Stable Cinemetrics : Structured Taxonomy and Evaluation for Professional Video Generation

Agneet Chatterjee[1,2,*] Rahim Entezari[1] Maksym Zhuravinskyi[1]
Max Lapin[1] Reshinth Adithyan[1] Amit Raj[3]
Chitta Baral[2] Yezhou Yang[2] Varun Jampani[1]

[1]Stability AI [2]Arizona State University [3]Google DeepMind

https://stable-cinemetrics.github.io/

Abstract

Recent advances in video generation have enabled high-fidelity video synthesis from user provided prompts. However, existing models and benchmarks fail to capture the complexity and requirements of professional video generation. Towards that goal, we introduce **S**table **Cine**metrics, a structured evaluation framework that formalizes filmmaking controls into four disentangled, hierarchical taxonomies: *Setup, Event, Lighting, and Camera*. Together, these taxonomies define 76 fine-grained control nodes grounded in industry practices. Using these taxonomies, we construct a benchmark of prompts aligned with professional use cases and develop an automated pipeline for prompt categorization and question generation, enabling independent evaluation of each control dimension. We conduct a large-scale human study spanning 10+ models and 20K videos, annotated by a pool of 80+ film professionals. Our analysis, both coarse and fine-grained reveal that even the strongest current models exhibit significant gaps, particularly in Events and Camera-related controls. To enable scalable evaluation, we train an automatic evaluator, a vision-language model aligned with expert annotations that outperforms existing zero-shot baselines. SCINE is the first approach to situate professional video generation within the landscape of video generative models, introducing taxonomies centered around cinematic controls and supporting them with structured evaluation pipelines and detailed analyses to guide future research.

1 Introduction

The field of video generative models has made significant progress in recent years [37], drawing substantial interest from both academia and industry. This can be evidenced by the growing number of benchmarks [25, 32, 2, 35], datasets [41, 58], and both open- [3, 56, 14, 65] and closed- [6, 54, 36] source models that have collectively driven the field forward. The foundational nature of these models makes them useful for several downstream tasks, including video editing [26], 3D generation [55] and robotics [69]. This widespread adoption of video generative models underpins the growing assertion that they represent a revolution for *professional video generation*.

Generative vision offers tremendous potential for media creation, but a fundamental question remains: how can we shift generative video from casual, exploratory synthesis to a medium that supports professional-grade, controllable cinematic outputs? The important distinction between casual and professional generative video lies in the critical gap of cinematic control [7]: while today's models can generate videos of "*an astronaut riding a horse*", professional creation necessitates granular control over cinematic elements such as the framing of the shot, position of the key light, and even whether the astronaut smiles before *or* after the horse gallops away - a truly professional video generation system must put every one of those cinematic choices back in the creator's hands. The need for this exact control over every cinematic element, from the timing of a smile to the quality of

*Work done during an internship at Stability AI.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

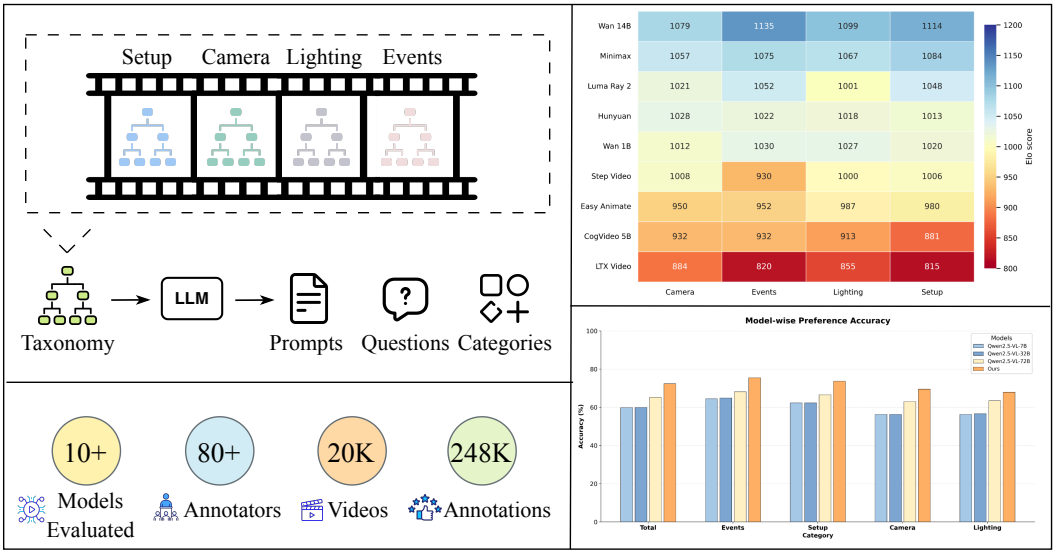


Figure 1: **Stable Cinemetrics** introduces structured taxonomies grounded in the controls required for professional video generation. These taxonomies form the foundation of our prompt based benchmark that mirrors real-world shot creation, progressing from scriptwriting to on-screen visuals. Every control element in a prompt is automatically categorized back to the taxonomy, enabling the generation of isolated evaluation questions for independent investigation into each element. This supports large scale human evaluation enabling both coarse and fine-grained insights into the capabilities of current models for professional video generation. To drive scalable annotations, we develop our own VLMs that outperform existing models in alignment with human judgements.

light, is the very reason filmmakers shoot multiple takes, selecting only the frames where everything comes together to tell the story most effectively [10].

In this work, we investigate the intersection of video generative models and the nuanced control mechanisms associated with professional video production. Despite rapid advancements in video generative modeling, the field still lacks both a clear definition of essential cinematic controls and standardized evaluation protocols for benchmarking progress at this crucial intersection. To bridge this gap, we present SCINE (**S**table **Cine**metrics) - an evaluation suite specifically designed to characterize this intersection, enabling us to directly address the question: "*Are Current Video Generative Models Ready for Professional Use?*"

The core focus of SCINE is to develop a taxonomy, at the intersection of generative vision and professional video generation, built on the principles of control and specificity [4, 28, 43]. These principles are important in film-making, as every decision carries cinematic meaning; a low angle conveys power while low-key lighting evokes drama. SCINE captures these principles by organizing control knobs into four cinematic pillars - Setup, Events, Lighting, and Camera - enabling the evaluation of video generative models along the same axes that a professional would. This disentanglement also allows evaluation around *personalization*: which aligns with the collaborative nature of real productions, in comparison to the monolithic nature of current video generation models. Our taxonomy is hierarchical, branching from coarse cinematic concepts to leaf-level controls that naturally map onto computer vision concepts such as object semantics and scene geometry [49].

Leveraging our taxonomy, we generate two prompt types: story-driven and visual exposition, to mirror professional workflows. Story-driven prompts act as mini-screenplays [17], specifying characters, dialogue, actions, and emotions. We enrich these with visual exposition cues, by sampling control nodes from our taxonomy, emulating the transition from script to shot in filmmaking [46]. Sampling control nodes allows automated (a) prompt categorization: mapping each control element to the taxonomy, and (b) generation of targeted evaluation questions for each element, allowing disentagled evaluation of each cinematic control. The structured nature of our taxonomy supports scalable human annotation: we evaluate 10+ models across 20K generated videos with feedback from 80+ professionals. This enables analysis across control dimensions where we observe substantial variance in performance across taxonomy pillars, even for top-performing models such as WAN-14B and Minimax. Our taxonomy facilitates both coarse insights, showing that models struggle most with

Events and Camera and fine-grained comparisons, such as better performance on shot size over camera framing, and on natural over artificial lighting. To support automatic evaluation of fine-grained cinematic controls in generated videos, we train a vision-language model (VLM) that aligns with the large-scale human annotations. Our model outperforms existing baselines, achieving an overall accuracy of 72.36% with human annotators. An overview of SCINE, outlining its contributions is shown in Figure 1.

2 Related Work

Video Generative Models and Evaluations. Video generative models can broadly be classified into two categories based on their input conditioning: image-to-video (I2V) and text-to-video (T2V). While I2V approaches such as Stable Video Diffusion (SVD) [3] have been widely adopted by the community, the focus of our work is to evaluate T2V models for professional use. We choose text as an input modality, since it is an effective and free-form way of describing the controls defined in our taxonomy. The Sora Preview [6] served as a catalyst for a wave of T2V model releases across the closed [54, 11] and open-source communities. State of the art open-source models include Wan [56], HunyuanVideo [14] and StepVideo [12]. Several T2V evaluation benchmarks have also emerged, with VBench [25, 67] gaining broad adoption. VBench evaluates T2V models by developing text prompts that evaluate generated videos across dimensions such as temporal flickering, aesthetic quality and, motion smoothness, while employing automatic metrics. Additional T2V benchmarks include - VideoPhy [2], which evaluates physical plausibility by measuring adherence of generated videos to real-world physics, and T2V-CompBench [48], which studies compositional consistency [23] in video generation. Prior work lacks the shot-level structure and cinematic detail needed for professional control; motivating our taxonomy and benchmark to capture the nuanced elements of industry-standard filmmaking. Existing benchmarks lack cinematic depth; for example, a prompt such as "*A man is walking*" from VBench-2 [67] omits key details like character appearance or camera movement; all essential to setup a cinematic shot. Existing benchmarks are static, relying on fixed prompt sets that limit extensibility. In contrast, SCINE's taxonomy-guided prompt generation enables future-proof evaluation, allowing prompt complexity to scale with model capabilities.

Structured Video Generation and Shot-Level Control. MovieNet [24] provides large-scale movie annotations including scene boundaries, cinematic styles, and character metadata. Storyboard-driven approaches like VDS [44] and VAST [66] generate structured video via intermediate pose/layout representations. Multi-shot generation has been addressed through systems like VideoGen-of-Thought [68] and MovieAgent [62], which use hierarchical reasoning to plan and synthesize sequences. However, these efforts overlook the fine-grained structure of a single shot which is core unit of cinematic composition. They often emphasize isolated factors, without modeling the full set of interdependent creative controls. In contrast, SCINE introduces a unified taxonomy that captures the complete spectrum of shot-level cinematic elements.

3 Stable Cinemetrics

The subsequent sections detail our proposed taxonomy (Section 3.1), and its underlying design principles. Next, we develop our benchmark comprising of prompts designed for professional use (Section 3.2). Section 3.3 describes how prompt categorization and question generation are performed, enabling large-scale evaluations of video generative models for professional use.

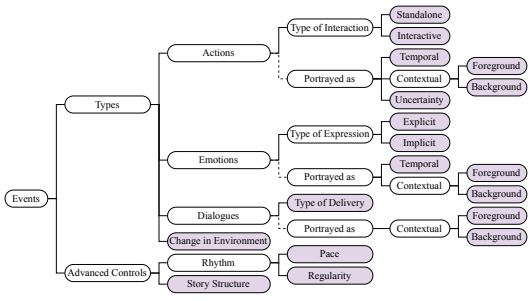
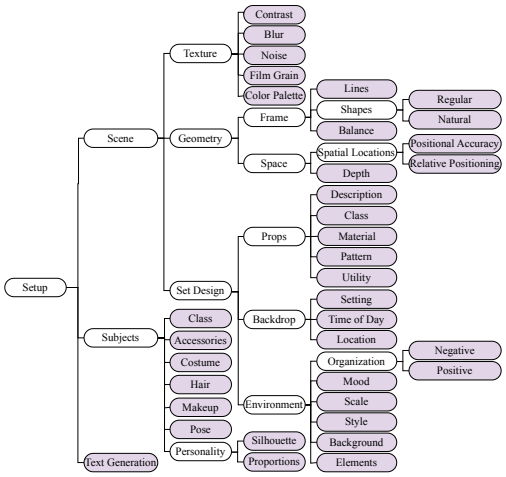


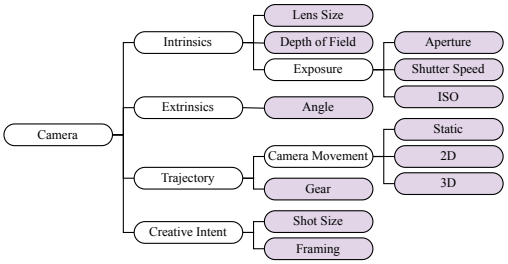
Figure 3: The **Events** taxonomy captures the narrative dimension of a shot which includes actions, emotions, and their fine grained portrayal as they evolve over time within a shot.

3.1 Taxonomy Design

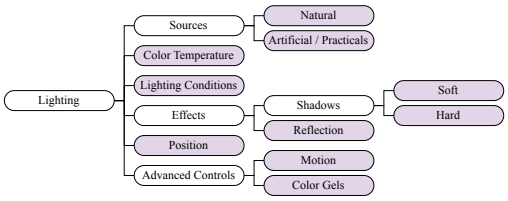
Our taxonomies are developed in iteration with industry professionals, including organizations established by the Big Five Studios [47], independent cinematographers and screenwriters, and an Academy Award winning Visual Effects



(a) The **Setup** taxonomy outlines the visual components within the frame, including subjects, props, and environmental context.



(b) The **Camera** taxonomy defines all controls related to camera configuration during a shot setup.



(c) The **Lighting** taxonomy specifies the illumination of shot, through light sources, their properties, and their interaction with the scene.

Figure 2: **Setup**, **Camera** and **Lighting** Taxonomies that structure the visual elements of a shot.

Artist. The central guiding question in our taxonomy development was: "*What **controls** do **professionals** require when setting up a **shot**?*".

A **shot** is the atomic unit of filmmaking; an uninterrupted sequence without cuts in which cinematic meaning emerges from the coordination of multiple cinematic choices. The average shot length (ASL) in feature films is 5–10 seconds [8], which closely aligns with the temporal limits of current video generation models. This duration is not a limitation; it is a compact canvas where rich and *enough* narrative and visual complexity can unfold. A single shot entails numerous controls to convey intent, emotion, and story. This motivates our decision to *design the taxonomy at the shot level*, where fine-grained control is paramount. **Control** is the key distinction between casual and professional video creation. In casual settings, users often accept model outputs with minimal intervention, delegating the *key* creative decisions to the model. In contrast, professional use demands precise, deliberate control at every stage of the generative process. In fact, pixel-level control is common in film-making [29, 61], underscoring the importance of fine-grained adjustments in achieving the desired visual effect. **Professionals** such as cinematographers and directors are primarily responsible for defining a shot's creative intent. While filmmaking is collaborative, these roles offer a practical abstraction for modeling control. A key insight is that, despite overlap, they are sufficiently *disentangled* to support distinct control dimensions: screenwriters rarely specify lighting or camera movement, and production designers are typically not associated with emotional tone or narrative pacing. These factors motivate the development of our 4 control pillars, each contributing to the composition of a shot. Our taxonomies are structured as hierarchical trees, where `leaf` nodes correspond to the most granular control parameters, each associated with a set of values. We describe each pillar below:

Setup. Setup (Figure 2a) encompasses all visible elements within the frame. We organize it into three top-level groups: (1) `Scene` aggregates environmental controls, including `Texture`, `Geometry`, and `Set Design`. `Texture` covers aspects such as `contrast` and `color palette`, which govern the surface feel of the shot. `Geometry` captures dominant `shapes` and the `spatial arrangement` of elements within the scene. `Set Design` comprises `Props` and their `attributes`; the `Backdrop`, which establishes the macro context of the set; and `Environment`, which defines micro-level elements contributing to the "feel" of the shot. (2) `Subjects` refer to the focal characters within a shot, defined by attributes such as `costumes` and `accessories`. (3) `Text Generation` refers to on-screen typography such as titles or lettering, designed to appear as integrated graphical elements. Each node in Setup has cinematic meaning: a dawn (`Time of Day`) setup combined with mist (`Elements`) can suggest danger, while a clean (`Organization`), symmetrical hallway (`Balance`) conveys order.

