	4.2	4.1	4.5	4.1
`Wan2.6`	4.4	**4.6**	4.7	4.6	4.7	4.1	4.0	3.8	4.0	3.9	4.6	4.0
`HYVideo1.5`	**4.4**	4.5	**4.8**	4.5	**4.7**	4.5	**4.6**	**4.4**	4.2	**4.3**	4.6	4.4
`Sora2-Pro`	4.1	4.4	4.7	4.5	4.6	**4.8**	4.2	4.3	**4.5**	4.1	4.5	4.4
`Veo3.1`	4.1	4.4	4.5	**4.6**	4.4	4.6	4.4	4.3	4.4	4.2	4.4	4.4
Average Score	4.3	4.5	4.7	4.5	4.6	4.3	4.2	4.1	4.2	4.1	4.5	4.2

Table 2: Video generation evaluation on the `ScriptBench` test set.

5 Experiment

5.1 Results of Script Generation

We compare against representative story visualization and screenplay generation methods in Table 1.

`ScripterAgent` sets a new state of the art for cinematic script generation. As shown in Table 1, the full `ScripterAgent` model outperforms all baseline methods across seven evaluation metrics spanning both automated AI assessments and human expert ratings. Compared to the strongest prior baseline, `MovieAgent`, our approach achieves consistent gains of +0.4 points or more on key structural and creative dimensions, including *Format Compliance*, *Content Completeness*, and *Dramatic Tension*. Importantly, human evaluators also award a substantially higher score in *Visual Imagery*, indicating that `ScripterAgent` more effectively transforms coarse dialogue into detailed, director-level cinematic instructions.

The RL-based preference alignment stage is critical for elevating artistic quality beyond structural correctness. Ablation against an SFT-only variant shows that supervised fine-tuning alone suffices to learn the formal structure of cinematic scripts, already surpassing prior baselines in *Format Compliance* and *Narrative Coherence*. However, the subsequent GRPO-based reinforcement learning stage yields pronounced improvements in more subjective dimensions, particularly *Dramatic Tension* and *Visual Imagery*. These results confirm that while SFT establishes structural competence, preference alignment is essential for capturing expert directorial aesthetics, enabling finer control over pacing, shot composition, and overall cinematic expressiveness.

5.2 Results of Video Generation

We evaluate SOTA text-to-video models in Table 2.

`ScripterAgent` enables faithful, temporally coherent video generation by translating dialogue into executable cinematic structure. Conditioning video models on structured scripts produced by `ScripterAgent` yields consistent improvements across all evaluated dimensions (Table 2). Aggregated results reveal a uniform uplift: the mean AI rating rises from 4.2 to 4.5, and the mean human rating increases from 3.7 to 4.2 . This efficacy is most pronounced in *Script Faithfulness*, with `Wan2.6` improving from 3.2 to 4.0 and `Sora2-Pro`

from 3.6 to 4.2, while HYVideo1.5 achieves the highest overall fidelity (4.6). Beyond semantic alignment, explicit shot-level blocking instructions enhance fine-grained execution, boosting AI-rated *Pace Timing* and *Body Blocking* by synchronizing motion with scene rhythm. Furthermore, gains in *Character Consistency* and *Narrative Coherence* validate DirectorAgent's Cross-Scene strategy; by coupling boundary-aware segmentation with frame-anchoring, the framework mitigates identity drift and extends coherent generation beyond single-model limits. Collectively, these results confirm that the domain-informed plans from ScripterAgent provide essential guidance absent in raw dialogue.

Our evaluation reveals a fundamental trade-off in SOTA models between visual spectacle and strict script adherence. The results expose a clear divergence in model capabilities that corroborates our third contribution. Sora2-Pro excels in visual impact, securing top scores in *Visual Appeal* (4.8) and *Physical Law* adherence (4.5), making it ideal for high-spectacle generation where realism is paramount. Conversely, HYVideo1.5 prioritizes narrative integrity, leading in *Script Faithfulness* (4.6), *Character Consistency* (4.4), and *Narrative Coherence* (4.3). This dichotomy suggests that current video models optimize along different axes: some prioritize perceptual realism, while others better maintain the semantic logic of a storyline when guided by structured scripts. This insight provides valuable guidance for practitioners selecting models for specific filmmaking applications.

6 Related Work

Video Generation. Recent video generation relies on diffusion models (Blattmann et al., 2023; He et al., 2022; Ho et al., 2022; Khachatryan et al., 2023; Singer et al., 2022; Bao et al., 2024; Brooks et al., 2024; Wan et al., 2025; Wu et al., 2025) and language models (Hong et al., 2022; Chang et al., 2022, 2023; Kondratyuk et al., 2023; Villegas et al., 2022). However, existing video generation systems still face significant limitations in managing long-form narrative coherence, especially when dealing with complex film scripts (Chen et al., 2025a). Our work addresses these challenges with a cross-scene generation strategy that mitigates fixed-duration constraints.

LLMs for Film Production. LLMs are increasingly used to automate film production tasks like scene generation, character planning, and cinematography (Lin et al., 2023; Wei et al., 2022). Systems such as Anim-Director (Li et al., 2024) help generate storylines and refine scenes, while MovieAgent (Wu et al., 2025) uses multi-agent collaboration for automated film creation. A key limitation of these models is their reliance on manual input for narrative and cinematographic planning. In contrast, our approach introduces a comprehensive end-to-end pipeline (Huang et al., 2025) that automates scene structuring and planning, reducing manual intervention and ensuring adherence to professional filmmaking standards.

