CineTechBench: A Benchmark for Cinematographic Technique Understanding and Generation

Xinran Wang[1*] **Songyu Xu**[1*] **Xiangxuan Shan**[2] **Yuxuan Zhang**[1] **Muxi Diao**[1]
Xueyan Duan[2] **Yanhua Huang**[2] **Kongming Liang**[1†] **Zhanyu Ma**[1]
[1]Beijing University of Posts and Telecommunications [2]China Mobile Research Institute
`{wangxr, xusongyu, zyx_hhnkh, dmx, liangkongming, mazhanyu}@bupt.edu.cn`
`{shanxiangxuan, duanxueyan, huangyanhua}@chinamobile.com`

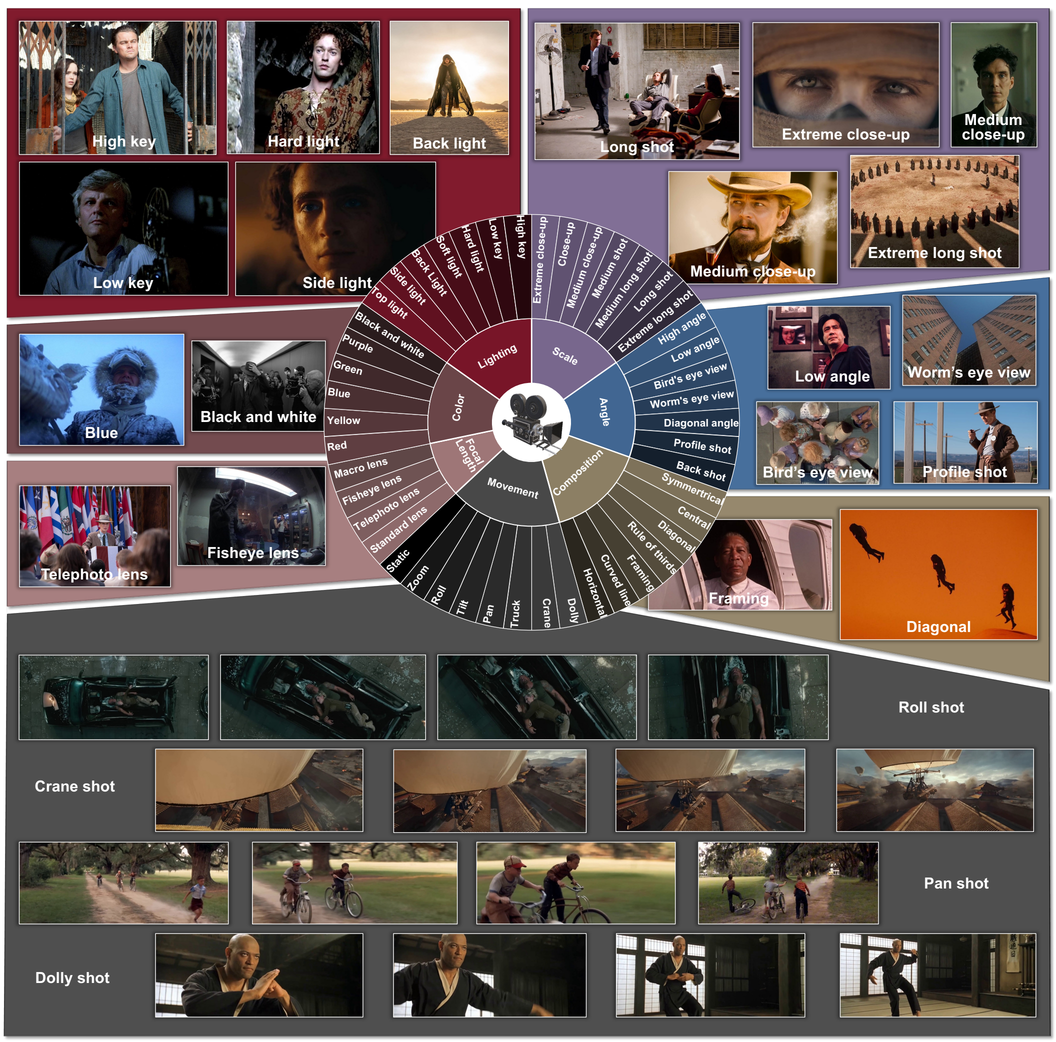


Figure 1: Cinematography taxonomy and data examples in our CineTechBench.

[*]equal contributions
[†]corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

Abstract

Cinematography is a cornerstone of film production and appreciation, shaping mood, emotion, and narrative through visual elements such as camera movement, shot composition, and lighting. Despite recent progress in multimodal large language models (MLLMs) and video generation models, the capacity of current models to grasp and reproduce cinematographic techniques remains largely uncharted, hindered by the scarcity of expert-annotated data. To bridge this gap, we present CineTechBench, a pioneering benchmark founded on precise, manual annotation by seasoned cinematography experts across key cinematography dimensions. Our benchmark covers seven essential aspects—shot scale, shot angle, composition, camera movement, lighting, color, and focal length—and includes over 600 annotated movie images and 120 movie clips with clear cinematographic techniques. For the understanding task, we design question–answer pairs and annotated descriptions to assess MLLMs' ability to interpret and explain cinematographic techniques. For the generation task, we assess advanced video generation models on their capacity to reconstruct cinema-quality camera movements given conditions such as textual prompts or keyframes. We conduct a large-scale evaluation on 15+ MLLMs and 5+ video generation models. Our results offer insights into the limitations of current models and future directions for cinematography understanding and generation in automatic film production and appreciation. The code and benchmark can be accessed at `https://github.com/PRIS-CV/CineTechBench`.

1 Introduction

Film production and appreciation play a vital role in both cultural expression and everyday entertainment. Whether through blockbuster movies, independent films, or short online videos, cinema shapes how people perceive stories, emotions, and experiences. Among the many elements that contribute to the impact of a film, cinematography serves as a powerful visual language. To convey mood, emotion, narrative, and other factors within a shot, cinematography is implemented by using different aspects within a film—ranging from camera movement and framing to lighting and composition [53]. With the rapid advancement of multimodal large language models (MLLMs) and video generation models, computer vision has made significant strides in analyzing and generating cinematic content. These models have demonstrated promising capabilities in recognizing scenes, describing plots, and even creating visually coherent video clips. However, there remains a critical gap: the lack of a standardized benchmark to assess whether MLLMs can truly understand the cinematographic techniques used in the film and video generation model can generate cinema-quality camera movements.