Table 1: Structured prompt upsampling with SCINE taxonomies. We show how control nodes from our taxonomies enable the generation of visually expressive (*SCINE Visuals*) prompts from narrative scripts (*SCINE Scripts*). The table also demonstrates how a single script can yield multiple visual interpretations, enabled by our taxonomy guided prompt generation pipeline. This aligns with filmmaking principles, where a script can be visually realized in diverse ways depending on the creative choices made by the filmmakers.

Baseline script: *A man serves dinner to his family.*		
Taxonomy Branch	Baseline choice → narrative impact	Alternative choice → narrative impact
`Depth of Field` (Camera)	Shallow → isolates food/serving hand, romantic warmth	Deep → every family member equally sharp, ensemble clarity
`Camera Movement` (Camera)	Static tripod + gentle dolly-in → calm focus on gesture; subtle emphasis	Handheld tracking → urgency, energetic family chaos
`Lighting Source` (Lighting)	Warm tungsten practicals → cozy, inviting domestic glow	Cool morning daylight through windows → brisk freshness and emotional distance
`Backdrop / Time of Day` (Setup)	Evening interior → nostalgic comfort, winding-down mood	Bright morning interior → optimism and upbeat tempo
`Props` (Setup)	Earth-tone wooden utensils → homely	Silver cutlery → formal, upscale
Upsampled prompt: *A man serves dinner to his family with shallow depth of field on a static tripod with a gentle dolly-in, under warm tungsten interior lighting in the evening, in a cozy earth-tone kitchen with wooden utensils.*		

Lighting. "*Lighting is the key to turning amateur footage into professional stories and presentation*" - Jay Holben [21]. Motivated by this principle, we define the following groups for Lighting (Figure 2b): (1) `Source`, the origin of illumination within the shot; (2) `Color Temperature`, which controls the warmth of the light; (3) `Lighting Conditions`, preset configurations describing scene-wide illumination; (4) `Effects`, visual outcomes resulting from light interacting with the scene; (5) `Position`, the spatial relation of the light source to the subject; and (6) `Advanced Controls` such as `flickering` modulation and the use of `color gels` to adjust lighting hue. Each control knob corresponds to distinct cinematic expressions: a shot with only a back light (`Position`) evokes mystery, while hard `shadows` are often used to amplify tension.

Camera. The camera taxonomy (Figure 2c) encompasses all camera-related control dimensions involved in a shot. We organize these into 4 high-level groups: (1) `Intrinsics`: optical and exposure parameters governing the light captured by the camera; (2) `Extrinsics`: position and orientation of the camera relative to the subject; (3) `Trajectory`: motion of the camera and the supporting gear that enables it; and (4) `Creative Intent`: compositional choices that shape the narrative or emotional tone of a shot. Prior works have primarily focused on camera motion control [19, 22]; however, we show that a much broader range of camera parameters can be independently manipulated while setting up a shot. Each parameter has tangible cinematic impact; for example, a shallow `depth of field` can isolate the subject from the background to direct emotional focus while, an insert `framing` spotlights narrative details with precision.

Events. Events (Figure 3) encodes the narrative substance of a shot, namely the depicted `actions`, `emotions`, and `dialogues` - which are further decomposed into dependent nodes for fine-grained control. These dependent nodes represent attributes that cannot exist independently of their parent categories; for instance, these nodes can specify the type of interaction or the delivery mode of a dialogue. Emotions may appear `explicitly` (visible tears) or `implicitly` (a clenched jaw), while actions can be `stand-alone` or `interactive`. The `Portrayed As` category captures aspects such as: `Temporal`, which refers to the unfolding pattern of the event (e.g., laughter erupting `simultaneously` vs. `sequentially`), and `Contextual`, which indicates whether the event occurs in the `foreground` or `background`. `Advanced Controls` refines pacing and the story structure of the shot, such as a turning point or climax. While recent works [59, 16] evaluate T2V models on sequential event generation, we show that, from a professional standpoint, Events encompass a much broader and richer evaluative space.

The taxonomies define a total of 76 leaf-level controls that can be independently adjusted when crafting a shot. We structure our taxonomies as hierarchical trees to enable disentanglement and multi-level abstraction of cinematic controls. Attributes within each branch are highly correlated, while branches remain independent, ensuring, for example, adjusting `Depth of Field` does not affect `Camera Movement`. The tree structure naturally supports multi-level abstraction, aligning with how filmmakers conceptualize scenes, starting from high-level intent and refining toward specific implementation. For example, a directive such as "set a tense alley at night" can be decomposed

into an EXT `Setting`, a cool `Color Temperature` and a Deep `Depth of Field`. This structure also allows easier scalability; adding a new detail like floating ember sparks, fits cleanly under (`Environment → Elements`) without disrupting the rest of the taxonomy. An alternate structure such as a flat, linear list would not support such extensibility.

Developing the taxonomy is a non-trivial task. This required multiple iterations with experts since professionals interpret and prioritize controls differently. Furthermore, shot creation is a multi-stage process, starting from script-writing to setup design to camera blocking, making unification under a single structured framework challenging. While some taxonomies [40] focus only on filmmaking aspects, they lack structure and are not aligned with generative modeling. Our goal instead was to impose structure, and develope a taxonomy that bridges professional filmmaking and generative video models. Details of each `control` node and their corresponding values are provided in Appendix A.1.

3.2 Designing Prompts for Professional Use

The taxonomies form the foundation for constructing prompts tailored to professional use. Our core approach in creating prompts involves sampling values from the `control` nodes and creating prompts that reflect realistic cinematic intent. Mirroring the filmmaking process, we first generate narrative scripts and then inject visual elements into these scripts :

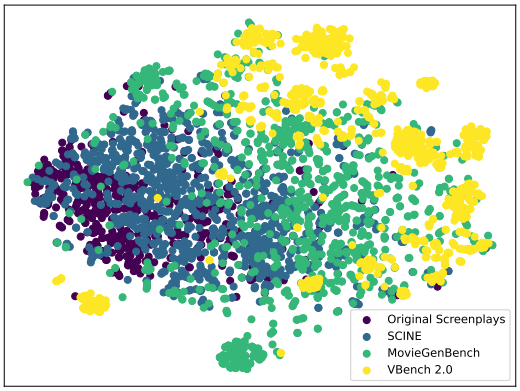


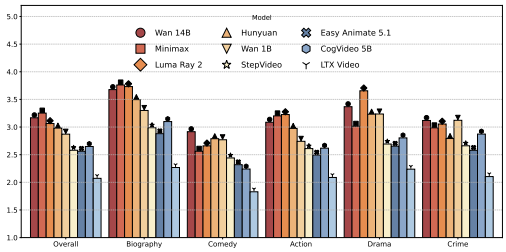
Figure 4: t-SNE visualization showing substantial overlap between ground truth screenplays and prompts in *SCINE Scripts*, in comparison to existing prompt based benchmarks such as VBench-2.0

Scripts. These prompts, referred to as *SCINE-Scripts*, contain the narrative content of individual shots. We collaborate with a professional screenwriter to create seed prompts that meet strict constraints: a single shot, under 10 seconds, with no reliance on off-screen elements. These seed prompts, along with sampled nodes from the Events taxonomy are provided as input to a LLM, for prompt generation. We use the Events taxonomy for these prompts because it directly encodes narrative beats i.e., what happens in a shot. It covers nodes such as physical dynamics (`actions`) and verbal interactions (`dialogues`), which are crucial to the story conveyed in a shot. To ensure prompt diversity, we vary parameters across multiple LLM invocations, sampling emotions from Plutchik's model [42], alternating actions, dialogue structures, genres, and subject composition. We use LLMs, as prior work [39, 50] have shown their effectiveness in screenplay generation. t-SNE visualization (Figure 4) of *SCINE-Scripts* embeddings [57] shows substantial overlap with ground-truth screenplays [45], whereas prompts from VBench-2.0 [67] and MovieGenBench [11] exhibit minimal overlap.

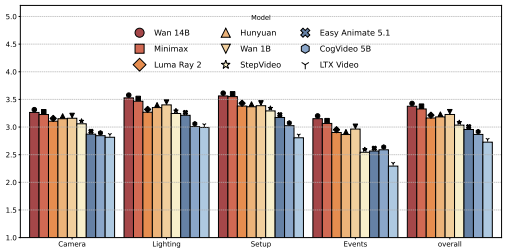
Visual Exposition. We refer to these set of prompts as *SCINE-Visuals*, which enrich *SCINE-Scripts* with visual elements from the Camera, Lighting, and Setup taxonomies. In contrast to *Events*, these taxonomies offer fine-grained control over the visual style and composition of a shot. For each base prompt in *SCINE-Scripts*, we sample values from one or more control nodes and inject them to expand the prompt with structured visual specifications. *SCINE-Visuals* highlight a key advantage of our taxonomy: structured prompt upsampling. Unlike existing prompt upsampling techniques that delegate all creative decisions to the LLM, our method constrains generation within the taxonomy, enabling more controlled and interpretable prompt expansion. Table 1 provides a breakdown of how *SCINE-Scripts* is subsequently upsampled via the taxonomy to generate *SCINE-Visuals*.

3.3 Category and Question Generation

Next, we extract *categories* and generate *questions* for each prompt. Categories link prompts back to the taxonomy, allowing fine-grained evaluations across different levels of abstraction; a single prompt can map to multiple categories across different taxonomies. For each category, we generate targeted questions that are shown to human annotators during video evaluations. These questions are specific



(a) Overall results on *SCINE Scripts*, across genres. Minimax and Wan 14B emerge as the strongest models. The strongest genre is Biography while all models consistently struggle at Comedy.



(b) Overall results on *SCINE Visuals* across four pillars of professional control. *Events* emerges as the most challenging category across models, while *Setup* yields the strongest performance.

Figure 5: Overall results on SCINE *Scripts* and *Visuals*.

in nature and target only a single control node, enabling its isolated evaluation. Unlike high-level prompt adherence questions, which lack fine-grained attribution, our framework supports per-control annotations. A minimal example is shown below :

Prompt: A tight close-up focuses on a fireplace, its embers flickering brightly.

- **Category** : Lighting → Advanced Controls → Motion | **Question** : Does the scene exhibit dynamic flickering effects in its lighting that align with the description?
- **Category** : Camera → Creative Intent → Shot Size | **Question** : Does the video include a tight close-up shot that captures the detailed framing?

Additional details are presented in Appendix A.3.

4 Are current Video Generative Models Ready for Professional Use?

We now evaluate state of the art text-to-video (T2V) models against the professional standards defined in our taxonomy pillars. Our analysis reveals both strengths and persistent challenges of current models, offering an overview of how current models align with professional quality expectations.

4.1 Experimental Setup

Prompts. The SCINE benchmark comprises two prompt categories, *Scripts* and *Visuals*, each aligned with distinct professional roles (Table 2). The *Visuals* prompts are created by systematically upsampling *Scripts* using our taxonomies leading to a total of 2,089 prompts. We categorize prompts by difficulty: we create basic prompts by limiting the number of sampled control nodes, while in advanced prompts, we do not impose any restriction.

Models. We evaluate 13 state-of-the-art T2V models, both open-source (WAN 1B/14B [56], Hunyuan-Video [14], Step Video [12], CogVideoX 5B [65], LTX-Video [18], Pyramid Flow [27], Easy Animate 5.1 [63], Mochi [52]), VChitect-2.0 [15] and closed source (Minimax [38], Luma Ray 2 [30], Pika 2.2 [31]). Our goal is to assess each model's suitability for role-specific professional tasks, like evaluating narrative fidelity from a screenwriter's perspective. Unless otherwise noted, we use default sampling parameters and maintain a consistent seed per prompt, across models for fair comparison.

Human Annotation Setup. To ensure high-quality evaluation, we work with a pool of 84 expert annotators with an average of 6.5 years of experience in film production, across roles such as cinematographers, film editors, screenwriters, visual communication designers, and directors. Annotators were shown a prompt along with two generated video samples. For each prompt, they were presented with taxonomy derived evaluation categories and

Table 2: The **SCINE** benchmark includes prompts tailored to professional roles, where each prompt is paired with multiple, fine grained evaluation questions.

Target Role	Target Taxonomies	# of Prompts	Avg. Questions per Prompt
	SCINE-Scripts		
Screenwriters	Events	1133	2.57 ± 0.98
	SCINE-Visuals		
Cinematographers	Camera, Lighting	355	5.42 ± 4.47
Production Designers	Setup	298	4.00 ± 3.35
Directors	All	303	10.48 ± 5.07

corresponding questions. Each video was rated independently on a 1–5 scale, where 1 indicated complete misalignment with the category and 5 indicated a perfect match. Although the evaluation was non-comparative, our UX ablation studies showed that displaying two videos side by side improved annotator calibration, especially when selecting middle range scores. To promote consistency and reduce subjectivity, we developed a comprehensive annotation guide covering each control node in the taxonomy. We collect 3 votes for every video-question pair across **13,457** unique questions, collecting a total of **248,536** pairwise annotations. We observe an intra-class correlation coefficient (ICC) [51] of 80.4% for the 1-5 ratings at the model-pair level, and 95.5% when the models are considered individually. Further, we conduct Wilcoxon signed-rank tests [60] across 45 model pairs where we observe statistically significant preferences in 37 of them, which highlights that the annotators agree on model preferences.

4.2 Results and Analysis

SCINE Scripts We first evaluate models on narrative event generation, i.e. the *story* dimension of a shot. Figure 5a compares model performance across different genres, focusing on the accuracy and coherence of generated events. Minimax and WAN-14B emerge as the overall top performers, while LTX-Video consistently underperforms. We observe that models generally perform better on the Biography genre, whereas Comedy proves challenging. In Figure 6, we zoom into sub-categories within *Events* $\rightarrow$ *Types*. Minimax leads in nearly all categories, showing the largest margins in *Dialogues* and *Change in Environment*, but falls short in *Advanced Controls*, where WAN-14B outperforms. Further, models are better at stand-alone actions compared to interactive actions and portray implicit emotions better than explicit ones. Within Event Types, *Actions* exhibit lowest variance across models. Performance on Causal and Sequential events is highly correlated ($\rho = 0.94$), as is performance on Concurrent and Overlapping events ($\rho = 0.86$) across models. Despite variation across models, all show limitations in multiple *Event* aspects, highlighting opportunities for improvement.