Story Visualization. Story visualization maps scripts to visual sequences. Early methods (StoryGAN (Li et al., 2019)) produced static, temporally incoherent images. Recent diffusion models (StoryDiffusion (Zhou et al., 2024), Magic-Me (Ma et al., 2024b)) improve temporal consistency and motion but still lack automated high-level planning for cinematography, scene structure, and character interactions, thus requiring manual guidance. We introduce a multi-agent, chain-of-thought (CoT) framework that uses hierarchical reasoning to automate long-form movie generation, ensuring temporal consistency, narrative integrity, and visual appeal over extended durations.

7 Conclusion

In this work, we propose a script-centric agentic framework for long-form dialogue-to-cinematic video generation. The system generates executable shot-level plans via ScripterAgent, maintains continuity during execution with DirectorAgent, and validates results using CriticAgentalongside human experts. Supported by our ScriptBenchbenchmark and a two-stage training regime (SFT followed by GRPO with a hybrid reward), ScripterAgentsurpasses strong baselines in quality. Conditioning video models on these scripts improves faithfulness and long-horizon consistency, revealing a trade-off between visual spectacle and directorial adherence. Future work targets finer-grained controllability, including lip synchronization, precise dialogue-action alignment, and broader cinematographic styles.

8 Limitations

While our approach demonstrates promising results in static environments, we acknowledge several limitations in handling complex, dynamic scenarios. First, visual and physical hallucinations remain prevalent; generated outputs suffer from severe degradation in multi-subject scenes and frequently violate physical laws (e.g., object clipping, anatomical distortion), indicating a lack of explicit 3D scene understanding. Second, temporal consistency is fragile. Due to the progressive nature of frame generation, we observe "identity drift" where character appearance fluctuates across shots, alongside abrupt scene transitions that disrupt narrative flow. Finally, fine-grained controllability is limited. The model struggles with precise audio-visual alignment (e.g., lip-sync) and complex interaction execution. Addressing these issues requires a holistic approach, potentially involving higher-quality annotated datasets, 3D-guided priors, and hierarchical planning mechanisms for long-horizon coherence.

References

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

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, and 1 others. 2024. Video generation models as world simulators. *OpenAI Blog*, 1(8):1.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, and 1 others. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, and 1 others. 2025a. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*.

Siyan Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Xuyan Chi, Jian Cong, Qinpeng Cui, Qide Dong, Junliang Fan, and 1 others. 2025b. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. 2022. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, and 1 others. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

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

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, and 1 others. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6329–6338.

Yunxin Li, Haoyuan Shi, Baotian Hu, Longyue Wang, Jiashun Zhu, Jinyi Xu, Zhen Zhao, and Min Zhang. 2024. Anim-director: A large multimodal model powered agent for controllable animation video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. 2023. Videodirectorgpt: Consistent multi-scene

video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*.

Yan Ma, Yu Qiao, and Pengfei Liu. 2024a. Mops: Modular story premise synthesis for open-ended automatic story generation. *arXiv preprint arXiv:2406.05690*.

Ze Ma, Daquan Zhou, Xue-She Wang, Chun-Hsiao Yeh, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. 2024b. Magic-me: Identity-specific video customized diffusion. In *European Conference on Computer Vision*, pages 19–37. Springer.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, and 1 others. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2022. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, and 1 others. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Xinpeng Wang, Han Jiang, Zhihua Wei, and Shanlin Zhou. 2022. Chae: Fine-grained controllable story generation with characters, actions and emotions. *arXiv preprint arXiv:2210.05221*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, and 1 others. 2025. HunyuanVideo 1.5 Technical Report. *arXiv:2511.18870*.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. 2025. Automated movie generation via multi-agent cot planning. *arXiv preprint arXiv:2503.07314*.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Ying-Cong Chen. 2025. Seed-story: Multimodal long story generation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1850–1860.

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

A Supplementary Experiment

A.1 Additional experiments for Table 2

Model	AI Rating (0-5)					Avg.
	Cam. Artic.	*Body Block.*	*Visual Fidelity*	*Emotion Arc*	*Pace Timing*	
Raw Dialogue (w/o `ScripterAgent`)						
`Vidu2`	3.9	3.8	4.3	4.3	4.1	4.1
`Seedance1.5-Pro`	4.0	4.1	4.4	4.4	4.2	4.2
`Kling2.6`	3.8	3.8	4.2	4.2	4.0	4.0
`Wan2.6`	4.1	4.0	4.5	4.5	4.3	4.3
`HYVideo1.5`	3.8	3.7	4.2	4.2	4.0	4.0
`Sora2-Pro`	3.8	3.8	4.3	4.3	4.1	4.1
`Veo3.1`	3.8	3.8	4.2	4.3	4.0	4.0
Average Score	3.9	3.9	4.3	4.3	4.1	4.1
w/ `ScripterAgent`						
`Vidu2`	4.6	4.6	4.4	4.7	4.7	4.6
`Seedance1.5-Pro`	4.6	4.5	4.4	4.7	4.7	4.6
`Kling2.6`	4.6	4.6	4.4	4.7	4.8	4.6
`Wan2.6`	4.6	4.6	4.4	4.7	4.7	4.6
`HYVideo1.5`	4.6	4.6	4.4	4.7	4.7	4.6
`Sora2-Pro`	4.6	4.5	4.3	4.6	4.7	4.5
`Veo3.1`	4.6	4.5	4.4	4.6	4.7	4.6
Average Score	4.6	4.6	4.4	4.7	4.7	4.6

Table 3: Video generation evaluation on the `ScriptBench` test set by Qwen3-VL.