Recent efforts in computational movie understanding have predominantly centered on high-level semantics. This includes tasks focused on narrative comprehension, such as story-based question-answering [43], analyzing human-centric situations [48], and long video understanding [42]. Other work has targeted coarse-grained visual tasks like scene recognition [5]. While some large-scale datasets, notably MovieNet [19], have begun to incorporate "cinematic style" annotations, these labels are often holistic and lack a fine-grained decomposition into their constituent elements. In contrast, cinematography—the visual core of cinematic storytelling—finds its essence in a series of professional and specific visual techniques, including camera movement, shot scale, lighting and other dimensions. These techniques fundamentally shape a film's tone and the audience's emotional experience, yet they are not readily available from online movie reviews and require expert knowledge for meticulous visual analysis and labeling. This new era of powerful MLLMs and video generation models holds immense potential for both analyzing and creating cinematic content. Whether these models can truly understand the nuances of cinematography or generate content that artistically employs them is a frontier that remains largely unexplored due to the lack of targeted evaluation frameworks. To address this gap, we have developed a benchmark that specifically targets these fine-grained cinematographic dimensions.

In this paper, we introduce CineTechBench, a benchmark designed to evaluate the understanding and generation capabilities of MLLMs and video generation models in the context of cinematographic techniques. Our benchmark encompasses the most important dimensions of cinematography, including **shot scale, shot angle, composition, camera movement, lighting, color, and focal length**.

These dimensions play a pivotal role in shaping the visual and emotional language of film, making them essential for evaluating model's cinematographic understanding and generation abilities. To assess the understanding capability across these dimensions, we collect more than 120 video clips featuring clear and intentional camera movements, along with over 600 curated images covering the remaining dimensions. Each sample is carefully selected or annotated to highlight key cinematographic elements, providing a rich and diverse testbed for evaluating multimodal large language models and video generation models in the context of cinematographic techniques.

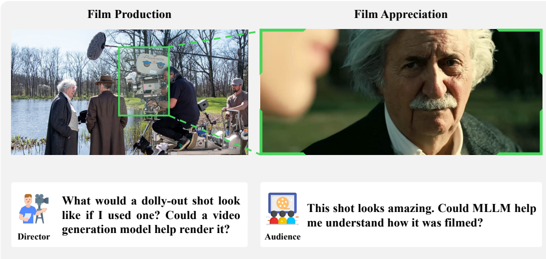
For the understanding task, we design a set of question–answer pairs and annotated cinematography-focused descriptions for both images and videos. These are used to evaluate how well multimodal large language models (MLLMs) can recognize, interpret, and describe cinematographic techniques. This task assesses the models' ability to not only identify visual elements but also articulate their narrative and emotional significance within a scene. For the generation task, we assess the ability of video generation models to recreate cinematic camera movements based on specific input conditions, e.g., textual description containing camera movement cues or the first and last frames of a clip. This setting allows us to measure how effectively video generation models can translate cinematographic intent into coherent visual outputs.

Our main contributions are as follows: (1) **We construct a taxonomy of cinematographic techniques covering 7 core dimensions**: shot scale, angle, composition, camera movement, lighting, color, and focal length. This taxonomy provides a structured foundation for the analysis and evaluation of cinematic visual understanding. (2) **We build a high-quality benchmark by collecting over 600 high-resolution film images and 120 flim clips** from critically acclaimed films, each exhibiting clear and professional cinematographic techniques. All data are manually annotated with relevant dimension labels. Based on these annotations, we further synthesize a set of cinematography-focused question–answer pairs and descriptive captions, forming a test set for evaluating both recognition and description generation. (3) **We evaluate the advanced MLLMs and video generation models on cinematographic technique understanding and camera movement generation, respectively.** Through experiments on over 15 MLLMs and 5 video generation models, we reveal that current MLLMs still struggle with fine-grained cinematograph understanding, and video generation models perform poorly on camera movement with intense rotation amplitude, highlighting the need for further research in this area.

2 Related Work

2.1 Movie Understanding Benchmarks

Previous datasets in the movie understanding domain have primarily focused on high-level semantic analysis, such as genre classification [70, 41], story comprehension [43, 42], situation recognition [48], content authenticity [13], and character detection and identification [42]. For instance, MND [30] introduced a dataset to classify scenes by their narrative function (e.g., Setup, Climax), advancing the study of macro-level story structures. To better enhance the audience's film comprehension experience, other works have focused on movie narration generation. For example, Movie101 [59] introduced a benchmark for generating role-aware narrations, which was subsequently improved and expanded into a large-scale bilingual dataset in Movie101v2 [60]. These tasks generally aim at understanding the plot or identifying key narrative elements in films, which are valuable for understanding a film's thematic content. In contrast, fewer works have explored cinematography-specific understanding, which is a crucial yet often overlooked aspect of visual storytelling [64]. Several notable efforts have explored specific cinematographic elements. For instance, MovieNet [19] provides a high-quality dataset focused on movie understanding, which includes annotations for shot scale and camera movements. MovieShots [39] offers a large-scale dataset for scale types and movement types classification. MotionSet [10] is a dataset centered around camera movement clips with movement types annotations. MovieCLIP [5, 19] utilizes CLIP [38] to automatically assign shot scale labels to shot clips, providing another perspective on annotation collection. Additionally, Camerabench [29] is focused on movement understanding, constructing a comprehensive taxonomy of camera motion primitives. However, these datasets address individual facets of cinematography, focusing on isolated aspects and lack a unified and comprehensive benchmark for evaluating fundamental cinematographic understanding across multiple core dimensions. To bridge this gap, both our work and the concurrent work ShotBench [31] have simultaneously focused on understanding the core dimensions of cinematographic techniques. While their work makes a valuable contribution to



Cinematography Benchmark Comparison

Understanding Dimensions	*Movie Shots*	*Movie Net*	*Cine Scale2*	*Camera Bench*	*Shot Bench*	*CineTech Bench*
Scale	✓	✓	×	×	✓	✓
Lighting	×	×	×	×	✓	✓
Angle	×	×	✓	×	✓	✓
Camera Movement	✓	✓	×	✓	✓	✓
Composition	×	×	×	×	✓	✓
Color	×	×	×	×	×	✓
Focal Length	×	×	×	×	✓	✓
Shot Framing	×	×	×	×	✓	×
Generation Dimensions						
Camera Movement	×	×	×	×	×	✓

Figure 2: Our benchmark focus on the **cinematographic techniques** in film production and appreciation. Compared with similar benchmarks, our benchmark include more core dimensions in cinematography.

cinematographic analysis, our work provides a broad and structured evaluation framework that is distinguished by also including an evaluation for camera movement generation.