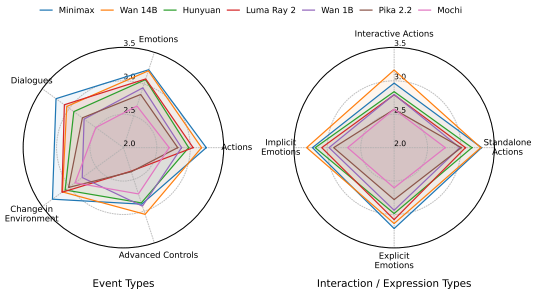


Figure 6: Fine-grained evaluation on **Events**. Models handle environmental changes well but struggle with dialogues and shot pacing. Standalone actions outperform interactive, and implicit emotions are easier than explicit.

SCINE Visuals Figure 5b demonstrates overall results on SCINE Visuals. Consistent with the pair-wise rankings in Figure 1, WAN-14B and Minimax emerge as top performers across all pillars. We find that current models struggle most with *Events* and *Camera*, while elements of *Setup* and *Lighting* are comparatively easier to capture. Only the top three models- WAN-14B, WAN-1B, and Minimax - reliably depict *Events*, with a substantial performance gap from the rest. While *Camera* scores are low across the board, the narrow spread suggests that all models face similar limitations. *Lighting* shows the most consistent performance, with most models achieving relatively high scores whereas *Setup* yields the highest absolute scores for the top-performing models.

Cinematographer. We evaluate this role by creating prompts that inject control nodes from the *Camera* and *Lighting* taxonomy. Within *Camera*, *Extrinsics* and *Trajectory* have the lowest average performance and the narrowest inter-model spread. For *Lighting*, the primary bottleneck is *Lighting Position*. We further present results, split by prompt difficulty in Figure 7. Across all models, performance degrades on advanced prompts, indicating that under conditions resembling professional workflows where cinematographers have a large amount of control, current models struggle. The biggest performance drops occur in *Lighting Source*, *Color Temperature*, and *Creative Intent*. *Lighting Position* and

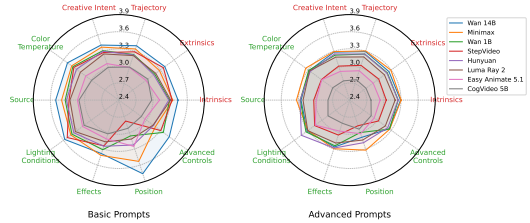
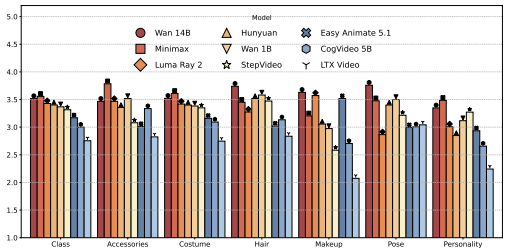
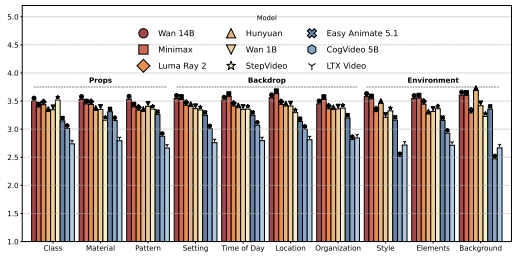


Figure 7: Split results on basic vs. advanced prompts for **Camera** and **Lighting**. All models show performance drops on advanced prompts, with the largest decline in Lighting Source.



(a) Fine-grained results on **Subjects**. Models perform well on hair and accessories, but struggle with personality and makeup.



(b) Fine-grained results on **Set Design**. Models perform better at Backdrop in comparison to Environment, and struggle most with styling the shot appropriately.

Figure 8: Fine-grained results on **Setup** across Subjects and Set Design.

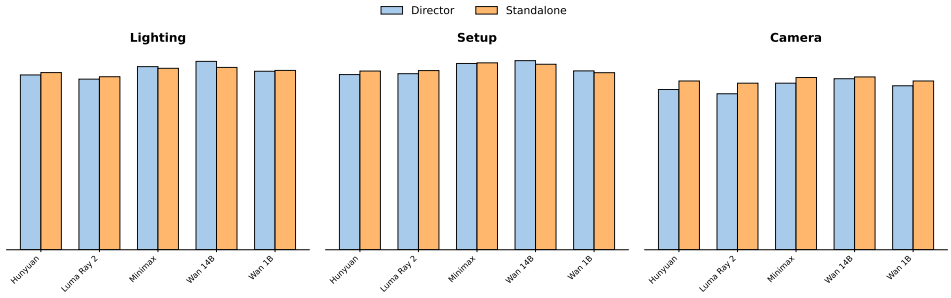


Figure 9: **Director** results: Joint specification of all controls, mirroring real-world shot creation leads to a performance drop on all models, compared to evaluation in a standalone manner.

Advanced Controls show the smallest performance drops, but remain the weakest categories overall, highlighting persistent challenges regardless of prompt complexity. Hunyuan and WAN-1B exhibit consistent performance across complexity levels.

Production Designer. Models perform strongest on the *Setup* pillar; within *Setup*, models achieve comparable performance on *Subject* and *Scene* generation, but show a drop in *Text Generation*. In *Subjects* (Figure 8a), model performances largely vary, with the highest scores for *Hair* and *Accessories*, and the weakest in *Personality* and *Make-up*. In *Set Design* (Figure 8b), performance trends follow the order: *Backdrop* >*Props* >*Environment*. Within *Props*, models perform well in *Material* but struggle at generating intricate *Patterns*. For *Environment*, current limitations lie in adhering to a coherent *Style* and *Backgrounds*, whereas better performance is seen in organizing *Space* within a frame.

Director. Prompts targeting this role differ from prior categories as they evaluate models across all taxonomies simultaneously. Model performance declines on average when all controls are defined jointly (Figure 9). The largest performance drop is observed in *Camera*, followed by *Setup* and *Lighting*. Wan 14B is the only model to show improved performance on *Director* prompts for Lighting and Setup, compared to its *Standalone* results.

Our evaluation identifies a three-tier hierarchy among current T2V models: Minimax and WAN-14B at the top, followed by Luma Ray 2, Hunyuan, and WAN-1B, with the remaining models forming the third tier. While overall performance varies, most models struggle with the fine-grained elements critical to professional video generation. For example, atomic events are handled reasonably well, but models falter on concurrent and causal events, which demand deeper temporal reasoning. Similarly, high-level cues like lighting conditions are better captured than nuanced aspects like precise light positioning. In summary, even top performing models exhibit substantial room for improvement across all dimensions of our taxonomy. No model achieves consistently strong performance across all aspects of shot composition, underscoring the challenge of aligning generative video models with professional standards. Additional results and analysis are presented in Appendix A.2.

5 Scalable Evaluation of Professional Videos

In the previous section, we evaluated video generative models using expert annotations across 76 control nodes defined by our taxonomy. While human evaluation remains the gold standard, it is costly and difficult to scale, and defining reliable automatic metrics for each control node is non-trivial. Recent advances in vision-language models (VLMs) [13, 33, 1] offer a scalable alternative, showing strong performance in video understanding tasks. In this section, we leverage these models to perform automatic evaluation of professional video generation.

Zero-shot VLM Evaluations. The rise of multimodal VLMs has enabled progress on vision-language tasks, including video understanding, making them natural candidates for evaluating professional video generation. We use expert annotations as ground truth and measure VLM alignment by prompting models with a video, its associated prompt, and a specific question tied to a taxonomy node, asking for a 1–5 rating similar to our user study. We explicitly instruct the VLM to ignore factors unrelated to the specified category when evaluating the video. We determine VLM preferences by independently scoring each video and selecting the higher-scoring one. This design mitigates hallucination and order-sensitivity issues commonly observed when prompting with both videos simultaneously. We use Qwen2.5-VL-Instruct models due to their strong video understanding capabilities. To study the effect of model scale, we evaluate 3 sizes: 7, 32 and 72B. Figure 10 shows that increasing model size does not yield significant improvements in alignment with human judgment. Our results reveal overall poor agreement, consistent with prior findings [20, 64] that highlight the need for fine-tuning on in-distribution, human-labeled data.

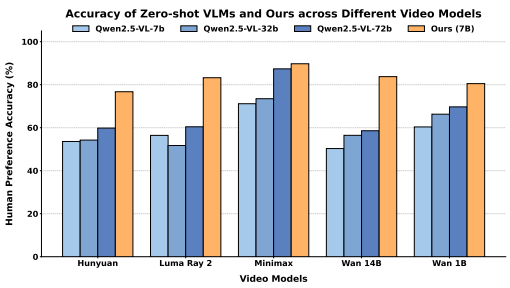


Figure 10: Our trained VLM shows consistent alignment with human annotations across video generation models, outperforming baselines, most notably on WAN-14B.

Aligning Human and VLM ratings. We adopt Qwen-2.5-VL-7B [53] as the base model for fine-tuning; our training and validation dataset consist of 44,062 and 12,763 samples, respectively. We aggregate annotator scores (3/video pair) into binary preferences, excluding ties. Videos are pre-processed at 2 fps at their native resolution. Each sample consists of a prompt, two videos, and a binary label. The trained model acts as a classifier; we modify the model architecture to output a scalar score using a linear projection over the final layer's last token. The model takes as input a single video, its prompt, and the evaluation question as input. For training, we use the Bradley–Terry objective [5] due to its sample efficiency over regression [34]. The model is trained for 1 epoch with a batch size of 8 and learning rate 6e-5. Similar to zero-shot evaluation, we compute pairwise preference accuracy against the average of the annotators as the target metric. Our fine-tuned model achieves an overall accuracy of 72.36%, outperforming all zero-shot VLM baselines. This represents an absolute improvement of $\sim$20%, over the baseline 7B model. Our model (Figure 10) shows consistent performance across videos generated by different models, highlighting its ability to generalize across different video qualities. Additional VLM results are presented in Appendix A.7.

6 Conclusion and Future Work

Stable Cinemetrics probes at the intersection of professional video generation and generative video models, grounding prompts, evaluations, and analysis in our structured taxonomies. Our findings reveal where current state-of-the-art models perform well, and where substantial improvements are needed. Our prompt suite offers a strong testbed for future video generative models and can be easily extended as models improve, owing to the flexibility of our taxonomy. We envision several extensions of our work; while our current focus is on evaluation, the taxonomy can also support analyzing video datasets for cinematic diversity or serve as a structure for video captioning. While today's text-to-video models are not yet usable in a fully zero-shot capacity, our findings identify the main challenging pillars for professional filmmaking, illuminating the need for potential solutions like fine-tuning and customization that can bring these models closer to real production use. We hope SCINE encourages deeper exploration at the intersection of filmmaking and video generative models, fostering closer collaboration between artists and models.

7 Acknowledgements

We thank Robert Legato, Hanno Basse and Heather Ferreira for their valuable input on our work. We are also grateful to the team at MovieLabs for their feedback on our taxonomies. A special thanks to Cedric Wagrez for his invaluable assistance with the human annotations!

8 Limitations

Although our taxonomy was developed in consultation with domain experts, it is limited by the scope of our collaborator network. Filmmaking terminology and interpretive nuance vary across regions and cultures, greater expert diversity would enable broader incorporation of global cinematic controls into the taxonomy. Some taxonomy nodes (e.g., Color Temperature, ISO) were abstracted for evaluation, as we found it difficult for annotators to consistently perceive fine-grained values (such as 2000K or ISO 800). Prompt generation is based on LLMs, whose proprietary nature and potential biases can influence the language and structure of the prompts. Our zero-shot VLM evaluations were bounded by compute and data resources, limiting the scale and scope of the experiments.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025. URL https://api.semanticscholar.org/CorpusID:276449796.

[2] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9D2QvO1uWj.

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

[4] Bruce Block. *The visual story: Creating the visual structure of film, TV, and digital media*. Routledge, 2020.

[5] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

[7] Blain Brown. *Cinematography: theory and practice: image making for cinematographers and directors*. Routledge, 2016.

[8] James E Cutting and Ayse Candan. Shot durations, shot classes, and the increased pace of popular movies, 2015.

[9] Google DeepMind. Veo 3 tech report. Technical report, DeepMind, 2025. Veo: a text-to-video generation system. Available at https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf.

[10] Edward Dmytryk, Andrew Lund, and Mick Hurbis-Cherrier. *On film editing: an introduction to the art of film construction*. Routledge, 2018.

[11] Adam Polyak et al. Movie gen: A cast of media foundation models, 2025. URL https://arxiv.org/abs/2410.13720.

[12] Guoqing Ma et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model, 2025. URL https://arxiv.org/abs/2502.10248.

[13] OpenAI Josh Achiam et al. Gpt-4 technical report. 2023. URL https://api.semanticscholar.org/CorpusID:257532815.

[14] Weijie Kong et al. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

[15] Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, Yi Wang, Yuming Jiang, Yaohui Wang, Peng Gao, Xinyuan Chen, Hengjie Li, Dahua Lin, Yu Qiao, and Ziwei Liu. Vchitect-2.0: Parallel transformer for scaling up video diffusion models, 2025. URL https://arxiv.org/abs/2501.08453.

[16] Weixi Feng, Jiachen Li, Michael Saxon, Tsu jui Fu, Wenhu Chen, and William Yang Wang. Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation, 2024. URL https://arxiv.org/abs/2406.08656.

[17] Syd Field. *Screenplay: The foundations of screenwriting*. Delta, 2005.

[18] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2024. URL https://arxiv.org/abs/2501.00103.

[19] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Z4evOUYrk7.

[20] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024.

[21] Jay Holben. *A Shot in the Dark: A Creative DIY Guide to Digital Video Lighting on (almost) No Budget*. Course Technology Press, 2011.

[22] Chen Hou and Zhibo Chen. Training-free camera control for video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KI1zldOFz9.

[23] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation, 2025. URL https://arxiv.org/abs/2307.06350.

[24] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer, 2020.

[25] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818, June 2024.

[26] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing, 2025. URL https://arxiv.org/abs/2503.07598.

[27] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling, 2025. URL https://arxiv.org/abs/2410.05954.

[28] Steven Douglas Katz. *Film directing shot by shot: visualizing from concept to screen*. Gulf Professional Publishing, 1991.

[29] MediaBee Color Lab. The art and impact of color grading in cinema, July 2024. URL https://www.mediabeecolorlab.com/post/the-art-and-impact-of-color-grading-in-cinema. Blog post.

[30] Luma Labs. Ray2, 2025. URL https://lumalabs.ai/ray. Accessed: 2025-05-02.

[31] Pika Labs. Pika 2.2, 2025. URL https://pikalabs.org/pika-2-2/. Accessed: 2025-05-02.

[32] Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 109790–109816. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c6483c8a68083af3383f91ee0dc6db95-Paper-Conference.pdf.

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023. URL https://api.semanticscholar.org/CorpusID:258179774.

[34] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

[35] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22139–22149, June 2024.