A.2 Ablation Study on Scripting and Agentic Components

A.2.1 Experimental Configuration

We evaluate the incremental impact of our framework across four stages:

- **Baseline (Raw Dialogue):** Generating videos directly from dialogue prompts without ScripterAgent.
- **w/ Script Only:** Conditioning on ScripterAgent's detailed scripts, but generating as a single long-horizon clip.
- **w/ Script + Segment:** Implementing shot-aware segmentation but without the frame-anchoring consistency mechanism.
- **Full Agent (Ours):** The complete pipeline featuring script-conditioning, segmentation, and *Frame-Anchoring* for visual continuity.

A.2.2 Quantitative Results Analysis

We evaluate the generated videos across five cinematic dimensions using a panel of three advanced LLMs as critics: Gemini 2.5 Pro, Qwen 3 VL, and GLM 4.6 V. As shown in Table 4, the **Full Agent** consistently outperforms baselines across all metrics and backbone models (Wan2.6 and HYVideo1.5).

The ablation study reveals a clear functional decoupling of our agents. *Visual Descriptive Fidelity* and *Kinetic Body Language & Blocking* show marked improvements with the introduction of *w/ Script Only*, attributing fine-grained visual control to the `ScripterAgent`. Conversely, the `DirectorAgent` is shown to be critical for temporal dimensions; the implementation of shot-aware segmentation and frame-anchoring drives scores in *Narrative Pacing & Timing* and *Cinematic Camera Articulation* to their highest levels (e.g., improving Pace Timing on Wan2.6 from 3.2 to 4.7). The high agreement among the three diverse critic models strongly corroborates the validity of these improvements.

B `CriticAgent`: The Evaluation Framework

To assess our system comprehensively, we introduce a multifaceted evaluation framework that examines both stages of our pipeline: script generation (dialogue-to-script) and video generation (script-to-video). This framework is essential because cinematic quality is inherently multidimensional, encompassing technical correctness, narrative fidelity, and subjective artistic merit. The framework combines objective metrics, automated scoring via our AI-powered `CriticAgent`, and qualitative evaluations by human experts. All scores are assigned on a 0–5 scale.

B.1 Script Generation Evaluation

For the dialogue-to-script stage, we evaluate the generated scripts on both their structural correctness, using our `CriticAgent`, and their artistic quality, via a panel of human experts.

AI Rating (`CriticAgent`) We employ our `CriticAgent`, powered by `gemini-2.5-pro`, to automatically assess generated scripts based on four criteria:

1. **Format Compliance.** Assesses strict adherence to the required JSON format, ensuring all key

Setting	Gemini 2.5 Pro					Qwen 3 VL					GLM 4.6 V				
	Cam. Artic.	*Body Block.*	*Visual Fidelity*	*Emotion Arc*	*Pace Timing*	*Cam. Artic.*	*Body Block.*	*Visual Fidelity*	*Emotion Arc*	*Pace Timing*	*Cam. Artic.*	*Body Block.*	*Visual Fidelity*	*Emotion Arc*	*Pace Timing*
Wan2.6															
Baseline	3.0	2.8	3.1	3.2	3.2	2.8	2.8	2.8	2.8	3.1	2.7	3.2	2.9	2.8	2.9
w/ Script Only	3.1	3.4	3.4	3.7	3.7	3.5	3.4	3.6	3.5	3.7	3.3	3.7	3.4	3.5	3.5
w/ Script+Seg.	3.8	4.0	4.1	4.1	4.1	4.0	3.8	3.9	3.8	4.2	3.9	4.1	3.9	4.0	4.0
Full Agent	**4.4**	**4.6**	**4.7**	**4.6**	**4.7**	**4.6**	**4.6**	**4.4**	**4.7**	**4.7**	**4.5**	**4.8**	**4.5**	**4.5**	**4.5**
HYVideo1.5															
Baseline	2.9	3.0	3.3	2.7	2.9	2.8	2.9	3.2	2.6	3.1	3.1	2.8	3.2	3.1	3.0
w/ Script Only	3.1	3.4	3.7	3.2	3.5	3.3	3.7	3.9	3.2	3.5	3.5	3.3	3.8	3.2	3.4
w/ Script+Seg.	3.7	3.9	4.3	3.9	4.1	3.7	4.1	4.3	3.8	4.2	4.0	3.6	4.3	4.1	4.0
Full Agent	**4.4**	**4.5**	**4.8**	**4.5**	**4.7**	**4.6**	**4.6**	**4.4**	**4.7**	**4.7**	**4.6**	**4.2**	**4.9**	**4.6**	**4.6**

Table 4: Ablation results on Wan2.6 and HYVideo1.5 across three AI evaluators. All scores are on a 0-5 scale.

fields (e.g., "Shot Type", "Camera Movement", "Description") are present and correctly structured.

2. **Shot Division Rationality.** Evaluates the logical segmentation of the script into shots, ensuring that breaks align with narrative beats and emotional shifts without being overly fragmented or lengthy.

3. **Content Completeness.** Measures whether the script provides rich, actionable details for filming and enriches the narrative with visual information absent from the source dialogue.

4. **Narrative Coherence.** Determines whether the sequence of shots is logically connected and if the visual storytelling flows smoothly to complement the dialogue's context.

Human Rating (Directors' Panel) To complement the automated assessment, a panel of professional directors and screenwriters evaluates the artistic quality of the scripts. They provide ratings on a 0–5 scale for three key creative aspects, which collectively indicate the script's potential for being filmed successfully:

1. **Character Portrayal Consistency.** Assesses whether each character's personality, speaking style, and behavior remain coherent and believable throughout the script.