2.2 Movie & Video Generation Benchmarks

The field of video generation has recently seen burgeoning growth, with a significant number of innovative works emerging that aim to produce movie-level visuals. These include foundational video generation models [50, 18, 7, 68, 4, 40, 69, 11, 22, 62, 22, 16], controllable visual generation framework [1, 12, 57, 51, 17, 66, 63], identity-preserving video generation [17] and audio and video synchronous generation [52]. Furthermore, to construct coherent narratives, the community has proposed multi-shot generation methods [54]. Several benchmarks are established to corresponding visual generation evaluation benchmarks [58, 33, 71] to evaluate these technologies. However, evaluating the film-level generation capabilities of these models—especially regarding cinematographic aspects such as camera movement—remains a challenging task. Several recent benchmarks have addressed general video generation evaluation. VBench [20, 21] provides a comprehensive benchmark suite that dissects video generation quality into hierarchical, disentangled dimensions with tailored prompts and evaluation protocols. DEVIL [27] focuses on the dynamics dimension, offering a detailed protocol for evaluating the temporal coherence of text-to-video (T2V) generation models. Meanwhile, MovieGen Video Bench [36] evaluates video generation models from the perspectives of visual quality, realism, and aesthetics. Concurrently, SCINE [6] focuses on prompt-driven T2V evaluation, measuring generated-video quality via filmmaking taxonomies with an automatic evaluator and a question-generation pipeline. Despite these advances, there is still a lack of benchmarks tailored specifically for the reconstruction of cinematographic technique, particularly camera movement, against original videos, in generated video content. Our benchmark fills this gap by focusing on the assessment of cinema-level camera movement generation.

3 CineTechBench

CineTechBench offers high-quality, expert-annotated data across multiple dimensions of cinematography. As illustrated in Figure 2, our benchmark focuses specifically on the domain of cinematographic techniques in film production and appreciation. Different from existing movie and camera understanding benchmarks, CineTechBench establishes the first comprehensive taxonomy that covers seven core dimensions of cinematography: shot scale, angle, composition, movement, lighting, color, and focal length. These dimensions reflect the visual language used by professional filmmakers and provide a structured foundation for evaluation.

3.1 Taxonomy Building

Establishing a rigorous taxonomy is essential for evaluating model performance in any specialized domain. As shown in Figure 3, we began by collecting keywords from online sources in film review websites, YouTube tutorials, and cinematography-focused educational content, such as videomaker[3],

[3] www.videomaker.com

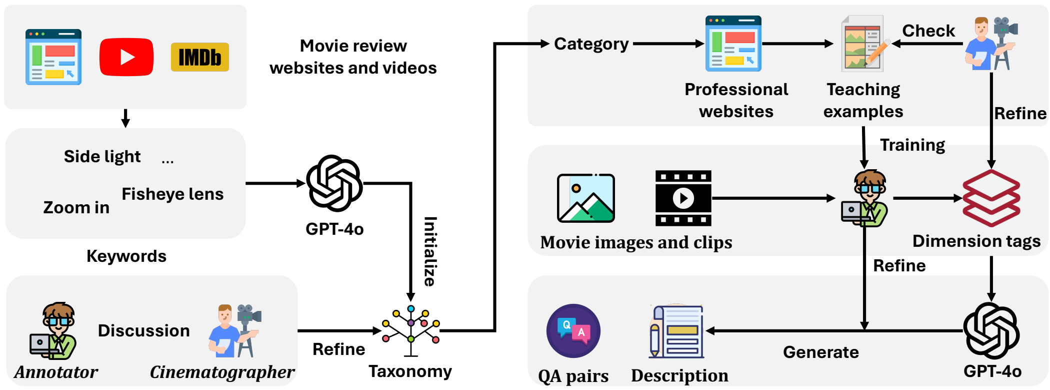


Figure 3: Overview of our benchmark building process.

studiobinder[4], and nofilmschool[5]. We then organized these keywords into a hierarchical taxonomy using GPT-4o, which was further refined through iterative feedback from professional cinematographers. Following are the seven core dimensions in our cinematographic taxonomy, a more detailed explanation of the categories within each dimension is provided in Appendix I.

Scale refers to the shot distance, which defines the spatial relationship between the subject and the frame. This dimension influences the viewer's perception of detail, context, and emotional intensity.

Angle describes the orientation of the camera relative to the subject, shaping the viewer's perspective and emotional response. Different angles can evoke varied psychological effects.

Composition concerns the arrangement of visual elements within the frame. It guides the viewer's attention, establishes visual harmony or tension, and enhances narrative expression.

Colors encompasses the hue, saturation, and tonal palette used in a shot. Colors are central to setting mood, evoking emotion, and reinforcing thematic motifs.

Lighting addresses the quality, direction, and intensity of illumination in a scene. It plays a critical role in establishing atmosphere, emphasizing form, and generating visual depth.

Focal Length pertains to the optical characteristics of the camera lens. This dimension affects spatial representation, subject emphasis, and visual aesthetics.

Camera Movement This dimension captures the dynamic motion of the camera during a shot. Following CameraBench [29], we categorize camera movements into five types: (1) **Translation**: lateral (truck), forward / backward (dolly), and vertical (pedestal) movements. (2) **Rotation**: angular movements including pan (horizontal), tilt (vertical), and roll (diagonal). (3) **Zoom**: optical zoom in and zoom out, altering framing without moving the camera. (4) **Static**: fixed shots where the camera remains completely stationary. (5) **Combined movement**: compositions involving multiple consecutive or simultaneous camera motions.

3.2 Data Collection & Annotation

Movie images and clips. Since no existing data source or website gather film clips and images showcasing clear, professional cinematographic techniques, it is very difficult for us to use automated methods to collect materials and corresponding annotations on a large scale. Therefore, we manually assembled our own benchmark. First, we gathered over 600 high-resolution stills, each illustrating a distinct static shot style (e.g., shot scale, composition, lighting) from IMDb's Top 250 films [6]and other curated movie databases. Every image was annotated across the relevant cinematographic dimensions. Second, we downloaded more than 120 short video clips from the YouTube channel Movieclips [7], specifically selecting segments that demonstrate clear camera movements (e.g., pan, tilt, dolly, zoom).

[4] `www.studiobinder.com`
[5] `www.nofilmschool.com`
[6] `https://www.imdb.com/`
[7] `https://www.youtube.com/@MOVIECLIPS`

Table 1: Accuracy of various MLLMs on static cinematographic technique question answering understanding. The best and second best results are highlighted by blue and green respectively.