[36] Luma Labs. Luma ray2 video model. https://lumalabs.ai/ray, 2025. Accessed: 2025-04-23.

[37] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. Artificial intelligence index report 2025, 2025. URL https://arxiv.org/abs/2504.07139.

[38] MiniMax. Video generation api, 2024. URL https://www.minimax.io/news/video-generation-api. Accessed: 2025-05-02.

[39] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals, 2022. URL https://arxiv.org/abs/2209.14958.

[40] Motion Picture Laboratories. Ontology for media creation: Part 3f: Images. Technical Report v2.6, MovieLabs, May 2024. URL https://mc.movielabs.com/docs/ontology/assets-images/images/. Accessed: 2025-05-15.

[41] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=j7kdXSrISM.

[42] Robert Plutchik. *The emotions*. University Press of America, 1991.

[43] Michael Rabiger. *Directing: Film techniques and aesthetics*. Routledge, 2013.

[44] Anyi Rao, Xuekun Jiang, Yuwei Guo, Linning Xu, Lei Yang, Libiao Jin, Dahua Lin, and Bo Dai. Dynamic storyboard generation in an engine-based virtual environment for video production. In *ACM SIGGRAPH 2023 Posters*, pages 1–2. 2023.

[45] Rohit Saxena and Frank Keller. Moviesum: An abstractive summarization dataset for movie screenplays, 2024. URL https://arxiv.org/abs/2408.06281.

[46] Linda Seger and Edward Jay Whetmore. *From Script to Screen: The collaborative art of filmmaking*. Lone Eagle Publishing Company, LLC, 2004.

[47] StudioBinder. What are the major film studios?, n.d. URL https://www.studiobinder.com/blog/what-are-the-major-film-studios/. Accessed: 2025-05-14.

[48] 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, 2025. URL https://arxiv.org/abs/2407.14505.

[49] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.

[50] Yuying Tang, Haotian Li, Minghe Lan, Xiaojuan Ma, and Huamin Qu. Understanding screenwriters' practices, attitudes, and future expectations in human-ai co-creation, 2025. URL https://arxiv.org/abs/2502.16153.

[51] Judith M. Tanur, George Casella, Richard Dykstra, Mark P. Finster, Donald P. Gaver, Joel Greenhouse, Gudmund R. Iversen, guillermina Jasso, Jan Kmenta, S. James Press, Seymour Sudman, Luke Tierney, Jessica Utts, Katherine K. Wallman, Stanley Wasserman, and Mark Watson. *Journal of the American Statistical Association*, 84(407):830–834, 1989. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2289675.

[52] Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024.

[53] Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

[54] Veo-Team. Veo 2. 2024. URL https://deepmind.google/technologies/veo/veo-2/.

[55] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion, 2024. URL https://arxiv.org/abs/2403.12008.

[56] Team Wan. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

[57] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report, 2024. URL https://arxiv.org/abs/2402.05672.

[58] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation, 2024. URL https://openreview.net/forum?id=dUDwK38MVC.

[59] Yiping Wang, Xuehai He, Kuan Wang, Luyao Ma, Jianwei Yang, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Is your world simulator a good story presenter? a consecutive events-based benchmark for future long video generation, 2024. URL https://arxiv.org/abs/2412.16211.

[60] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer, 1992.

[61] Aylish Wood. Pixel visions: Digital intermediates and micromanipulations of the image. *Film Criticism*, 32(1):72–94, 2007.

[62] Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent cot planning, 2025. URL https://arxiv.org/abs/2503.07314.

[63] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture, 2024. URL https://arxiv.org/abs/2405.18991.

[64] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024.

[65] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

[66] Chi Zhang, Yuanzhi Liang, Xi Qiu, Fangqiu Yi, and Xuelong Li. Vast 1.0: A unified framework for controllable and consistent video generation, 2024. URL https://arxiv.org/abs/2412.16677.

[67] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness, 2025. URL https://arxiv.org/abs/2503.21755.

[68] Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. Videogen-of-thought: Step-by-step generating multi-shot video with minimal manual intervention, 2025. URL https://arxiv.org/abs/2412.02259.

[69] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robo-Dreamer: Learning compositional world models for robot imagination. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 61885–61896. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhou24f.html.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately capture the paper's contribution and scope by clearly stating the claims made and outlining primary contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss the limitations of our current work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not present theoretical results, assumptions, or proofs. Instead, it proposes a taxonomy for professional movie elements and provides statistical analysis on how current state-of-the-art text-to-video (T2V) models perform across different pillars of the taxonomy.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: The paper provides all necessary information to reproduce the main experimental results. Details for creating the taxonomy are available in Section 3. The setup for VLM evaluation is described in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release our proposed taxonomy upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides all necessary information to reproduce the main experimental results. Details for creating the taxonomy are available in Section 3. The setup for VLM evaluation is described in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report statistically significant annotator agreements within the annotators. Additional details are present in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Yes, we report these numbers in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The paper complies with NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Yes, we report the impacts of our current work in the Supplementary.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not anticipate any high risk misuse of our current work.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite and credit papers that introduce the pre-trained models we train on, as well as the LLM used for generating prompts, and the metrics we use to compare across models.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We will release our taxonomy which is well documented.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We discuss the full text of instructions given to participants and include screenshots in the Supplementary.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: The user study involves participants score/rank videos generated with a predetermined list of prompts and quetions in the professional pillars. As such, we believe the study does not pose any risk or harm to participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We provide information on using LLMs in helping us creating prompt set and questions in Section 3.2 and 3.3.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Appendix

A.1 Taxonomy Details

We provide additional details on control nodes and their associated values in the taxonomies. Some nodes accept open-ended values, for example, a range of stand-alone actions. To simplify evaluation, we abstract certain values that can be fine-grained in future works. For instance, we group `Aperture` into wide/medium/narrow, though exact f-stop values can also be studied in the future. Similarly, `color palette` is treated as a discrete value in our current work, but can be decomposed into hue, brightness, and saturation. Table 3 - 6 details the control nodes and their values of the Camera, Lighting, Setup and Events Taxonomies, respectively.

Table 3: **Camera Taxonomy** Control Nodes and Values

Name	Description	Potential Values
Lens Size	Defines the focal length and field of view of the camera lens.	Standard, Fisheye, Wide, Medium, Long Lens, Telephoto
Depth of Field	Controls the range of focus in the image, affecting subject isolation.	Deep, Shallow, Soft, Rack, Split Diopter, Tilt Shift
Aperture	The camera lens opening that controls the amount of light propagated through the camera.	Wide, Medium, Narrow
Shutter Speed	The duration for which the camera sensor is exposed to light.	Slow, Medium, Fast
ISO	Sensitivity of the camera sensor to light.	Low, Medium, High
Angle	Defines the camera's viewpoint in relation to the subject.	Low, High, Aerial, Overhead, Dutch, Eye-level, Shoulder, Hip, Knee, Ground, Continuous Values
Static	A fixed camera position without any movement.	-
2D	Camera movements restricted to horizontal or vertical axes.	Pan left, Pan right, Tilt up, Tilt down, Zoom in, Zoom out
3D	Camera movements that incorporate spatial depth and multi-axis motion.	Push In, Pull Out, Dolly Zoom, Camera Roll, Tracking, Trucking, Arc, Crane
Gear	Specifies the support systems and stabilization equipment used to facilitate camera movement.	Handheld, Tripod, Pedestal, Cranes, Overhead Rigs, Dolly, Stabilizer, Snorricam, Vehicle Mount, Drones, Motion Control, Steadicam
Shot Size	Determines how much of the subject and surroundings are visible in the frame.	Establishing, Master, Wide, Full, Medium–Full, Medium, Medium–Close-up, Close-up, Extreme Close-up
Framing	Placements and composition of subjects within the frame.	Single, Two Shot, Crowd, OTS, PoV, Insert

Table 4: **Lighting Taxonomy** Control Nodes and Values

Name	Description	Potential Values
Natural Light	Natural sources of light, such as sunlight, moonlight, or firelight.	Sunlight, Moonlight, Firelight
Artificial/Practicals Light	Man-made light sources that illuminate the scene.	LED, HMI, Tungsten, Fluorescent, HID
Color Temperature	Defines the hue of the light, typically measured in Kelvin, affecting the scene's mood.	Warm, Cool, Cold
Lighting Conditions	Describes various lighting scenarios or ambient conditions present in a scene.	Candlelight, Golden Hour, White Fluorescent, Clear Daylight, Overcast
Soft Shadows	Subtle and diffused shadows resulting from indirect or scattered light.	Diffused Light, High Key Lighting, Reflectors
Hard Shadows	Sharp, well-defined shadows generated by a direct light source.	Direct Light, Low Key Lighting
Reflection	The effect of light bouncing off surfaces to create a reflective appearance.	-
Lighting Position	Specifies the placement or direction of the light source relative to the subject.	Back Light, Fill Light, Top Light, Side Light, Key Light
Motion	Dynamic changes or movement in the lighting effect.	Flickering, Pulsing
Color Gels	Colored filters applied to lights to modify or enhance the color of the illumination.	-

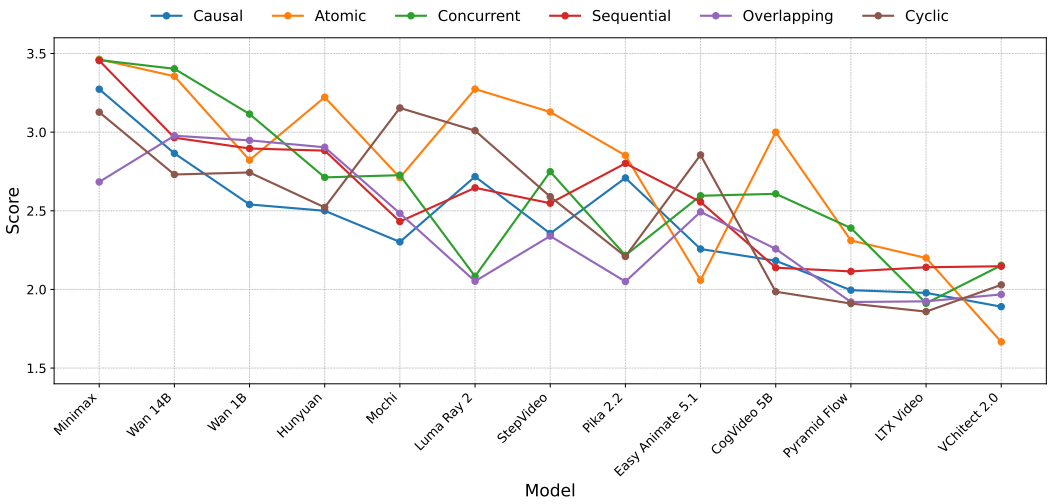


Figure 11: **Model performance on Events across temporal portryal of Actions**. Atomic actions are handled well, whereas models struggle with causal and overlapping Events.

A.2 Additional Results and Analysis

A.2.1 Events

Figure 11 shows Events performance across 13 models and 6 Temporal portrayal of Actions. Models handle atomic and concurrent actions well, but struggle with causal, overlapping, and cyclic events. Figure 12 shows Events performance across 13 models and 12 genres. Biography and Adventure are strongest whereas Comedy and Horror are the weakest. Minimax leads in 6/12 genres, Luma Ray 2 tops Action and Drama, and WAN-14B is the most consistent, with the lowest standard deviation.

Figure 13 shows performance of 10 open-source models across 19 emotion classes. Models perform best on remorse and ecstasy, but fare poorly on aggressiveness and rage. As shown in Figure 14, dialogue performance is weaker in comparison to Emotions and Actions. Models particularly struggle with multi-turn dialogues or when non-verbal reactions follow. Since T2V models do not generate audio, we evaluate whether the correct character delivers the line and/or with appropriate visual expression. Most models fail to localize the speaker, often attributing a single dialogue to multiple characters.

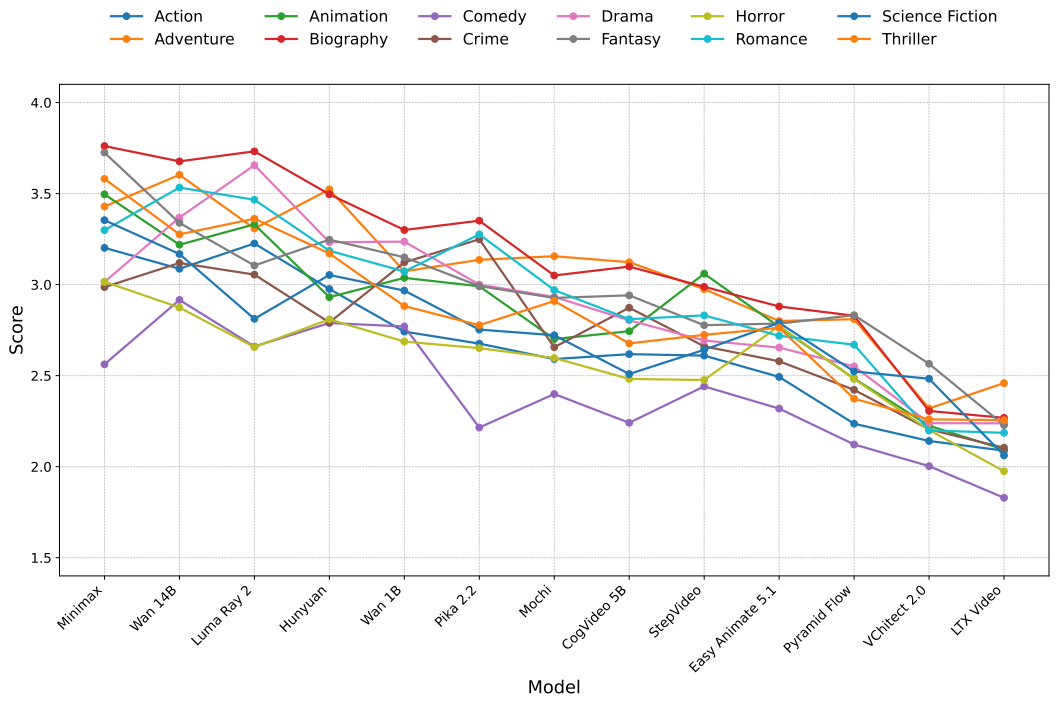


Figure 12: **Model performance on Events across genres**. Across 13 models and 12 genres, portrayal of Events in Biography and Adventure are the strongest, while Comedy and Horror are the weakest.