2. **Dramatic Tension & Rhythm.** Measures the script's effectiveness in building, sustaining, and releasing dramatic tension, as well as the naturalness and engagement of its pacing.

3. **Visual Imagery & Cinematic Expressiveness.** Assesses how vividly the script conveys visual information and how effectively it employs cinematic language (e.g., shots, staging, atmosphere) to support the narrative.

B.2 Video Generation Evaluation

We evaluate the script-to-video generation stage on two primary axes: script-video alignment and overall video quality. In addition, we conduct automatic evaluation that combines standard video quality metrics with a novel measure of script alignment.

AI Rating (`CriticAgent`) For the video generation stage, `CriticAgent`evaluates the cinematic quality and faithfulness of the generated video to the source script and reference audio across five dimensions:

1. **Cinematic Camera Articulation.** Measures the sophistication of camera work, including shot types, framing transitions, and dynamic movements that support the scripted narrative.

2. **Kinetic Body Language & Blocking.** Assesses whether character motions, physical interactions, and spatial arrangements are specific, expressive, and consistent with the scripted actions.

3. **Visual Descriptive Fidelity.** Evaluates how well the visual details (e.g., character appearance, clothing textures, scene layout, lighting) match the descriptive cues in the script.

4. **Emotional Arc & Micro-Expressions.** Examines whether the facial expressions, subtle gestures, and temporal evolution reflect the intended emotional progression in the script and audio delivery.

5. **Narrative Pacing & Timing.** Measures the alignment of shot timing, action beats, and pauses with the narrative structure and rhythm implied by the script and audio.

Human Rating Human annotators also assess the final generated videos, providing ratings on five dimensions that collectively offer a comprehensive view of video quality:

1. **Visual Appeal.** Evaluates the realism, aesthetic quality, and rendering stability of the video.

2. **Script Faithfulness.** Assesses how accurately the video adheres to the provided script in terms of scenes, actions, and plot progression.

3. **Narrative Coherence.** Measures whether the video forms a logically consistent and easy-to-follow story, with reasonable scene transitions and pacing.

4. **Character Consistency.** Evaluates whether characters maintain a stable identity and appearance throughout the video.

5. **Physical Law Adherence.** Assesses whether motions and interactions in the video plausibly adhere to real-world physical laws, contributing to natural-looking dynamics.

C Experimental Details

C.1 Training Infrastructure and Cost

All experiments were conducted on a compute node equipped with 8 × H20 GPUs, fully utilizing all available GPUs in parallel. The Supervised Fine-Tuning (SFT) stage for `ScripterAgent` required approximately 8 GPU hours, while the subsequent GRPO-based Reinforcement Learning stage consumed 192 GPU hours to converge.

C.2 Baselines

Script Generation We compare against representative story visualization and screenplay generation methods:

- *MoPS* (Ma et al., 2024a) proposes a modular framework for automated story premise synthesis by decomposing premises into theme, background, persona, and plot modules, then recombining them via a nested dictionary and LLM-based integration to generate diverse, high-quality story foundations.

- *CHAE* (Wang et al., 2022) enables fine-grained controllable story generation by allowing users to specify characters, their actions, and emotions through a structured input format, enhanced with a copy mechanism and character-wise emotion loss for precise narrative control.

- *SEED-Story* (Yang et al., 2025) extends multimodal large language models to generate long, coherent narratives interleaved with images, utilizing a multimodal attention sink mechanism to maintain consistency and enable generation beyond training sequence lengths.

Video Generation We evaluate SOTA text-to-video models:

- `Vidu2` (Bao et al., 2024) (Shengshu Technology, Tsinghua University): A U-ViT-based model excelling in temporal consistency and generation speed.

- `Seedance1.5-Pro` (Chen et al., 2025b) (ByteDance Seed): A high-fidelity diffusion model specialized in generating professional-grade videos with enhanced dynamic coherence and visual detail.

- `Kling2.6` (Kuaishou Technology): A Transformer-based generative model capable of synthesizing high-resolution videos with improved motion fluidity and scene understanding.

- `Wan2.6` (Wan et al., 2025) (Alibaba): A diffusion-based framework synthesizing realistic scenes with intricate details and smooth temporal dynamics.

- `HYVideo1.5` (Wu et al., 2025) (Tencent): A large-scale video generation framework featuring a dual-stream diffusion transformer architecture that achieves state-of-the-art performance in prompt following and 4K-resolution synthesis.

- `Sora2-Pro` (Brooks et al., 2024) (OpenAI): A diffusion transformer model generating high-fidelity, physically plausible videos with complex scenes.

- `Veo3.1` (Google DeepMind): A generative model creating high-resolution (e.g., 1080p) videos with strong cinematic quality and motion coherence.