MLLMs	Params	Overall	Scale	Angle	Composition	Color	Lighting	Focal Length
Commercial								
GLM-4V-Plus [15]	–	60.00	50.71	69.14	67.50	83.33	56.36	31.67
Qwen-VL-Plus	–	61.36	40.71	73.33	67.50	81.67	66.36	43.33
Gemini-2.0-Flash	–	59.34	46.43	74.17	40.83	91.67	70.91	43.33
Gemini-2.5-Pro	–	69.67	71.43	83.33	67.50	88.33	62.73	36.67
Doubao-1.5-vision-pro	–	56.07	42.86	68.33	41.67	78.33	60.00	61.67
GPT-4o [34]	–	70.16	75.00	82.50	57.50	93.33	71.82	33.33
Open-source								
Kimi-VL [45]	3B	46.39	32.14	63.33	31.67	73.33	55.54	31.67
Phi3.5 [47]	4B	40.82	20.00	49.17	41.67	61.67	56.36	21.67
Gemma3-it [44]	4B	39.18	17.86	45.00	41.67	58.33	52.73	28.33
Qwen2.5-VL [2]	7B	50.66	30.00	61.67	43.44	83.33	62.73	36.67
Qwen2.5-Omni [55]	7B	54.75	45.00	65.83	61.67	70.00	49.09	36.67
LLaVA-OneVision [25]	7B	45.90	31.43	54.17	42.50	75.00	54.55	25.00
LLaVA-NeXT [24]	8B	38.69	22.86	42.50	39.17	63.33	44.55	31.67
MinCPM-V-2.6 [56]	8B	45.90	32.86	57.50	35.00	80.00	50.91	31.67
InternVL2.5 [8]	8B	54.59	39.29	63.33	65.00	90.00	52.73	20.00
InternVL3 [72]	8B	55.25	45.00	66.67	53.33	76.67	57.27	35.00
Llama-3.2-Vision [46]	11B	47.21	33.57	48.33	50.83	78.33	45.45	41.67

These clips form our test set for video generation and motion-understanding tasks. By restricting our selection to films in IMDb's Top 250, we ensure that all materials exhibit exemplary technical craftsmanship, visual storytelling, and enduring cinematic value. More statistical information about our benchmark can be found in Appendix A.

Annotation To support high-quality annotation, we first searched for professional websites using cinematography-related category keywords (e.g., "extreme close up shot", "camera movement"). These websites typically include visual examples, images or video snippets, corresponding to each category. We curated five representative examples per category (a sample website is provided in Appendix A) and used them to train a team of annotators with a foundational understanding of cinematography. After training, the annotators labeled the collected images and video clips according to the relevant cinematographic dimensions. During annotation, any instance that was ambiguous or difficult to classify was either escalated to a professional cinematographer for review or discarded to maintain the overall quality of the dataset. Building on basic category annotations across key cinematographic dimensions, we further enriched the data set by generating question-answer pairs and descriptive annotations using GPT-4o. GPT-4o was guided by our predefined taxonomy and the existing category labels to ensure relevance and consistency. All generated content was manually reviewed and refined by trained annotators to ensure accuracy, clarity, and alignment with professional cinematography standards. More annotation details are shown in Appendix B. This process result in 610 image QA pairs, 128 video QA pairs, 100 detailed image descriptions (average length $\approx$176 words) and 128 detailed video descriptions (average length $\approx$168 words).

4 Evaluation

In this section, we evaluate both understanding and generation tasks using our proposed CineTech-Bench. For the understanding task, we assess over 15 advanced MLLMs on both dynamic aspects (e.g., camera movement) and static aspects (e.g., shot angle, shot style) of visual content, through both question-answering and description generation tasks (see Section 4.1). These evaluations leverage movie images and clips to comprehensively examine MLLMs' ability to interpret various cinematographic dimensions. For the generation task, we benchmark over five advanced video generation models on the camera movement generation task (see Section 4.2) to assess their ability to generate coherent camera movements. The detailed experiment settings are shown in Appendix D.

4.1 Cinematographic Technique Understanding

Metrics For question-answering tasks, we report overall accuracy as well as accuracy broken down by each cinematography dimension. For description generation tasks, we use four reference-based metrics. Three of these—BLEU [35], METEOR [3], and ROUGE [28]—are based on n-gram overlap. However, such metrics are limited in evaluating fine-grained, detailed descriptions [14]. To address

Table 2: Accuracy of various MLLMs on camera movement question answering understanding. The best and second best results are highlighted by blue and green respectively.

MLLMs	Params	Frames	Overall	Static	Translation	Rotation	Zoom	Combined
Commercial								
GLM-4V-Plus [15]	–	1fps	52.34	100.00	40.74	41.94	57.14	68.00
Qwen-VL-Plus	–	8fps	52.40	100.00	56.60	33.33	57.14	43.48
Doubao-v1.5-vision-pro	–	2fps(>=8)	40.00	100.00	40.74	16.13	14.29	48.00
GPT-4o	–	2fps(>=8)	50.00	90.91	61.11	25.81	28.57	44.00
Gemini-2.0-Flash	–	1fps	49.22	27.27	61.11	32.26	28.57	60.00
Gemini-2.5-Pro	–	1fps	56.69	81.82	66.04	45.16	14.29	52.00
Open-source								
Phi3.5 [47]	4B	1fps(>=4)	27.19	10.00	33.33	31.03	40.00	26.32
gemma3-it [44]	4B	1fps(>=4)	33.33	60.00	36.54	16.67	14.29	45.83
Qwen2.5-VL [2]	7B	1fps	50.78	100.00	55.56	19.35	71.43	52.00
Qwen2.5-Omni [55]	7B	–	46.09	72.73	46.30	19.35	42.86	68.00
LLaVA-NeXT-Video [24]	7B	64	28.00	45.45	29.63	19.35	14.29	32.00
LLaVA-OneVision [25]	7B	32	36.00	90.91	35.19	16.13	42.86	36.00
LLaVA-NeXT [24]	8B	4	29.13	63.64	31.48	13.33	28.57	28.00
MinCPM-V-2.6 [56]	8B	1fps	35.94	27.27	42.59	25.81	0.00	48.00
InternVL2.5 [8]	8B	all	34.38	72.73	40.74	16.13	57.14	20.00
InternVL3 [72]	8B	all	41.41	81.82	35.19	29.03	42.86	52.00
Llama-3.2 [46]	11B	4	31.25	18.18	27.78	35.48	14.29	44.00

Table 3: Performance of various MLLMs on cinematographic technique description. The best and second best results are highlighted by blue and green respectively.