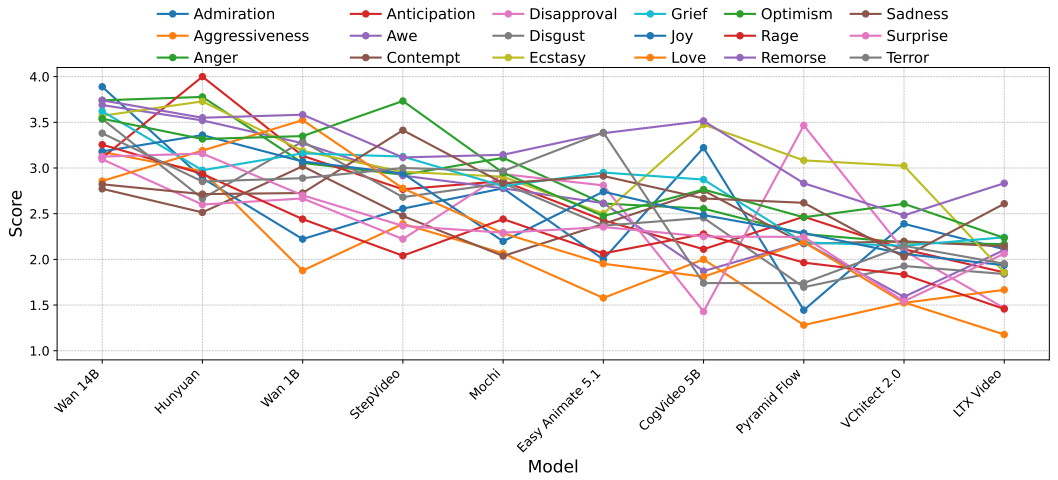


Figure 13: **Model performance on Emotions**. Among 10 models and 19 emotions, Remorse is best portrayed, while Aggressiveness is the weakest.

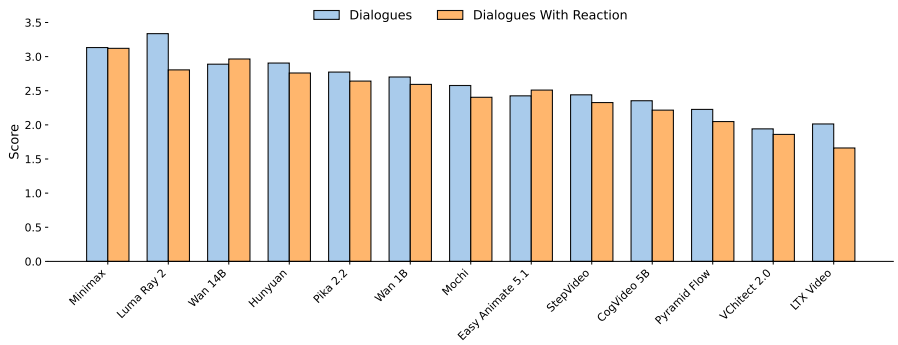


Figure 14: **Model performance on Dialogues**. Compared to Actions and Emotions, models struggle at Dialogues. Within Dialogues, performance drop is seen during multi-turn conversations.

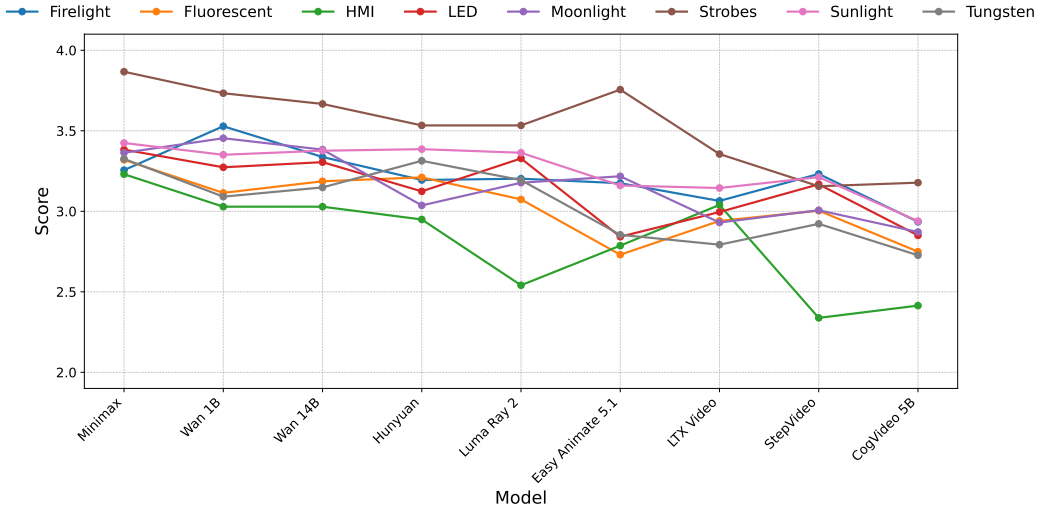


Figure 15: **Model performances across Lighting Source**. Strobes and Sunlight emerge strongest, whereas HMI and Fluorescent are points of weaknesses.

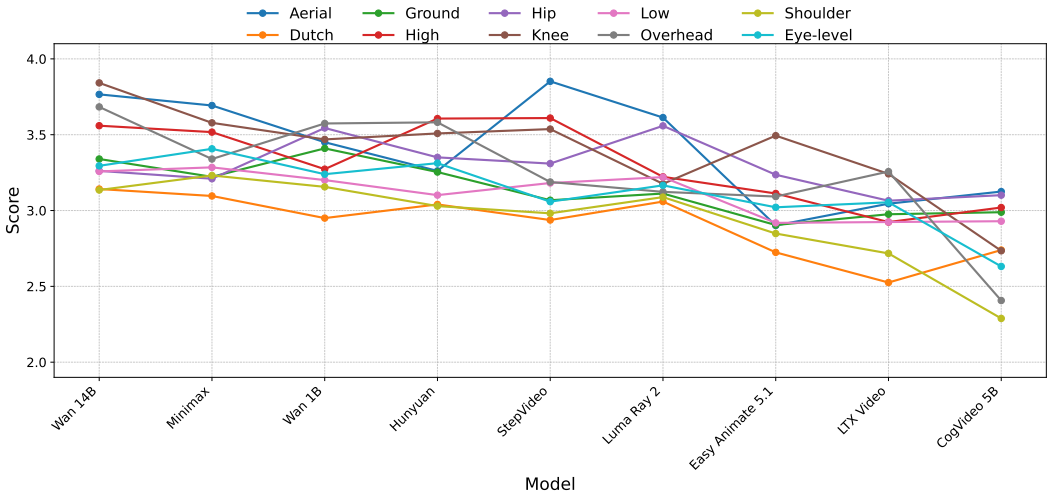


Figure 16: **Model performances across Camera Angles**. The Dutch angle poses a common challenge to all current video generative models

Table 5: **Setup Taxonomy** Control Nodes and Values

Name	**Description**	**Potential Values**
Contrast	Determines the difference between light and dark areas to enhance visual impact.	Low, High
Blur	Introduces softness to parts of the image to guide focus or create mood.	Gaussian, Radial, Motion
Noise	Adds random variations in brightness or color, mimicking film grain or digital sensor noise.	Gaussian, Salt and Pepper, Poisson
Film Grain	Emulates the granular texture of traditional film photography for a classic look.	-
Color Palette	Defines the overall range and harmony of colors in the scene, influencing its mood.	*Open Set*
Lines	Directional elements that guide the viewer's gaze within the shot.	Horizontal, Vertical, Diagonal
Regular Shapes	Structured, geometric forms such as squares, circles, and triangles that add order to the design.	Square, Circle, Triangle
Natural Shapes	Unstructured shapes that naturally emerge in the scene, without any geometric constraints.	Water-like, Cloud-like
Frame Balance	Refers to the distribution of visual weight across the composition, ensuring a harmonious layout.	Rule of Thirds, Symmetry, Right Heavy, Left Heavy
Positional Accuracy	The absolute position of an object or a subject in a scene.	*Open Set*
Relative Positioning	The relative positioning of an object in relationship to other objects in the scene.	*Open Set*
Depth	Controls the perception of distance between elements, enhancing the three-dimensional feel of the scene.	Deep, Flat, Limited, Ambiguous
Setting	Defines if the scene is happening indoors or outdoors.	INT/EXT
Time of Day	The time of day the scene is set in.	DAY, NIGHT, MORNING, EVENING, DAWN, DUSK, LATE NIGHT, MIDDAY, SUNRISE, SUNSET, AFTERNOON
Location	The specific place or setting of the scene.	*Open Set*
Negative Space	Defines if there a lot of empty space.	-
Positive	Defines how the space is occupied in the environment.	Clean, Cluttered
Mood	The emotional atmosphere or feeling created by the environment.	*Open Set*
Scale	The relative size or extent of the environment.	*Open Set*
Style	The artistic or visual style of the backdrop.	*Open Set*
Background	The part of the scene that is behind the main subject and does not need to be exactly described.	*Open Set*
Elements	The natural or artificial components of the backdrop.	Rain, Snow, Fog, Wind, Thunder, Smoke, Dust, Ash, Fire
Prop Description	A general description of the prop.	*Open Set*
Prop Class	The category or type of the prop.	*Open Set*
Prop Material	The substance(s) the prop is made of.	Wood, Glass, Gold, Paper, Plastic
Prop Pattern	The design on the prop.	Grid, Checker, Stripes, Zigzag, Dots, Bricks, Metal, Hexagons
Prop Utility	The purpose or function of the prop, whether it just exists in the scene or will it be used by the subject.	Decorative, Functional
Subject Class	The category or type of the subjects.	*Open Set*
Subject Accessories	Items worn or carried by the subjects that enhance their appearance or functionality.	*Open Set*
Subject Costume	The clothing worn by the subjects, especially for a performance or to create a specific character.	*Open Set*
Subject Hair	The style and appearance of the subjects' hair.	*Open Set*
Subject Makeup	Cosmetics applied to the subjects' face or body to enhance or alter their appearance.	*Open Set*
Subject Pose	The position or stance of the subjects, especially for a photograph or portrait.	*Open Set*
Subject Silhouette	The outline or shape of the subjects against a light background.	*Open Set*
Subject Proportions	The relative size and scale of the subjects' body parts or features.	*Open Set*
Text Generation	The process of creating written content to be displayed on the video.	*Open Set*

A.2.2 Camera and Lighting

On Lighting Source (Figure 15), Sunlight, Strobes, and Firelight are handled more reliably, while HMI, Fluorescent, and Tungsten lighting show lower performance. As shown in Figure 16, Aerial and Knee level camera angles are depicted better, while the Dutch and Shoulder-level angles show lower performance. On Shot Sizes (Figure 17), Medium-Wide and Master shots have stronger performance in comparison to Full and Extreme Close-Up shots.

Table 6: **Events Taxonomy** Control Nodes and Values

Name	**Description**	**Potential Values**
Standalone Actions	If the action is stand-alone	*Open Set*
Interactive Actions	If the action involves subject-subject or object-subject interaction	*Open Set*
Temporal (Actions)	How actions unfold across time.	Atomic, Concurrent, Sequential, Causal, Overlapping, Cyclic, Reverse
Foreground (Actions)	Describes if the action is taking place in the foreground	Local, Global, Focal
Background (Actions)	Describes if the is taking place in the background	-
Uncertainty	The probabilistic nature of the action outcome	Probabilistic, Deterministic, Mixed
Implicit Emotions	Emotions that are suggested or implied rather than directly stated.	*Open Set*
Explicit Emotions	Emotions that are clearly and directly shown or stated within the scene.	*Open Set*
Temporal (Emotions)	How emotions evolve across time.	Atomic, Concurrent, Sequential, Overlapping, Causal
Foreground (Emotions)	Describes if the emotion is taking place in the foreground	Local, Global, Focal
Background (Emotions)	Describes if the emotion is taking place in the background	-
Type of Dialogue Delivery	How the dialogue is delivered	Dash, Ellipsis, Monologue
Foreground (Emotions)	Describes if the dialogue is being spoken in the foreground	Local, Global, Focal
Background (Emotions)	Describes if the the dialogue is being spoken the background	-
Change in Environment	Change of environment or occurrences within a shot	*Open Set*
Story Structure	Key narrative elements that shape the scene's progression.	Turning Point, Climax, Foreshadowing, Conflict
Pace	How fast the events are happening in a shot	Slow, Fast
Regularity	How regularly the events are happening in a shot	Regular, Irregular

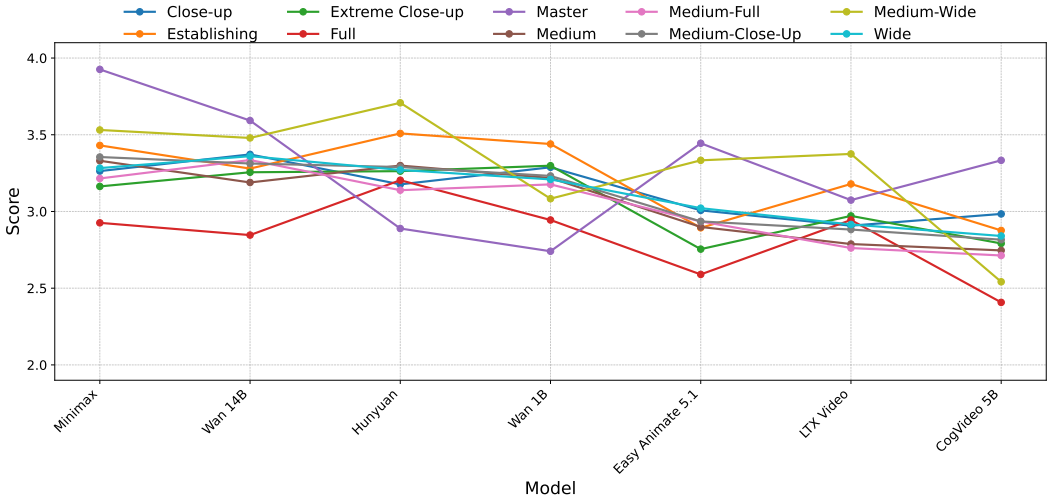


Figure 17: **Model performances across Camera Shot Size**. Models perform well on Master and Establishing shots and struggle at medium-wide and extreme-close-up shots.

A.2.3 Setup

For the Setup taxonomy, we also analyze performance at the value level. In Balance (Figure 18), models handle rule of thirds framing more effectively but struggle with symmetrical compositions. For Time of Day (Figure 19), among 11 categories, Sunrise and Morning are portrayed well, while Afternoon remains challenging.

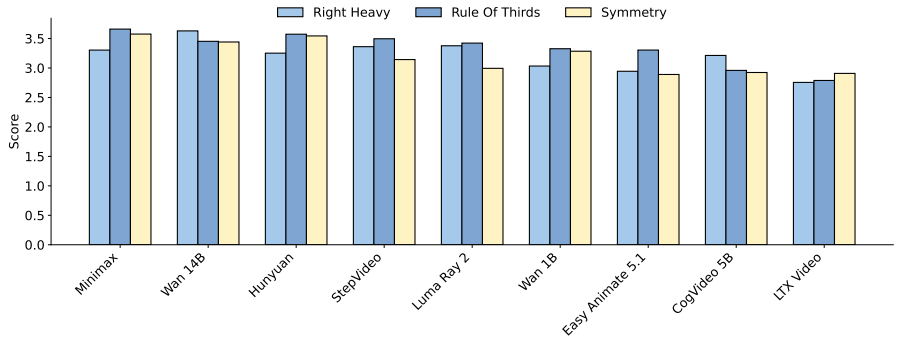


Figure 18: Between different frame compositions, models are better at Rule of Thirds but struggle at maintaining Symmetry.