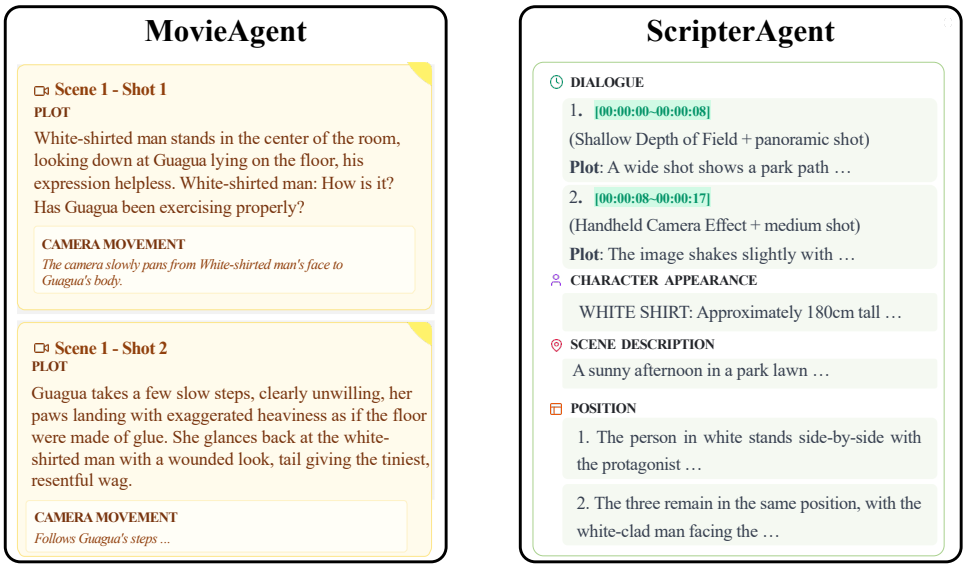

Figure 4: Qualitative comparison: `ScripterAgent` vs. MovieAgent (Wu et al., 2025).

C.3 Case Study

Script Generation Qualitative analysis further highlights `ScripterAgent`'s ability to generate a detailed, executable filmmaking plan. A direct comparison of generated outputs reveals the practical superiority of our approach, as shown in Figure 4. The script from `MovieAgent` provides a simple plot summary, such as "Camera slowly pans...". In stark contrast, the `ScripterAgent` output constitutes a complete, fine-grained cinematic blueprint. It specifies precise technical details like camera settings ("Shallow Depth of Field + panoramic shot"), timestamps for synchronization ("[00:00:00 00:00:08]"), detailed character appearance ("Approximately 180cm tall"), atmospheric scene descriptions, and exact character positioning or "blocking". This richness confirms that `ScripterAgent` successfully generates the professional-quality, executable script needed to guide automated video production, directly addressing the core challenge outlined in our introduction.

Video Generation Detailed scripts also improve fine-grained temporal alignment. Beyond macro-level coherence, scripts from `ScripterAgent` lead to marked enhancements in fine-grained synchronization. This indicates that shot-level planning helps models better align visual pacing with narrative beats. Qualitative examples of these improvements are provided in Figure 5.

D Details on Evaluation Metrics

In this section, we detail the human evaluation criteria used in our experiments and all scores are assigned on a 0–5 scale.

D.1 Human Evaluation for Script Generation

Annotators read the source dialogue and the generated script, then assign 0–5 scores for each metric according to the following criteria.

Character Portrayal Consistency. This metric evaluates whether each character's personality, speaking style, and behavior remain coherent and believable throughout the script.

- *Score 0 (Invalid or Off-Topic):* The script is largely unrelated to the given dialogue, missing key characters or scenes, or is so fragmented that character portrayal cannot be meaningfully judged.
- *Score 1 (Completely Inconsistent):* Characters frequently change personality, tone, or goals without any narrative justification, making them feel arbitrary or incomprehensible.
- *Score 2 (Severe Inconsistencies):* Major contradictions in speech style or behavior appear, and characters often feel unstable or unconvincing.
- *Score 3 (Noticeable Variations):* Core traits are roughly maintained, but several noticeable shifts in dialogue or actions break immersion.
- *Score 4 (Good Consistency):* Characters are generally consistent in personality and voice, with

(a) Example1

(b) Example2

Figure 5: Examples of video generation using `Sora2-Pro`.

only minor deviations that do not seriously affect believability.

- *Score 5 (Perfect Consistency):* Each character maintains a stable and well-defined identity, with coherent speech and behavior in all scenes.

Dramatic Tension & Rhythm. This metric measures how well the script builds, sustains, and releases dramatic tension, and whether the pacing of events feels natural and engaging.

- *Score 0 (Invalid or No Story):* The text does not form a recognizable story (e.g., mostly noise, repetition, or unrelated fragments), so dramatic structure cannot be evaluated.
- *Score 1 (No Dramatic Structure):* The script lacks recognizable conflicts or turning points, and the pacing feels random or monotonous.
- *Score 2 (Weak and Uneven Tension):* Some conflicts exist but are poorly set up or resolved, with abrupt or dragging sections that hurt engagement.
- *Score 3 (Basic but Uneven Rhythm):* The overall arc is understandable, but there are noticeable pacing issues (e.g., rushed climaxes or overlong trivial scenes).
- *Score 4 (Good Dramatic Arc):* Conflicts, climaxes, and resolutions are well-structured, with generally appropriate pacing and emotional build-up.
- *Score 5 (Strong and Compelling Tension):* The script offers a clear and engaging dramatic curve, with well-timed beats that sustain viewer interest throughout.

Visual Imagery & Cinematic Expressiveness. This metric assesses how vividly the script conveys visual information and how well it uses cinematic language (shots, staging, atmosphere) to support filming.

- *Score 0 (Not Filmable / Unreadable):* The script is so incomplete, disorganized, or off-topic that it provides no meaningful basis for imagining shots or scenes.
- *Score 1 (Vague and Non-Visual):* Descriptions are extremely abstract, offering almost no cues for camera work, staging, or visual composition.
- *Score 2 (Sparse Visual Guidance):* Only a few scattered visual hints are provided; most scenes are difficult to imagine concretely on screen.
- *Score 3 (Basic Visual Clarity):* Key scenes and actions are described clearly enough to imagine, but shot types or cinematic details remain generic.
- *Score 4 (Good Cinematic Guidance):* The script includes clear descriptions of scenes, actions, and rough shot intentions, enabling straightforward visualization.
- *Score 5 (Highly Cinematic and Filmable):* The script shows rich visual imagination and appropriate use of film language, making it easy to translate into professional storyboards.