MLLMs	Params	BLEU@4	METEOR	ROUGE-L	CAPability			
					HR	AP	AR	F1
Commercial								
GLM-4V-Plus [15]	–	4.33	18.63	25.41	84.43	50.15	40.45	43.18
Qwen-VL-Plus	–	0.72	15.24	12.97	81.29	48.25	36.24	40.38
Doubao-v1.5-vision-pro	–	4.02	19.44	24.93	82.91	53.01	39.27	42.67
GPT-4o	–	6.08	19.76	27.13	86.18	56.86	45.66	49.08
Gemini-2.0-Flash	–	4.17	19.07	25.14	85.42	51.28	40.75	44.43
Gemini-2.5-Pro	–	6.12	21.64	25.35	88.81	57.82	48.67	52.27
Open-source								
Phi3.5 [47]	4B	2.24	15.76	21.97	11.72	73.33	8.59	15.38
gemma3-it [44]	4B	2.11	17.53	21.15	89.14	44.75	36.70	39.07
Qwen2.5-VL [2]	7B	3.14	17.73	23.28	86.71	52.30	42.05	44.39
Qwen2.5-Omni [55]	7B	3.67	17.89	24.80	85.92	45.81	37.13	39.92
LLaVA-OneVision [25]	7B	2.58	17.42	22.16	81.31	46.69	34.68	37.32
LLaVA-NeXT [24]	8B	1.89	17.11	21.64	81.39	45.38	32.93	35.62
MinCPM-V-2.6 [56]	8B	3.06	16.15	23.05	77.07	45.73	30.89	34.48
InternVL2.5 [8]	8B	3.44	17.56	24.08	85.72	51.21	39.83	42.16
InternVL3 [72]	8B	4.10	19.12	25.38	86.91	55.64	45.91	47.86
Llama-3.2 [46]	11B	2.66	17.41	23.65	85.60	45.51	37.57	39.58

this, we additionally incorporate evaluation metrics from the CAPability benchmark [32] based on our taxonomy, which reliably assess both the correctness and thoroughness of MLLM-generated descriptions using hit rate (HR), average precision (AP), average recall (AR) and F1-score.

Results We first evaluate MLLMs' understanding of static cinematographic techniques—scale, angle, composition, color, lighting, and focal length using annotated image question-answer pairs. Results are shown in Table 1. Among commercial models, GPT-4o and Gemini-2.5-Pro achieve the highest and second-highest overall scores (70.16% and 69.67%, respectively), primarily due to their strong performance on scale (75.00%, 71.43%) and angle (82.50%, 83.33%). Gemini-2.0-Flash, while slightly lower in overall accuracy (59.34%), exhibits the leading color understanding performance (91.67%) and strong lighting perception (70.91%). Doubao-1.5-Vision-Pro, although underperforming across most dimensions, achieves the highest focal length accuracy (61.67%) among all MLLMs. Open-source MLLMs lag significantly behind, averaging about 15 percentage points lower in overall accuracy. Among them, InternVL3 leads with 55.25%, showing relative strength in angle (66.67%), scale (45.00%), and lighting (57.27%). Notably, Qwen2.5-VL-7B achieves the best lighting perception (62.73%) among open-source models, outperforming even some commercial counterparts. We next assess models' understanding of camera movement using video question answering pairs. As shown in Table 2, Gemini-2.5-Pro achieves the best overall performance (56.69%). Among open-source models, Qwen2.5-VL and Qwen2.5-Omni rank first and second, respectively. Surprisingly, several open-source MLLMs struggle to recognize fixed shots, resulting in poor performance on the "static" category—e.g., LLaVA-NeXT-Video. Across all models, camera rotation remains a particularly challenging dimension, with consistently low accuracy. To evaluate overall comprehension, we test each MLLM's ability to generate comprehensive descriptions. As shown in Table 3, Gemini-2.5-Pro achieves the highest average precision (AP), average recall (AR), and F1 score, indicating its outputs are both accurate and complete. Among open-source models, InternVL3 performs best—surpassing even some commercial MLLMs such as Gemini-2.0-Flash. More understanding results are shown in Appendix E.

Qualitative Analysis We further illustrate these findings with qualitative examples in Figure 4. In example (b), which tests shot angle recognition, both Llama-3.2 and GLM-4V-Plus misclassify the scene as Diagonal instead of the correct Profile. Example (d), evaluating color palette understanding, shows Gemma3 and LLaVA-OneVision incorrectly focusing on a local object (a desk lamp) rather than assessing the overall scene color. In example (e), where the ground truth is Side Light, all MLLMs fail, with Gemini-2.0-Flash misclassifying it as Back Light. Example (f) further reveals widespread difficulty across models in recognizing lighting and focal length. Examples (g) and (h) illustrate challenges in camera movement understanding, even GPT-4o misinterprets camera rotation direction. In example (i), generated descriptions from all MLLMs fail to accurately reflect the ground truth, highlighting limitations in comprehensive and correct description generation.

Table 4: Cinematic camera motion control performance of different image-to-video models. F, L and T means the first frame, the last frame and textual description of the movie clip, respectively. The best results are highlighted in blue.

I2V models	Condition	RotError ↓	TransError ↓		CamMC ↓		CLIP-IS ↑
			Rel.	Abs.	Rel.	Abs.	
Commercial							
Klingv1.6	FLT	21.68	48.49	196.14	62.57	207.65	90.15
Gen4turbo	FT	23.61	49.84	102.32	64.47	117.07	86.96
Open-source							
Wan2.1-FLF-14B-720P [50]	FLT	27.80	48.31	99.61	67.82	115.76	89.65
FramePack-FLF2V [62]	FLT	23.88	58.10	82.00	71.98	95.62	89.30
FramePack-I2V [62]	FT	26.93	61.94	192.08	78.17	208.78	82.70
Hunyuan-Video-I2V [22]	FT	33.42	71.65	268.62	91.87	289.36	83.98
SkyReels-V2-I2V-1.3B-540P [7]	FT	40.05	74.86	423.52	100.96	442.34	78.42

4.2 Camera Movement Generation

Metrics In this section, we use video generation models to reconstruct the camera movement in the original film clip by inputting the first frame, the last frame (if applicable), and textual description. Following [49, 26, 66], we quantify trajectory similarity between the generated and the original video clips via three metrics: rotation error (RotErr), translation error (TransErr), CamMC. The TransErr and CamMC metrics are reported in two forms: relative (Rel) and absolute (Abs). The relative error normalizes each video by its own scene scale, focusing purely on the correctness of the camera path. It provides a more stable and reasonable evaluation than the absolute error, which also penalizes inaccuracies in the overall scene scale. Consequently, our subsequent analysis focuses primarily on the relative metrics. We use MonST3R [61] to estimate the camera trajectory of the generated and original movie clip. Finally, we also report a CLIP-based frame similarity score (CLIP-IS) to capture visual consistency. The detailed introduction of these metrics are in Appendix C.