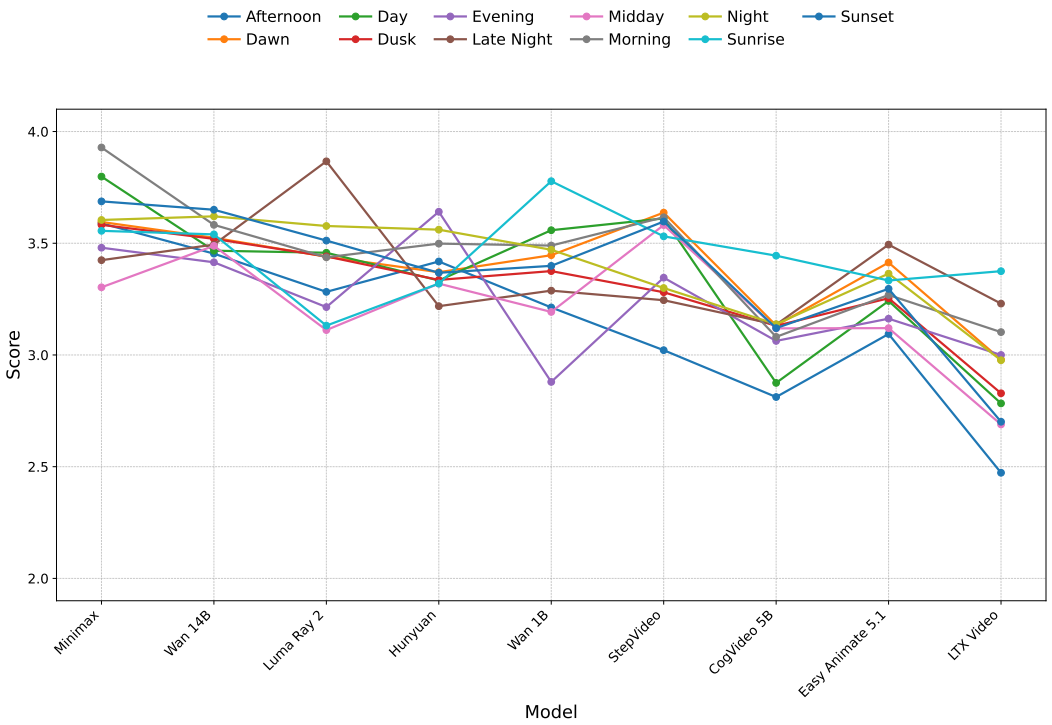


Figure 19: Across Time of Day setups, Sunrise shots are handled better, while Afternoon remains more challenging for models.

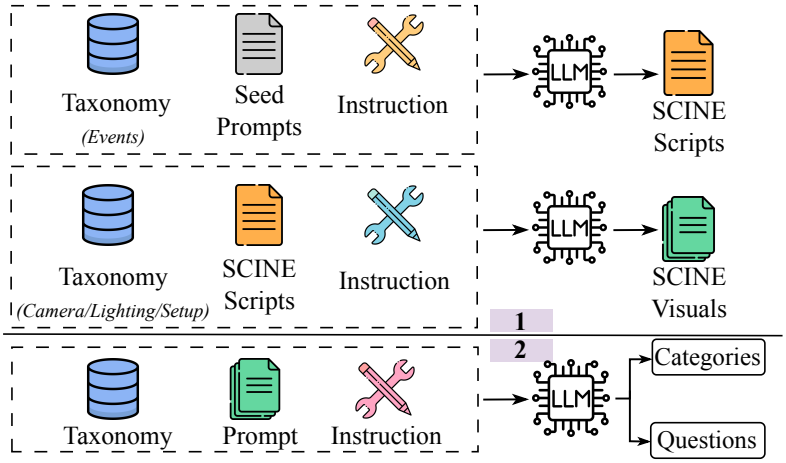


Figure 20: **1. Prompt Generation Pipeline** SCINE Scripts are created by passing seed prompts and sampled Events taxonomy nodes to an LLM, forming the narrative component of our benchmark. SCINE Visuals are then generated through structured upsampling, where nodes from the Camera, Lighting, and Setup taxonomies are sampled and injected into each SCINE Script to create prompts that capture visual exposition. 2. **Automatic Categorization and Question Generation** Given a SCINE prompt and taxonomy, we *categorize* each taxonomy element present in the prompt and generate a corresponding *question* to enable isolated evaluation of each control node.

Table 7: A working example of a prompt with its corresponding categories and questions. Each question targets a single control node from the taxonomy, enabling human annotators to perform fine-grained, independent evaluations per node.

Prompt	**Final Category**	**Question**
In a stark white laboratory illuminated by cool LEDs casting clinical precision, a scientist carefully drops a single blue chemical into a beaker, the camera framing an intimate close-up as soft depth of field blurs the sterile environment behind. A back light carves a subtle halo around the glassware moments before the liquid erupts into bright green, intensified by a strategic neon-tinted color gel that makes the reaction glow like bottled lightning.	Camera → Creative Intent → Shot Size	Does the generated video clearly exhibit a well-executed close-up shot that captures the subject with the intended intimacy and detail?
	Camera → Intrinsics → Depth of Field	Does the video effectively showcase a soft depth of field that isolates the subject while smoothly blurring the background?
	Lighting → Sources → Artificial/Practicals Light	Is the effect of artificial LED source clearly visible and does it emulate the clinical, cool lighting effect as described in the scene?
	Lighting → Color Temperature	Does the video convey a cool color temperature in its lighting setup that reinforces the clinical precision suggested in the prompt?
	Lighting → Lighting Position	Is a back lighting effect evident in the video, such that it effectively carves a halo or outline around the subject as described?
	Lighting → Advanced Controls → Color Gels	Does the video incorporate a neon-tinted color gel effect that intensifies the lighting during the chemical reaction as detailed in the prompt?

A.3 Details on Prompt Generation

The overall evaluation pipeline, depicting prompt generation, categorization and question generation, is presented in Figure 20.

Table 7 presents an example of a prompt and the corresponding categories and generated questions.

Below, we show an example of the instruction given to an LLM to upsample a SCINE-Script into SCINE-Visuals by incorporating control nodes from the Camera and Lighting taxonomies.

SCINE Scripts - Cinematographer

System Prompt :
You are a world-class cinematographer known for your visionary storytelling, mastery of light, and camera. You have decades of experience working on award-winning films across genres, collaborating with top directors and production teams. Your insights blend technical expertise with artistic sensibility. When describing scenes or advising on visual storytelling, you use cinematic terminology with clarity and inspiration. Think like Roger Deakins, Emmanuel Lubezki, and Greig Fraser—your visual choices always elevate the emotional tone and narrative arc of a project.
User Prompt :

GOAL
You will be given a prompt and 2 taxonomies that define camera and lighting controls commonly used by cinematic professionals. Your objective is to enrich the given prompt by sampling relevant nodes from both the taxonomies. As a cinematographer, your role is to "shoot" this scene using the best possible cinematic expression, utilizing the camera and lighting control options provided in the taxonomy.

PROMPT: {prompt}

MOST IMPORTANT INFORMATION
1. Only Use Nodes from the Provided Taxonomies : - You must never introduce nodes that are not present in the given taxonomies. - While the values within each node can be flexible—allowing for creativity and imagination grounded in your professional experience. For example, the node "Color Gel" is defined, but has no values. It is upto you to define these values. - The structure must strictly adhere to the nodes defined in the taxonomy. Think expansively within the bounds of each node, but never go beyond them.
2. Preserve the Original Prompt Content : - Do NOT remove or add any of the original content from the input prompt. - Your only task is to enrich the prompt by layering in camera and lighting related information. The core semantics and narrative of the prompt must remain entirely intact.
3. Do NOT include the path through which you sample the nodes in the prompt. That is, do NOT add the paths from the taxonomy using '->'.

GUIDELINES 1. Input Prompt – The input prompt describes a single continuous event, intended to occur within one uninterrupted shot. Therefore, do not include any cuts or multiple camera setups. Assume this is a one-shot sequence.
2. Each node in the taxonomies contains: - Description: A definition of what the node represents.
- Example: An example of how the node may appear in a prompt.
- Values: A non-exhaustive list of possible values for the node. Some notation: a. OPEN SET – Indicates the node supports a wide range of possible values.
b. [] – Indicates the node may have multiple values, which are not predefined and should be selected based on your reasoning and cinematic knowledge.
3. Enriched Prompt – Your enriched version will serve as input to a text-to-video model. It must be fluent, natural, and interpretable by the model, while incorporating cinematic elements effectively.

CAMERA TAXONOMY The Camera Taxonomy defines elements related to the camera's intrinsics, extrinsics, and its cinematic use : {camera_taxonomy}

LIGHTING TAXONOMY The Lighting taxonomy broadly defines all elements of lighting, including source, position of lighting, along with its effects such as shadows and reflections, along with color temperature, lighting motion such as flickering etc : {lighting_taxonomy}
When incorporating lighting into your enriched prompt, remember that a cinematographer can shape the look and feel of a shot by selectively illuminating different depth planes of the scene. Lighting can be applied to the foreground, mid-ground, background, and the subject itself—either individually or in combination. Your choices should support the emotional tone, visual focus, and narrative intent of the shot.

Below, we show an example of the instruction given to an LLM to categorize and generate evaluation questions for an input prompt using the Camera taxonomy.

Camera Categorization and Question Generation

GOAL
You are an expert prompt evaluator. Your task is to analyze a video generation prompt and categorize it based on a predefined taxonomy.

PROMPT: {prompt}
Available Categories (with Examples)
The category presented to you is that of Camera. The Camera taxonomy broadly defines everything related to the camera - the intrinsics, the extrinsics and the cinematic use of camera. {camera_taxonomy}

Notes about the Taxonomy

Each node in the taxonomy contains :
1. Description : Definition of what that node represents.
2. Example : An example of the presence of a node in the form of a prompt.
3. Values : A non-exhaustive list of values of these nodes. Values are a list of values that this node can have. Some nomenclature :
a. OPEN SET indicates that this node contains a large number of values.
b. [] indicates that this node may have multiple values, but are not defined explicitly and it is upto your reasoning and knowledge.

Examples of Categorization

1. Static Medium-Close-Up of David's face showing quiet devastation. Quick Push In as tears well up in his eyes. Shot with a medium ISO to capture the dim apartment lighting.

Static - Camera -> Trajectory -> Camera Movement -> Static
Push In - Camera -> Trajectory -> Camera Movement -> 3D
Medium-Close up - Camera -> Creative Intent -> Shot Size
Medium ISO - Camera -> Intrinsics -> Exposure -> ISO

2. Wide shot of a bustling city street at night. The neon lights of the shops and restaurants cast a colorful glow on the wet pavement. People walk by, their faces illuminated by the bright signs. The camera pans up to reveal the towering skyscrapers that loom overhead, their windows reflecting the city lights.

Wide shot - Camera -> Creative Intent -> Shot Size
Pans Up - Camera -> Trajectory -> Camera Movement -> 2D

TASK
Analyze the given prompt and return the following structured output in a valid JSON format:

Words: Extract important keywords or key phrases from the prompt using the following guidance:
- Identify named entities related to a camera in professional use as you would in NER (Named Entity Recognition).
- Extract noun phrases or descriptive terms that relate to a camera.
- Prefer multi-word expressions where meaningful related to a camera.
- Avoid generic or uninformative words like "a", "video", "the", etc.

Categories: For each word or phrase, assign the most appropriate category from the taxonomy. A dictionary of relevant categories from the taxonomy.
- For each relevant category, assign a score between '0' and '1' representing how strongly the prompt matches the category.
- Provide a reason for each score, referring to the words or phrases extracted and how they relate to the category.
- Generate a question that helps a human evaluator determine whether this category is visually present in the generated video. Use your reasoning to guide the question. The

Table 8: Lexical Diversity of SCINE Scripts. Compared to existing prompt-based benchmarks, SCINE-Scripts demonstrate higher lexical diversity across multiple metrics.

Benchmark	TTR ↑	Distinct Bi-Grams ↑	Jaccard Distance ↑
VBench [25]	0.1489	0.4605	0.9384
MovieGenBench [11]	0.1660	0.5311	0.9285
EvalCrafter [35]	**0.2270**	<u>0.6038</u>	<u>0.9413</u>
T2V-CompBench [48]	0.1435	0.4781	0.9350
SCINE Scripts	<u>0.1760</u>	**0.6177**	**0.9445**

```
evaluator will use this question to rate the video on a scale from 1 (not at all) to 5
(strongly represented).
- The generated question should evaluate quality, consistency and presence of the node
in the video.

Important Guidelines:
- The camera information should be explicitly mentioned in the prompt. Do NOT imply,
assume or derive anything.  Only consider a word or a phrase a match, if it is
explicitly mentioned in the prompt.
- Each prompt can have multiple nodes of the Camera taxonomy. You should capture all
of the nodes in the prompt and map it back to the taxonomy.
- You must always traverse from the root node, which is Camera in this case. That is,
the 'category' should always start as (Camera -> ..)
- You will never create a node that is not in the taxonomy.  These nodes can have
multiple values, as previously explained and you are expected to be imaginative about
the values. But the nodes, should always come from the given taxonomy.
- Since the taxonomy is of Camera, we do not care about objects, subjects, lighting,
events, actions or emotions.  Your sole focus should be about camera terms that are
present in the prompt in accordance with the taxonomy. You will NOT ask any question
related to objects, subjects, lighting, events, actions or emotions.
```

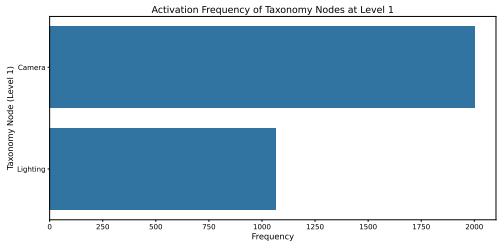
Table 8 compares SCINE-Scripts with existing prompt-based video generation benchmarks. We compute token level metrics: Type-Token Ratio (TTR), Distinct Bi-Grams, and average pairwise Jaccard Distance, and find that SCINE-Scripts exhibits strong lexical diversity.

A.4 Distribution of Taxonomy Categories in SCINE Prompts

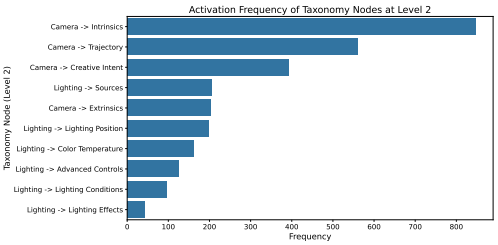
Figures 21, 22, and 23 show the distribution of activated nodes in SCINE Visuals, aggregated at the node level, across the roles of Cinematographer, Production Designer, and Director, respectively. As shown, our prompts cover a broad distribution of nodes across all the taxonomies.