D.2 Human Evaluation for Video Generation

Annotators watch each generated video with access to the corresponding script (or dialogue) as reference, and then give 0–5 scores according to the following guidelines.

Visual Appeal. This metric evaluates the realism, aesthetic quality, and rendering stability of the video.

- *Score 0 (Missing or Corrupted Video):* The video cannot be properly viewed (e.g., file error, almost entirely static or black frames), so visual quality is not assessable.
- *Score 1 (Severe Artifacts):* Heavy distortions, glitches, or unrecognizable content dominate most frames, making the video difficult to watch.
- *Score 2 (Poor Visual Quality):* Frequent flickering, unstable textures, and obvious inconsistencies, though core content is still somewhat interpretable.
- *Score 3 (Moderate Issues):* Overall content is clear, but noticeable artifacts in textures, lighting, or motion transitions reduce visual quality.
- *Score 4 (Good Visual Quality):* Generally realistic and stable visuals with only minor distortions or occasional flickering that do not strongly affect viewing.
- *Score 5 (High-Quality Rendering):* Smooth, realistic, and aesthetically pleasing visuals, comparable to professionally produced footage, with minimal visible artifacts.

Script Faithfulness. This metric assesses how accurately the video follows the intended script or ScriptAgent-generated plan in terms of scenes, actions, and plot progression.

- *Score 0 (No Relation or No Reference):* The video has almost no identifiable connection to the given script (or the script is missing/invalid), making faithfulness impossible to judge.
- *Score 1 (Almost Unrelated):* The video bears little or no resemblance to the script, missing key scenes, characters, or events.
- *Score 2 (Major Deviations):* Some script elements appear, but important settings, actions, or turning points are missing, misplaced, or heavily altered.
- *Score 3 (Partial Alignment):* The main storyline can be recognized, but there are noticeable inaccuracies or omissions in details and shot arrangement.
- *Score 4 (Good Faithfulness):* Most key moments and settings match the script, with only minor deviations that do not significantly affect overall story understanding.
- *Score 5 (Almost Perfect Adherence):* The video closely follows the script in both content and structure, accurately reflecting specified scenes, actions, and plot beats.

Narrative Coherence. This metric measures whether the video forms a logically consistent and easy-to-follow story, with reasonable scene transitions and pacing.

- *Score 0 (No Discernible Narrative):* The video is so fragmented, repetitive, or random that no coherent storyline or temporal order can be inferred.
- *Score 1 (Completely Incoherent):* Scenes appear random and disconnected, with no understandable storyline or causal relations.

- *Score 2 (Frequent Confusion):* Some story intention is visible, but abrupt cuts and illogical transitions make the plot hard to follow.
- *Score 3 (Partially Coherent):* A basic story can be inferred, yet inconsistencies or awkward pacing often disrupt narrative flow.
- *Score 4 (Well-Structured Story):* The video presents a mostly clear and coherent narrative, with only minor issues in pacing or transitions.
- *Score 5 (Fully Coherent and Engaging):* The plot develops smoothly and logically, with seamless scene transitions and a clear, engaging storytelling arc.

Character Consistency. This metric evaluates whether characters maintain a stable identity and appearance across the entire video.

- *Score 0 (Characters Not Identifiable):* Human or main characters are almost entirely missing, severely deformed, or indistinguishable, so consistency cannot be meaningfully evaluated.
- *Score 1 (Completely Inconsistent):* Character faces, bodies, or outfits change drastically between shots, making them hard to recognize as the same person.
- *Score 2 (Severe Inconsistencies):* Frequent noticeable changes in facial features, clothing, or proportions, even if some continuity is preserved.
- *Score 3 (Moderate Variations):* Characters are generally recognizable, but variations in appearance or style appear multiple times and affect immersion.
- *Score 4 (Good Consistency):* Character identity and look are mostly stable, with only small visual fluctuations that do not seriously disrupt continuity.
- *Score 5 (Perfect Consistency):* Characters keep a highly stable appearance and identity across all shots, making them visually coherent throughout the video.

Physical Law Adherence. This metric assesses whether motions and interactions in the video roughly follow real-world physical laws, contributing to natural-looking dynamics.

- *Score 0 (Motion Not Assessable):* The video is nearly static, heavily corrupted, or too abstract, so physical plausibility of motion cannot be reasonably judged.
- *Score 1 (Highly Unrealistic):* Objects or characters frequently move in impossible ways (e.g., floating, severe limb distortions) without any narrative justification.
- *Score 2 (Many Violations):* Multiple obvious physical inconsistencies (e.g., unnatural collisions, gravity-defying motion), although not in every scene.
- *Score 3 (Partially Plausible):* Most movements are acceptable, but there are several noticeable errors in weight, balance, or continuity of motion.
- *Score 4 (Mostly Realistic):* Characters and objects move in a generally natural way, with only minor physical anomalies that are easy to overlook.
- *Score 5 (Highly Plausible Physics):* Motion and interactions appear smooth and physically convincing, with no obvious violations of basic physical laws.

D.3 AI-Based Rating Metrics

To complement objective metrics, we use LLM-based evaluators to score both *script generation* and *video generation* on a 0–5 scale aligned with film production standards. For scripts, Gemini-2.5-Pro focuses on structural and logical correctness; for videos, Gemini-2.5-Pro focuses on perceptual and technical audio-visual quality. Below we show the full prompts and the detailed scoring rubric for each dimension.