Results The overall results are shown in Table 4. Among the commercial video generation models, Klingv1.6 with first and last frame control achieves the best performance on both RotError and TransError. Among open-source models, Wan2.1 and FramePack support first frame and last frame control, obtain relatively good performance compared to the models conditioned purely on the first frame, such as HunyuanI2V and FramePack-I2V. We further divide the test examples by their camera movement translation speed and camera rotation angular velocity, and average their translation error and

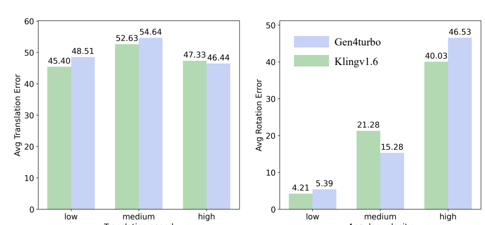


Figure 6: Average TransError and RotError on different translation speed and angular velocity.

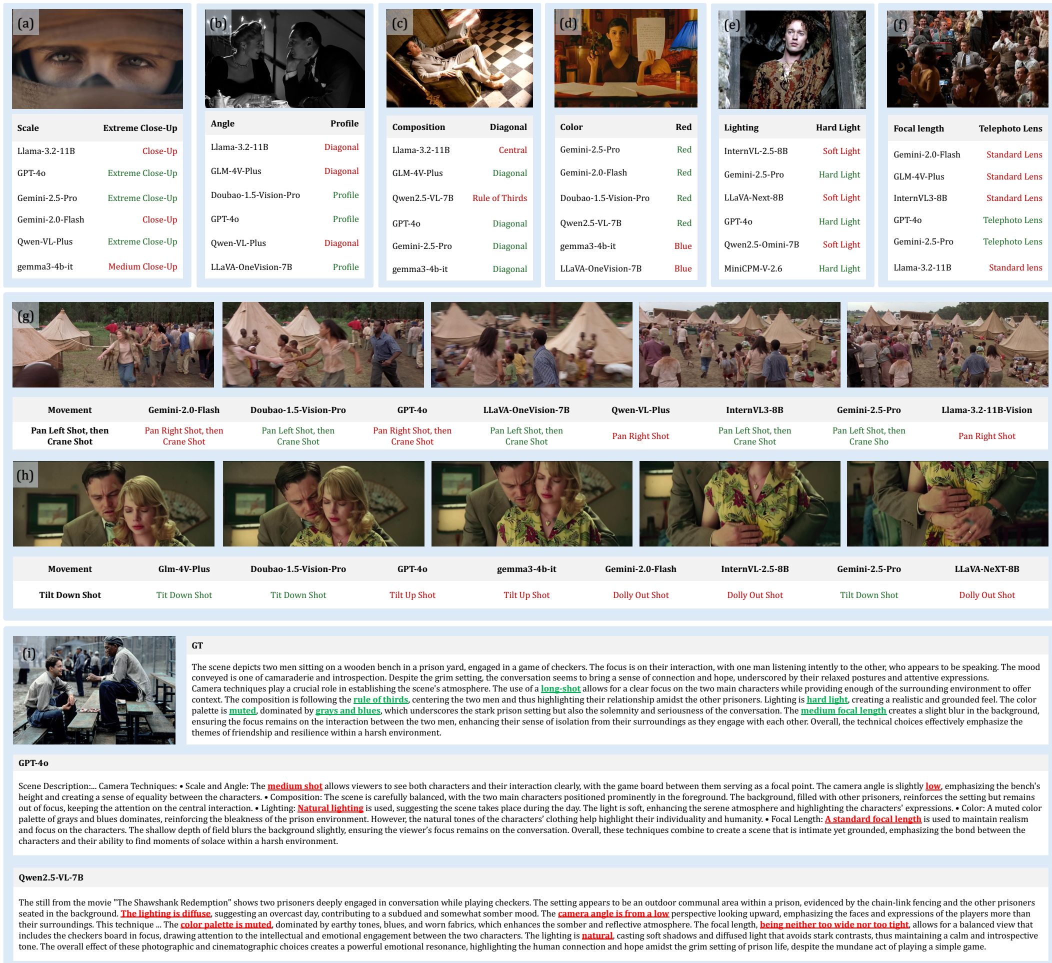


Figure 4: Visualization of MLLMs' answers on cinematographic technique question answering task. The red text highlights the wrong answers and the green text highlights the correct answers. More visualization examples can be seen in Appendix G.

Figure 5: Generated movie clips by different video generation models and the corresponding camera trajectory estimated by Monst3r [61]. More examples are shown in Appendix G.

rotation error respectively, the results are shown in Figure 6. The video generation models usually have a higher error on examples with high camera rotation angular velocity, which is mainly used in shots with intense fighting scenes. We show generation results of different video generation models in the Figure 5. The original clip applies a counter-clockwise roll camera movement. Among the three models, Wan2.1 doesn't generate a roll camera movement at all. Although the video generated by Kling has a sense of rotation, its roll direction is clockwise, which is the opposite of the intended motion. Only Gen4turbo generates the correct camera movement with correct direction. We further analyze more generation examples in Appendix G.

5 Future Direction

Future extensions of this work could deepen the evaluation of cinematographic understanding by establishing explicit connections between camera techniques and narrative structure. For example, models could be assessed on their ability to recognize how specific shot types—such as over-the-shoulder angles, tracking shots, or extreme close-ups—contribute to character development, emotional tone, or plot progression. A richer understanding of film language would also benefit from expanding the diversity and scale of the underlying video corpus, incorporating a broader range of genres, cultures, and directorial styles to reduce bias and improve generalization. On the generation side, current evaluation tasks focus on reconstruction—that is, whether models can reproduce specific cinematographic techniques in a visually coherent manner. While this serves as a useful starting point, it represents only a constrained form of generation. Future work could explore more advanced tasks such as cinematographic re-composition, where models are required to modify or re-edit videos based on high-level stylistic and narrative instructions (e.g., changing shot scale, adjusting lighting, or reconfiguring spatial composition). With the emergence of more capable video models, such as Runway's Aleph, this line of evaluation is becoming increasingly feasible. These directions would help move the field closer to assessing and developing models with not only visual fluency but also narrative and stylistic awareness—key components of true cinematic intelligence.

6 Conclusion

In this work, we introduce CineTechBench, the first benchmark that evaluates MLLM understanding across seven core dimensions (shot scale, angle, composition, camera movement, lighting, color, focal length) and video generation models on camera movement generation. We curated and annotated over 600 still images and 120 video clips from acclaimed films, each paired with targeted QA pairs and descriptions. Our evaluation of over 15 state-of-the-art models, reveals key limitations in current models on understanding and generation of cinematographic techniques. Specifically, for understanding, we found that multimodal large language models profoundly struggle with complex, relational concepts like lighting direction and camera movement. This is demonstrated by a significant score gap between high hit rates and low F1 scores, as shown in Figure 12, suggesting that models often resort to heuristic guessing over robust interpretation. We trace this failure to the scarcity of technical terms in pre-training corpora (e.g., "focal length" in 0.05% of LLaVA-Video-178k captions [65]). Fundamentally, this weak performance highlights the limited capacity of current models for spatial reasoning and coherent dynamic change perception in visual media. For generation, we found that video generation models struggle to synthesize dynamic camera motions. While conditioning on first and last frames improves control for simple movements, models largely fail to render intense camera rotations, such as those common in action sequences. By providing this benchmark, we aim to drive multimodal large language models and video generation models with more nuanced cinematic analyzing and robust motion synthesis capabilities. Future work might focus on scaling these annotations in a more efficient way to further elevate model performance.