A.5 Annotation Details

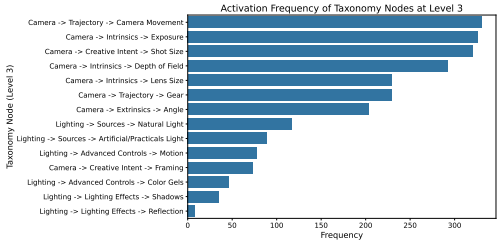
Figure 24 shows the annotation interface used by human annotators during evaluation. We also present the distribution of annotators' years of experience in film production in Figure 25. While annotations for cinematic controls can be subjective, especially given the large number of control nodes, we try our best to mitigate this by providing clear rating guidelines to annotators for each control node. Table 9 presents a minimal example of the rating guidelines shared with the annotators.



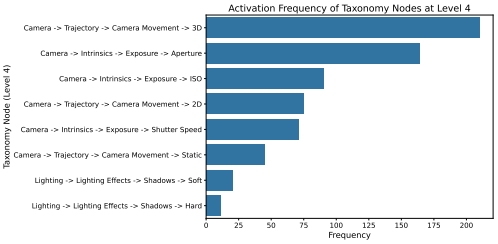
(a) Level 1 Activations



(b) Level 2 Activations

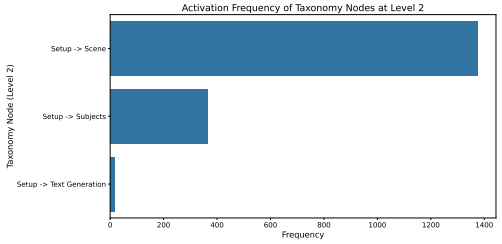


(c) Level 3 Activations

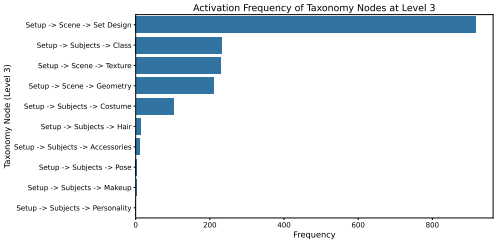


(d) Level 4 Activations

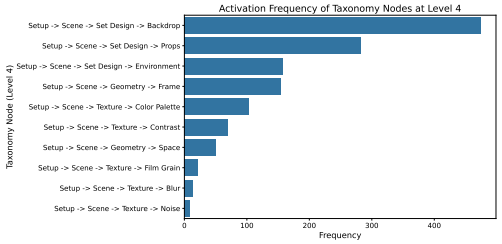
Figure 21: Node activations in Camera and Lighting taxonomies for the Cinematographer role in SCINE Visuals.



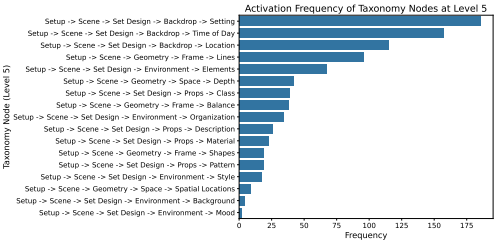
(a) Level 2 Activations



(b) Level 3 Activations

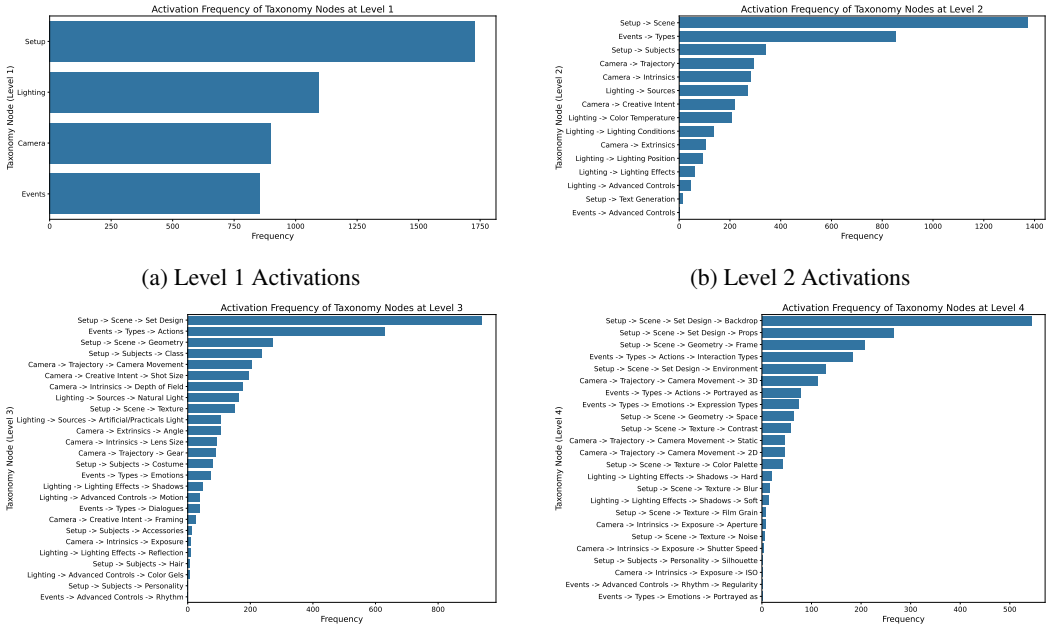


(c) Level 4 Activations



(d) Level 5 Activations

Figure 22: Node activations in Setup taxonomy for the Production Designer role in SCINE Visuals.



(a) Level 1 Activations

(b) Level 2 Activations

(c) Level 3 Activations

(d) Level 4 Activations

Figure 23: Node activations in All taxonomies for the Director role in SCINE Visuals.

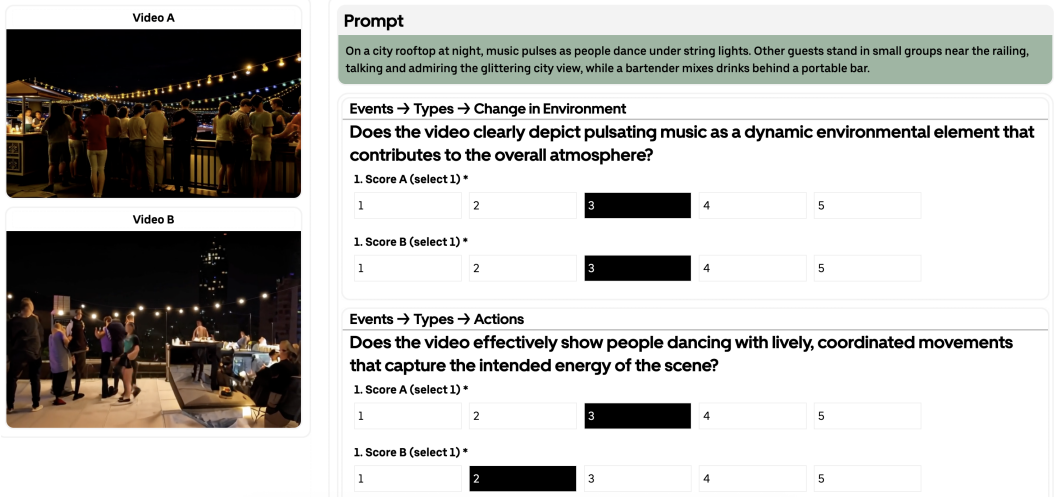


Figure 24: User Interface used by annotators to perform evaluations.

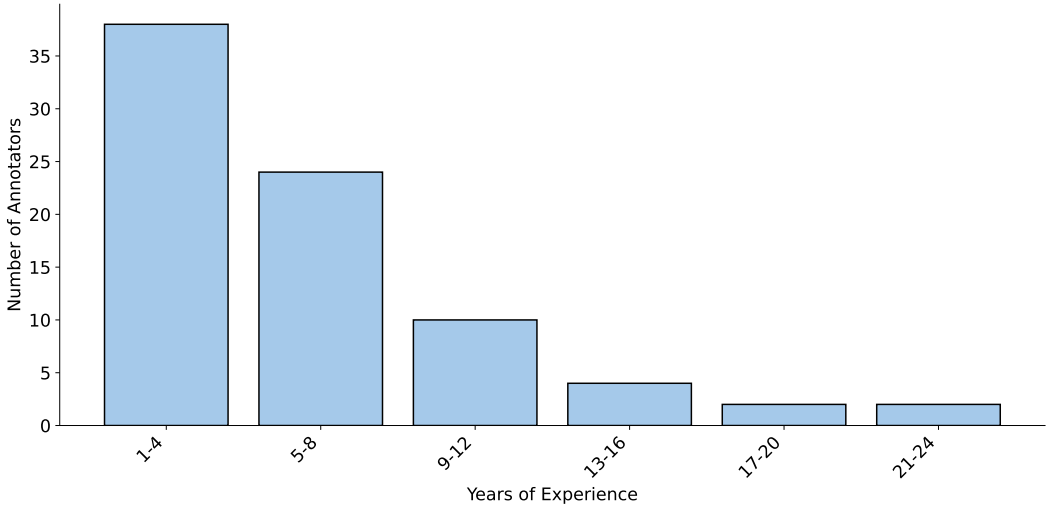


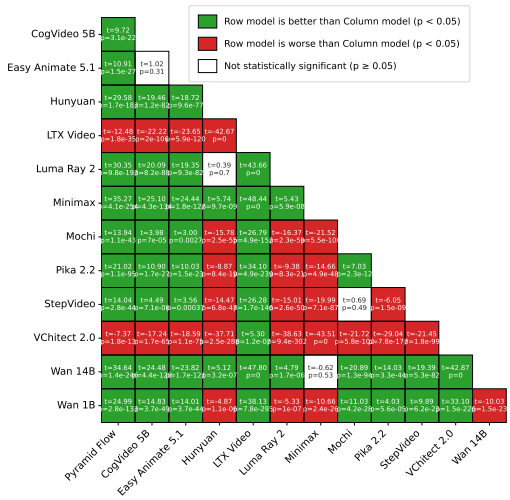
Figure 25: Distribution of the years of film production experience amongst human annotators in our evaluation setup.

Table 9: Examples of rating guidelines provided to human annotators for different control nodes, across all taxonomies.

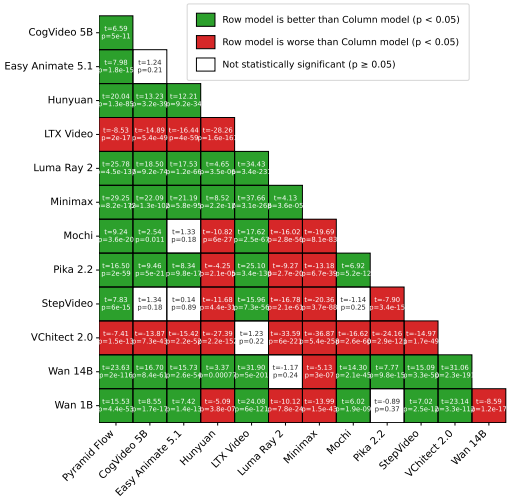
Dimension	**Score**	**What to Look For**
Low Angle	1	Camera is at or above eye level, not low angle at all.
	2	Slight upward tilt, but still feels neutral.
	3	Below subject, mild upward view, light impact.
	4	Clear low angle, subject looks larger or imposing.
	5	Strong low angle, subject dominates, towering presence.
Overlapping Actions	1	No action or one action is present.
	2	Actions are isolated or unrelated.
	3	Timing is off, they start or end awkwardly.
	4	Some overlap, but hard to follow.
	5	Fluid overlap, actions feel natural and dynamic together.
Back Light Position	1	Light clearly comes from front or side, no rim light or background separation.
	2	Some edge lighting, but not consistent or strong, subject may still blend into background.
	3	Back light is partially visible, outline is hinted but not clear on full subject.
	4	Back light is clearly present, rim light separates subject from background.
	5	Strong back light effect, glowing edges around hair or shoulders. Subject clearly pops against the background. Perfect match.
Symmetrical Frame Balance	1	Composition is clearly asymmetrical.
	2	Some repeating elements, but no visual mirror.
	3	Partial symmetry or mirrored clutter that's not clean.
	4	Almost perfect symmetry, small inconsistencies exist.
	5	Clear and precise symmetry, mirrored subjects, reflections, or centered framing. Strong and intentional.

A.6 Statistical tests

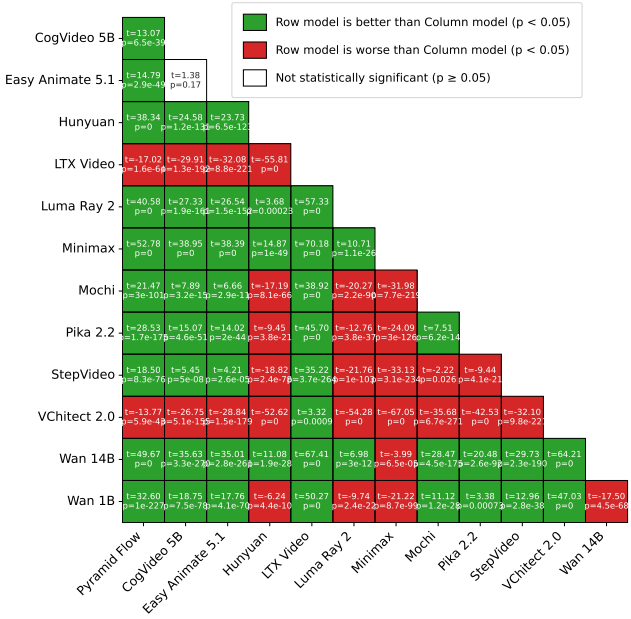
Pairwise t-tests show that the vast majority of model comparisons in our human evaluation, across all taxonomies are statistically significant at the 5% level (p <0.05). Figure (26 - 28) presents the t-test results of Events, Lighting and Camera, and Setup, respectively.