Script Generation Evaluation Prompt

You are a professional film director and script supervisor. Your task is to evaluate the quality of a generated shooting script based on the provided coarse-grained dialogue and context.
Input Data:

- **Source Dialogue:** {Insert Origin Dialogue Here}
- **Generated Script:** {Insert Generated JSON Script Here}

Evaluation Criteria: Please score the generated script on a scale of 0 to 5 for each of the following dimensions. For each dimension, use the following general guideline:

- *Score 0:* Completely unusable or fails the requirement.
- *Score 1:* Very poor quality; severe issues in most parts.
- *Score 2:* Clearly below acceptable quality; many issues.
- *Score 3:* Acceptable but with noticeable issues.
- *Score 4:* Good quality with only minor issues.
- *Score 5:* Excellent quality; no meaningful issues.

Then, judge each dimension more concretely as follows:

1. **Format Compliance (0–5):** Does the output strictly follow the required JSON format? Are all key fields (Shot Type, Camera Movement, Description, etc.) present and correctly structured?
   - *0:* Not valid JSON or completely ignores the requested schema.
   - *1:* Severe structural errors; many missing or malformed fields.
   - *2:* Multiple structural problems; only partially follows the schema.
   - *3:* Mostly follows the schema but with some missing fields or minor format issues.
   - *4:* Fully follows the schema with only very small formatting inconsistencies.
   - *5:* Perfectly formatted JSON with all fields correctly present and structured.
2. **Shot Division Rationality (0–5):** Is the script segmented into shots reasonably? Do the shot breaks align with narrative beats and emotional shifts without being too fragmented or too long?
   - *0:* No meaningful shot division; essentially a single block or random splitting.
   - *1:* Very unreasonable segmentation; shots break the flow and ignore story structure.
   - *2:* Many inappropriate shot boundaries; frequent over- or under-segmentation.
   - *3:* Basic correspondence to narrative beats, but with several awkward or suboptimal shot splits.
   - *4:* Mostly well-aligned with emotional and narrative shifts, with only minor segmentation issues.
   - *5:* Shot division is highly reasonable, closely following narrative and emotional structure throughout.
3. **Content Completeness (0–5):** Does the script provide rich, actionable details for filming? Does it supplement necessary visual information that was missing in the source dialogue?
   - *0:* Almost no additional visual or staging information beyond the raw dialogue.
   - *1:* Very sparse detail; crucial information for filming is largely missing.
   - *2:* Some useful details, but important aspects (scene, actions, camera) remain underspecified.
   - *3:* Contains enough information to stage the scene, but important visual details are still missing.
   - *4:* Generally rich and specific, covering most necessary visual, spatial, and action details.
   - *5:* Highly complete and specific, providing clear and thorough guidance for key visual choices.
4. **Narrative Coherence (0–5):** Is the sequence of shots logically connected? Does the visual storytelling flow smoothly and match the context of the dialogue?
   - *0:* Completely incoherent sequence; shots appear random and unrelated to the dialogue.
   - *1:* Very confusing progression; frequent contradictions or abrupt jumps.
   - *2:* A rough story is visible, but there are many logical gaps, contradictions, or unnatural transitions.
   - *3:* Overall story is understandable, but several transitions or details break the narrative flow.
   - *4:* Mostly coherent and smoothly flowing narrative with only minor inconsistencies.
   - *5:* Fully coherent, well-structured visual narrative that aligns closely with the dialogue context.

Output Format: Return the result in the following JSON format:

```
{
  "Format Compliance": [Score],
  "Shot Division Rationality": [Score],
  "Content Completeness": [Score],
  "Narrative Coherence": [Score]
}
```

Video Generation Evaluation Prompt

You are an expert AI Film Critic and Cinematographer with deep expertise in visual storytelling, camera techniques, and cinematic language. Your task is to evaluate the video's cinematic quality and adherence to complex directorial instructions.
Input Data:

- **Reference Script:** {Insert Reference Script Here}
- **Generated Video:** {Video File to be Evaluated}

Evaluation Criteria: Please score the video on a scale of 0.0 to 5.0 for each of the following cinematic dimensions. You can assign ANY decimal score (e.g., 2.3, 3.7, 4.2). The integer benchmarks (0, 1, 2, 3, 4, 5) serve as REFERENCE POINTS for quality boundaries.
IMPORTANT: Simple dialogue videos with minimal movement should receive LOW scores (1-2). HIGH scores (4-5) are reserved ONLY for videos demonstrating sophisticated cinematic techniques.

- *Score 0:* Completely fails the requirement; no evidence of the evaluated quality.
- *Score 1:* Minimal/default quality; severe problems throughout.
- *Score 2:* Basic quality; many noticeable issues.
- *Score 3:* Competent/functional quality; acceptable but uninspired.
- *Score 4:* Advanced/dynamic quality; well-executed with minor issues.
- *Score 5:* Master-level quality; exceptional execution indistinguishable from professional cinema.