Acknowledgment

This work was supported by the National Nature Science Foundation of China (Grant 62476029, 62225601, U23B2052), funded by Beijing University of Posts-China Mobile Communications Group Co.,Ltd. Joint Institute, the Fundamental Research Funds for the Beijing University of Posts and Telecommunications under Grant 2025TSQY08, and sponsored by Beijing Nova Program.

References

[1] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. Recammaster: Camera-controlled generative rendering from a single video, 2025.

[2] 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, 2025.

[3] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

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

[5] Digbalay Bose, Rajat Hebbar, Krishna Somandepalli, Haoyang Zhang, Yin Cui, Kree Cole-McLaughlin, Huisheng Wang, and Shrikanth Narayanan. Movieclip: Visual scene recognition in movies. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2082–2091, 2023.

[6] Agneet Chatterjee, Rahim Entezari, Maksym Zhuravinskyi, Maksim Lapin, Reshinth Adithyan, Amit Raj, Chitta Baral, Yezhou Yang, and Varun Jampani. Stable cinemetrics : Structured taxonomy and evaluation for professional video generation, 2025.

[7] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhiheng Xu, Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. Skyreels-v2: Infinite-length film generative model, 2025.

[8] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025.

[9] LMDeploy Contributors. Lmdeploy: A toolkit for compressing, deploying, and serving llm. `https://github.com/InternLM/lmdeploy`, 2023.

[10] Robin Courant, Christophe Lino, Marc Christie, and Vicky Kalogeiton. High-level features for movie style understanding. *Le Centre pour la Communication Scientifique Directe - HAL - Diderot,Le Centre pour la Communication Scientifique Directe - HAL - Diderot*, Oct 2021.

[11] Karan Dalal, Daniel Koceja, Gashon Hussein, Jiarui Xu, Yue Zhao, Youjin Song, Shihao Han, Ka Chun Cheung, Jan Kautz, Carlos Guestrin, Tatsunori Hashimoto, Sanmi Koyejo, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training, 2025.

[12] Jiayi Gao, Zijin Yin, Changcheng Hua, Yuxin Peng, Kongming Liang, Zhanyu Ma, Jun Guo, and Yang Liu. Conmo: Controllable motion disentanglement and recomposition for zero-shot motion transfer, 2025.

[13] Yueying Gao, Dongliang Chang, Bingyao Yu, Haotian Qin, Lei Chen, Kongming Liang, and Zhanyu Ma. Fakereasoning: Towards generalizable forgery detection and reasoning. *arXiv preprint arXiv:2503.21210*, 2025.

[14] Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Michael Baldridge, and Radu Soricut. ImageInWords: Unlocking Hyper-Detailed Image Descriptions. In *Proc. EMNLP 2024*, pages 93–127, Miami, Florida, USA, 2024. Association for Computational Linguistics.

[15] Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

[16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024.

[17] Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation, 2025.

[18] Haoyang Huang, Guoqing Ma, Nan Duan, Xing Chen, Changyi Wan, Ranchen Ming, Tianyu Wang, Bo Wang, Zhiying Lu, Aojie Li, Xianfang Zeng, Xinhao Zhang, Gang Yu, Yuhe Yin, Qiling Wu, Wen Sun, Kang An, Xin Han, Deshan Sun, Wei Ji, Bizhu Huang, Brian Li, Chenfei Wu, Guanzhe Huang, Huixin Xiong, Jiaxin He, Jianchang Wu, Jianlong Yuan, Jie Wu, Jiashuai Liu, Junjing Guo, Kaijun Tan, Liangyu Chen, Qiaohui Chen, Ran Sun, Shanshan Yuan, Shengming Yin, Sitong Liu, Wei Chen, Yaqi Dai, Yuchu Luo, Zheng Ge, Zhisheng Guan, Xiaoniu Song, Yu Zhou, Binxing Jiao, Jiansheng Chen, Jing Li, Shuchang Zhou, Xiangyu Zhang, Yi Xiu, Yibo Zhu, Heung-Yeung Shum, and Daxin Jiang. Step-video-ti2v technical report: A state-of-the-art text-driven image-to-video generation model, 2025.

[19] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 709–727, Cham, 2020. Springer International Publishing.

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

[21] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models, 2024.

[22] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025.

[23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023.

[24] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024.

[25] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025.

[26] Teng Li, Guangcong Zheng, Rui Jiang, Shuigenzhan, Tao Wu, Yehao Lu, Yining Lin, and Xi Li. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control, 2025.

[27] 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.

[28] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[29] Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, Rushikesh Zawar, Xue Bai, Yilun Du, Chuang Gan, and Deva Ramanan. Towards understanding camera motions in any video, 2025.

[30] Chang Liu, Armin Shmilovici, and Mark Last. Mnd: A new dataset and benchmark of movie scenes classified by their narrative function. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 610–626, Cham, 2023. Springer Nature Switzerland.

[31] Hongbo Liu, Jingwen He, Yi Jin, Dian Zheng, Yuhao Dong, Fan Zhang, Ziqi Huang, Yinan He, Yangguang Li, Weichao Chen, Yu Qiao, Wanli Ouyang, Shengjie Zhao, and Ziwei Liu. Shotbench: Expert-level cinematic understanding in vision-language models, 2025.

[32] Zhihang Liu, Chen-Wei Xie, Bin Wen, Feiwu Yu, Jixuan Chen, Boqiang Zhang, Nianzu Yang, Pandeng Li, Yinglu Li, Zuan Gao, Yun Zheng, and Hongtao Xie. What is a good caption? a comprehensive visual caption benchmark for evaluating both correctness and thoroughness, 2025.

[33] Yibo Miao, Yifan Zhu, Lijia Yu, Jun Zhu, Xiao-Shan Gao, and Yinpeng Dong. T2vsafetybench: Evaluating the safety of text-to-video generative models. 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 63858–63872. Curran Associates, Inc., 2024.

[34] OpenAI. Gpt-4 technical report, 2024.

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[36] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena

Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2025.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[39] Anyi Rao, Jiaze Wang, Linning Xu, Xuekun Jiang, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A unified framework for shot type classification based on subject centric lens. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 17–34, Cham, 2020. Springer International Publishing.