(a) Pairwise t-tests across models on emotions

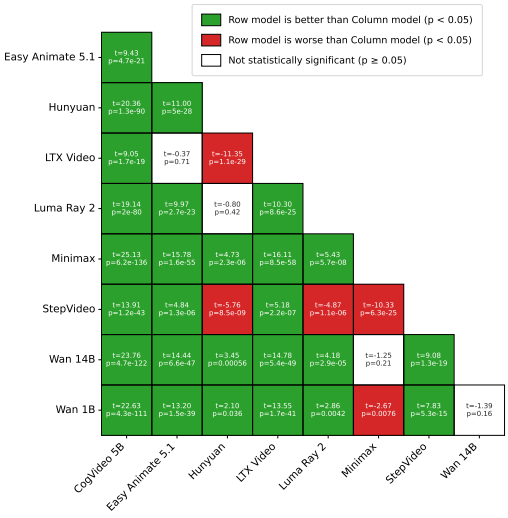


(b) Pairwise t-tests across models on dialogues

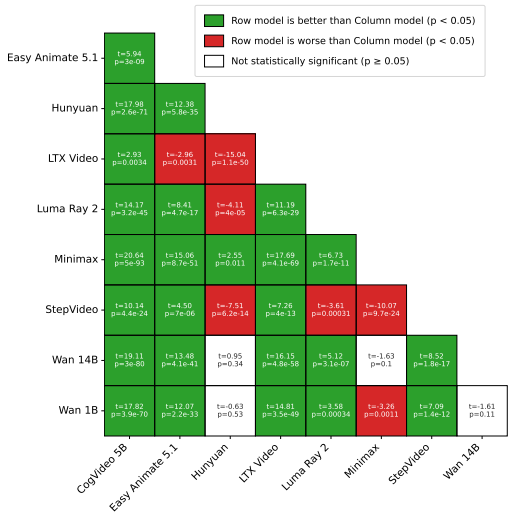


(c) Pairwise t-tests across models on actions

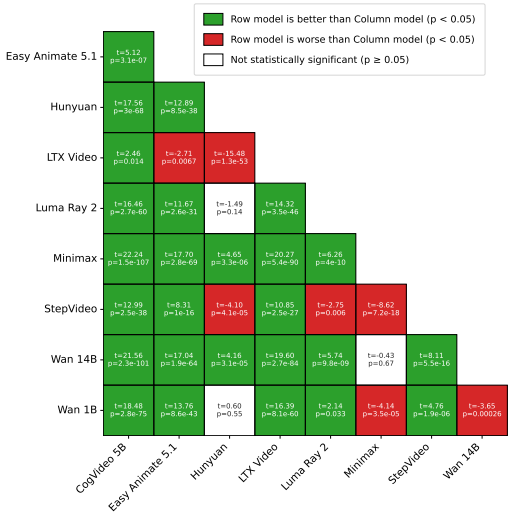
Figure 26: Statistical comparison matrices for Events: Emotions, Dialogue, and Actions.



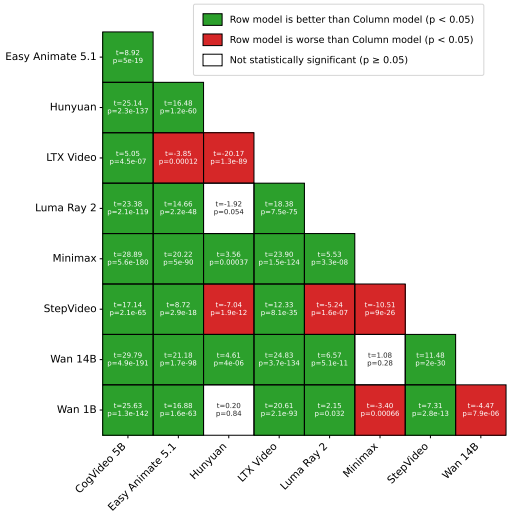
(a) Pairwise t-tests across models on Color Temperature



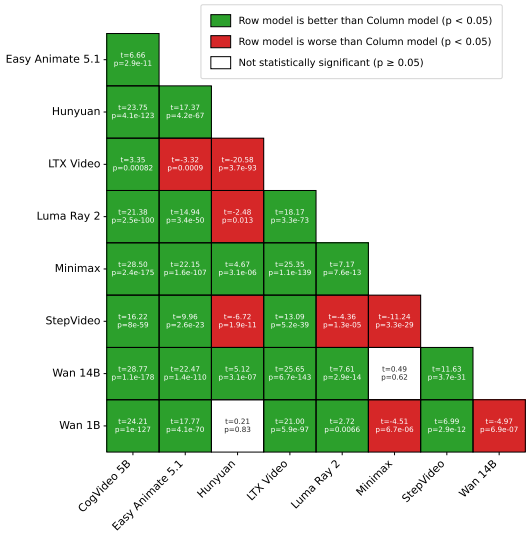
(b) Pairwise t-tests across models on Lighting Effects



(c) Pairwise t-tests across models on Extrinsics

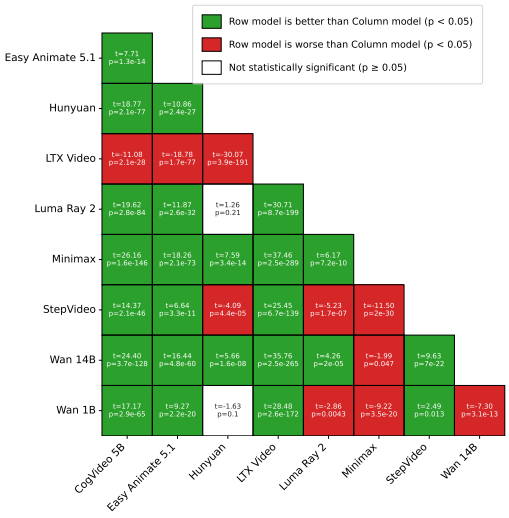


(d) Pairwise t-tests across models on Intrinsics

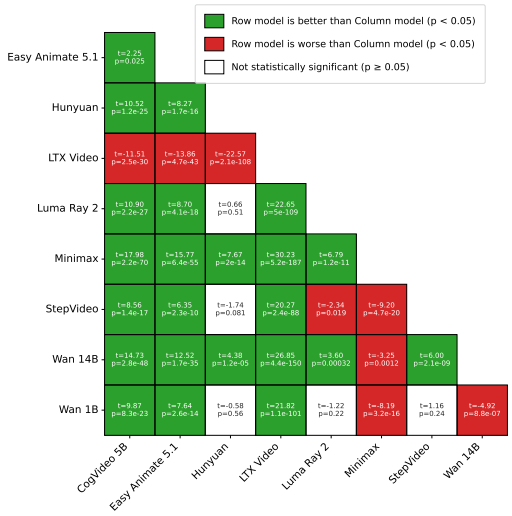


(e) Pairwise t-tests across models on Trajectory

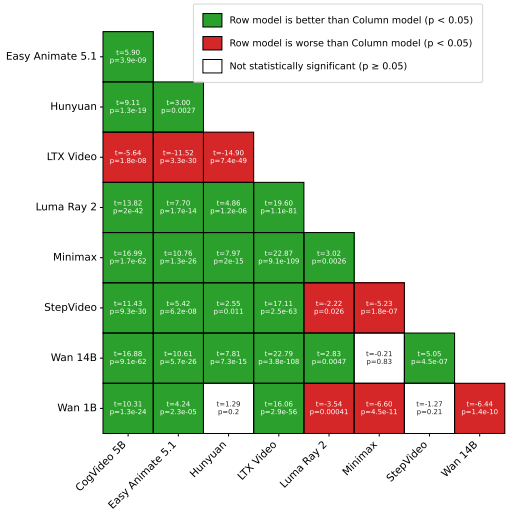
Figure 27: Statistical comparison matrices for Camera and Lighting



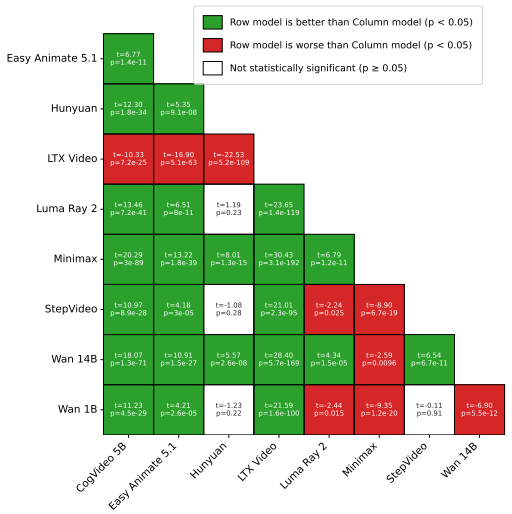
(a) Pairwise t-tests across models on Subject Class



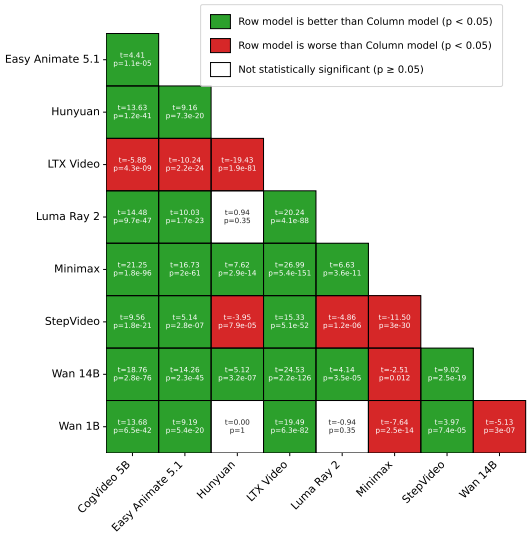
(b) Pairwise t-tests across models on Subjet Costume



(c) Pairwise t-tests across models on Elements



(d) Pairwise t-tests across models on Time of Day



(e) Pairwise t-tests across models on Location

Figure 28: Statistical comparison matrices for Setup

Table 10: Top 10 nodes where the VLM shows the strongest performance. *Score* indicates the accuracy of the VLM's choice with the most common preference among human annotators.

Node	Score
Setup→Scene→Set Design→Environment→Mood	0.88
Events→Adv.Controls→Rhythm→Pace	0.82
Setup→Scene→Set Design→Environment→Style	0.82
Setup→Scene→Set Design→Props→Utility	0.80
Events→Types→Emotions→Exp.Types→Explicit	0.79
Setup→Scene→Geometry→Frame→Shapes→Regular	0.78
Lighting→Lighting Effects→Shadows→Soft	0.76
Setup→Scene→Geometry→Space→Spatial Loc.→Rel.Pos.	0.75
Events→Types→Actions→Int.Types→Standalone	0.75
Events→Types→Emotions→Exp.Types→Explicit	0.75

Table 11: Bottom 10 nodes where the VLM shows the weakest performance. *Score* indicates the accuracy of the VLM's choice with the most common preference among human annotators.

Node	Score
Setup→Subjects→Makeup	0.33
Setup→Scene→Set Design→Environment→Background	0.33
Events→Types→Actions→Portrayed as→Contextual→Background	0.46
Setup→Subjects→Accessories	0.53
Lighting→Lighting Effects→Reflection	0.55
Setup→Scene→Texture→Color Palette	0.55
Camera→Intrinsics→Exposure→Shutter Speed	0.57
Lighting→Adv.Controls→Color Gels	0.57
Setup→Scene→Texture→Blur	0.57
Lighting→Color Temperature	0.58

A.7 Additional VLM results

Node-specific results of VLM evaluator Table 10 and 11 list the set of nodes on which our VLM evaluator has the strongest and weakest performance, respectively.

Comparison with Closed Source Models We extend our validation to closed-source, flagship SOTA models. Specifically, we evaluate two recent models from the Gemini family with distinct purposes: Gemini-2.0-Flash, optimized for fast inference, and Gemini-2.5-Pro-Preview-05-06, optimized for complex reasoning. We use the same human-aligned preference accuracy metric as with open-source models. Due to the lack of public details on model sizes, we cannot draw conclusions about scaling effects. However, Gemini-2.5-Pro consistently outperforms open-source models, including QwenVL-2.5-72B, across all categories. Notably, as shown in Figure 29, our 7B model outperforms Gemini-Flash across all categories and performs competitively with Gemini-2.5-Pro. This highlights the strength and scalability of our approach for professional video evaluation.

Reliability in VLMs A reliable VLM-as-a-Judge should produce consistent scores when given the same video, prompt, and focus aspect. In this analysis, we evaluate the raw scores generated by VLMs rather than preference rankings, and measure their stability under Best-of-5 sampling. Since VLMs are probabilistic, we evaluate reliability via the standard deviation of scores across runs. We use temperature=0 to sample to make ensure that the highest probability is selected at each sampling step. We exclude our model from this analysis, as its architecture includes a dedicated value head, unlike zero-shot VLMs that produce rewards as text. Our results show that Qwen-2.5VL-3B exhibits a high variance, making it unreliable under repeated sampling. In contrast, the flagship models and the strongest open-source model, QwenVL-2.5-72B, demonstrate high reliability, with consistently low variance (Table 12).

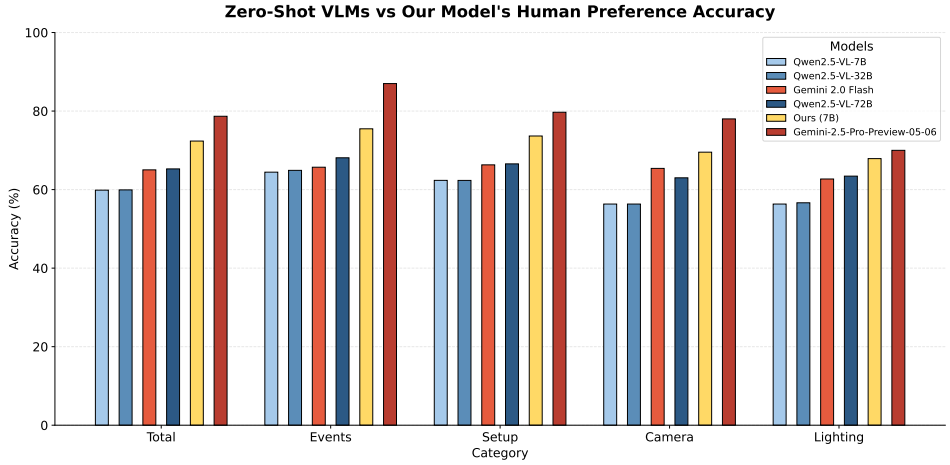


Figure 29: Preference Accuracy of open and closed-sourced VLMs in rating videos generated for Professional Use

Table 12: Measuring VLM Reliability across best-of-5 sampling

Model	**Standard-Deviation** ↓	**Krippendorff-alpha** ↑
Qwen2.5-VL-3B	2.34	0.36
Qwen2.5-VL-7B	0.37	0.84
Qwen2.5-VL-32B	0.47	0.65
Qwen2.5-VL-72B	0.23	**0.95**
Gemini2.5-Flash	0.20	0.90
Gemini2.5-Pro	**0.14**	**0.95**

Table 13: Average scores across taxonomy categories on recently released video generative models.

Model	**Camera**	**Events**	**Lighting**	**Setup**	**Overall Avg.**
Veo 3 Fast	**3.686**	**3.382**	3.974	4.078	**3.780**
Wan 2.2 14B	3.584	3.060	**4.009**	**4.114**	3.692
Wan 2.1 14B	3.239	2.821	3.827	3.975	3.466
Wan 2.2 5B	3.194	2.705	3.838	3.913	3.412
Wan 2.1 1B	3.240	2.579	3.698	3.742	3.315

A.8 Additional Results on Recent Models

We also perform small-scale evaluations on recently released models, specifically Veo 3 (Fast) [9] and the Wan 2.2 family of models. Results are presented in Table 13.

A.9 Broader Impact

We hope SCINE encourages the generative AI and computer vision communities to engage more deeply with the elements required to produce a professional cinematic shot. While our taxonomy is currently used for evaluation, it also offers a structured foundation for broader tasks such as captioning, creating control aware video datasets, and guiding model training toward explicit cinematic intent. We also envision that our fine-grained control nodes can drive new directions in open computer vision problems, such as estimating camera intrinsics from video, inferring lighting position, and understanding frame compositionality. In conclusion, we see our taxonomy as a first step toward systematically understanding cinematic controls in generative models, and we are hopeful it will be meaningfully adopted and extended across both generative AI and filmmaking communities.