Then, judge each dimension more concretely as follows:

1. **Cinematic Camera Articulation (0.0–5.0):** Evaluates the sophistication and intentionality of camera work, including movement, framing transitions, and visual storytelling techniques.
   - *0:* Completely static camera; single unchanging framing; feels like a frozen screenshot.
   - *1:* Predominantly static with occasional accidental shifts; simple linear zoom with no artistic purpose; AI default setting.
   - *2:* Simple panning/tilting with uniform speed; basic zoom uncorrelated with narrative; abrupt transitions.
   - *3:* Clear shot variety (Wide/Medium/Close-up); camera movements motivated by action; demonstrates basic cinematic grammar but mechanical execution.
   - *4:* Purposeful dynamic techniques (handheld shake, tracking shots, focal shifts); smooth transitions aligned with narrative peaks; camera positioning creates visual tension.
   - *5:* Exceptional sophisticated camera language (fluid handheld, crane shots, dolly moves); perfect composition; focal shifts precisely timed; every camera decision serves narrative purpose.
2. **Kinetic Body Language & Blocking (0.0–5.0):** Assesses physical performance quality, spatial relationships (blocking), and how bodies express narrative and emotion.
   - *0:* Characters completely static like mannequins; no gestures or facial movement; zero physicality.
   - *1:* Only basic lip movement; stiff/repetitive gestures; no spatial repositioning; AI-generated feel with no human quality.
   - *2:* Simple gestures lacking fluidity; mechanical walking; spatial relationships accidental; gestures don't match emotional context.
   - *3:* Characters move with basic purpose (A to B); contextually appropriate but unspecific gestures; basic blocking present; believable but not expressive.
   - *4:* Highly specific actions (running and stopping at precise point, leaning forward, catching breath); intentional spatial blocking creates tension; body language evolves through scene.
   - *5:* Every action precise and motivated; micro-movements (fidgeting, weight shifts); complex sequences; blocking choreographed to perfection reflecting power dynamics; culturally-specific gestures.
3. **Visual Descriptive Fidelity (0.0–5.0):** Measures how accurately visual output matches script descriptions, including character appearance, clothing textures, environmental details, and lighting.
   - *0:* Characters look random or change appearance; environment blank/incoherent; lighting broken.
   - *1:* Characters vaguely human but bear no resemblance to descriptions; generic clothing contradicts script; lighting ignores time-of-day cues.
   - *2:* Gender/age match but specific features wrong; clothing category correct but textures/colors incorrect; environment thematically correct but lacks details.

- *3:* Major descriptors match (gender, age, clothing style); environment includes key elements but simplified; lighting matches time-of-day but lacks detailed effects.
- *4:* Characters closely match detailed descriptions (hair style, clothing textures, body type); environment shows specific details (pavement texture, metal railings); sophisticated lighting with proper shadows.
- *5:* Photorealistic precision with every descriptor present; micro-details (fabric wrinkles, button placement); environmental lighting interacts realistically; atmospheric depth; indistinguishable from high-end cinematography.

4. **Emotional Arc & Micro-Expressions (0.0–5.0):** Evaluates range and authenticity of emotional performance, including facial expressions, emotional transitions, and psychological subtext.
   - *0:* Faces blank/frozen/mask-like; no visible emotional state; characters appear lifeless.
   - *1:* Single unchanging expression; clearly looped frames; no emotional reaction to events; contradicts narrative context; robotic feel.
   - *2:* One or two basic emotions expressed simplistically/exaggerated; abrupt transitions; lacks nuance; doesn't align with dialogue tone; cartoonish.
   - *3:* Emotional states generally match dialogue; at least one clear shift; basic facial movements (eyebrow raises, mouth changes); recognizable but generic; lacks micro-expressions.
   - *4:* Multiple distinct emotional states with clear transitions (laugh→serious→questioning); nuanced details (eyebrow furrows, eye contact changes); micro-expressions present; emotional intensity varies appropriately; feels "acted" not generated.
   - *5:* Rich layered emotional journey; complex arcs (playful→realization→concern→resolve); exceptional micro-expressions (1-2 frame fleeting expressions); emotions blend naturally; character-consistent and psychologically motivated; subtext visible; indistinguishable from professional actor performance.

5. **Narrative Pacing & Timing (0.0–5.0):** Assesses whether video executes clear narrative structure with appropriate timing, action sequencing, and rhythmic flow matching script's story beats.
   - *0:* Video incoherent; no discernible beginning/middle/end; actions occur randomly; timing completely broken.
   - *1:* Duration wildly mismatches dialogue; actions in wrong order or omitted; no logical flow; feels like random clips stitched together.
   - *2:* Length approximately matches but internal pacing off; key actions happen at wrong times; some narrative beats present but sequencing confused; rhythm monotonous or chaotic.
   - *3:* Duration matches dialogue; basic narrative sequence present (setup→event→conclusion); actions in correct order; pacing acceptable but lacks dynamic variation; competent but uninspired.
   - *4:* Dialogue and action timing precisely synchronized; clear purposeful structure (setup→action→escalation→resolution); pacing creates rhythm; timing builds/releases tension appropriately; action sequences follow believable physics.
   - *5:* Perfect narrative timing with cinematic rhythm; three-act structure compressed into scene; actions timed with precision to the second; rhythmic variation creates emotional texture; timing builds and releases tension masterfully; pacing feels inevitable and organic; indistinguishable from professionally edited film.

Output Format: Return ONLY a JSON structure with decimal scores (0.0-5.0), detailed reasoning for each dimension (referencing which benchmarks the video falls between), Final Cinematic Grade (average of all 5 scores), and Overall Assessment.
Scoring Reminders:

- Use decimal precision (e.g., 2.3, 3.7, 4.5) to distinguish quality levels
- Reference integer benchmarks but don't feel limited to them
- Explain in reasoning which benchmarks the video falls between and why
- Simple dialogue videos with minimal movement should score 1.0-2.5
- Only sophisticated, cinema-quality videos should receive scores of 4.0-5.0