[40] Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, Feng Cheng, Feilong Zuo Xuejiao Zeng, Ziyan Yang, Fangyuan Kong, Zhiwu Qing, Fei Xiao, Meng Wei, Tuyen Hoang, Siyu Zhang, Peihao Zhu, Qi Zhao, Jiangqiao Yan, Liangke Gui, Sheng Bi, Jiashi Li, Yuxi Ren, Rui Wang, Huixia Li, Xuefeng Xiao, Shu Liu, Feng Ling, Heng Zhang, Houmin Wei, Huafeng Kuang, Jerry Duncan, Junda Zhang, Junru Zheng, Li Sun, Manlin Zhang, Renfei Sun, Xiaobin Zhuang, Xiaojie Li, Xin Xia, Xuyan Chi, Yanghua Peng, Yuping Wang, Yuxuan Wang, Zhongkai Zhao, Zhuo Chen, Zuquan Song, Zhenheng Yang, Jiashi Feng, Jianchao Yang, and Lu Jiang. Seaweed-7b: Cost-effective training of video generation foundation model, 2025.

[41] Gabriel S. Simões, Jônatas Wehrmann, Rodrigo C. Barros, and Duncan D. Ruiz. Movie genre classification with convolutional neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 259–266, 2016.

[42] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

[43] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4631–4640, 2016.

[44] Gemma Team. Gemma 3 technical report, 2025.

[45] Kimi Team. Kimi-vl technical report, 2025.

[46] Llama3 Team. The llama 3 herd of models, 2024.

[47] Phi-3 Team. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[48] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[49] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, New York, NY, USA, 2024. Association for Computing Machinery.

[50] WanTeam, :, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025.

[51] Runpu Wei, Zijin Yin, Shuo Zhang, Lanxiang Zhou, Xueyi Wang, Chao Ban, Tianwei Cao, Hao Sun, Zhongjiang He, Kongming Liang, and Zhanyu Ma. Omnieraser: Remove objects and their effects in images with paired video-frame data, 2025.

[52] Shuchen Weng, Haojie Zheng, Zheng Chang, Si Li, Boxin Shi, and Xinlong Wang. Audio-sync video generation with multi-stream temporal control, 2025.

[53] Wikipedia. Cinematography, April 2025. Page Version ID: 1286012571.

[54] Junfei Xiao, Ceyuan Yang, Lvmin Zhang, Shengqu Cai, Yang Zhao, Yuwei Guo, Gordon Wetzstein, Maneesh Agrawala, Alan Yuille, and Lu Jiang. Captain cinema: Towards short movie generation, 2025.

[55] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025.

[56] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone, 2024.

[57] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Ming Gong, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. NUWA-XL: Diffusion over diffusion for eXtremely long video generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1309–1320, Toronto, Canada, July 2023. Association for Computational Linguistics.

[58] Shenghai Yuan, Jinfa Huang, Yongqi Xu, YaoYang Liu, Shaofeng Zhang, Yujun Shi, Rui-Jie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[59] Zihao Yue, Qi Zhang, Anwen Hu, Liang Zhang, Ziheng Wang, and Qin Jin. Movie101: A new movie understanding benchmark. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4669–4684, Toronto, Canada, July 2023. Association for Computational Linguistics.

[60] Zihao Yue, Yepeng Zhang, Ziheng Wang, and Qin Jin. Movie101v2: Improved movie narration benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17081–17095, Vienna, Austria, July 2025. Association for Computational Linguistics.

[61] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3r: A simple approach for estimating geometry in the presence of motion. In *The Thirteenth International Conference on Learning Representations*, 2025.

[62] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation, 2025.

[63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023.

[64] Ruihan Zhang, Borou Yu, Jiajian Min, Yetong Xin, Zheng Wei, Juncheng Nemo Shi, Mingzhen Huang, Xianghao Kong, Nix Liu Xin, Shanshan Jiang, Praagya Bahuguna, Mark Chan, Khushi Hora, Lijian Yang, Yongqi Liang, Runhe Bian, Yunlei Liu, Isabela Campillo Valencia, Patricia Morales Tredinick, Ilia Kozlov, Sijia Jiang, Peiwen Huang, Na Chen, Xuanxuan Liu, and Anyi Rao. Generative ai for film creation: A survey of recent advances, 2025.

[65] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[66] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model, 2024.

[67] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs, 2024.

[68] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024.

[69] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models, 2023.

[70] Howard Zhou, Tucker Hermans, Asmita V. Karandikar, and James M. Rehg. Movie genre classification via scene categorization. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 747–750, New York, NY, USA, 2010. Association for Computing Machinery.

[71] Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, et al. Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 56–66, 2025.

[72] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: In the inrtoduction section, we clearly outlined our main contributions across three key aspects.

   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: In the limitations section, we discussed the limitations of this work, specifically the absence of ground-truth camera trajectories for the collected movie clips and potential inconsistency of the annotation process

   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: No, we don't include theoretical results.

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: In experiment settings section, we showed our detailed experiment settings.

   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 provided our dataset which includes data links and annotations and we also provide the code link.

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: Yes, we fully showed our metrics and experimet settings.

Guidelines: On

- 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: Please see the error bar part in the Appendix E.

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: In experiment settings section, we fully showed our computer resources.

   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: Yes, we strictly conduct in the paper conform.

   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: [NA]

    Justification: No, there is no societal impact.

    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: [Yes]

    Justification: We fully respect the copyright of all films and do not use any clips for commercial purposes. Instead of distributing or hosting video content, we only provide links to publicly available, authorized sources (e.g., official studio or distributor channels). This approach ensures that we neither infringe on copyright nor redistribute protected materials. All assets are credited to their original rights holders, and our use of these links falls under fair-use provisions for non-commercial, academic research.

    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 fully respect the copyright of all films and do not use any clips for commercial purposes. Instead of distributing or hosting video content, we only provide links to publicly available, authorized sources (e.g., official studio or distributor channels). This approach ensures that we neither infringe on copyright nor redistribute protected materials. All assets are credited to their original rights holders, and our use of these links falls under fair-use provisions for non-commercial, academic research.

    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: The paper submitted a dataset. All the assets are well documented and have a copyright statement in appendix H.

    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: In appendix B we showed our annotation instruction.

    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: [NA]

    Justification: No, the paper does not involve research with human subjects.

    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 use LLM to process dataset, please see the method part. The usage is under human supervision.

    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.

Appendix

A Benchmark Statistical Information

As shown in Figure 7, our benchmark spans 93 years of cinematic history (1931–2024) and includes 48 distinct film genres, from classic Hollywood dramas to contemporary global art house cinema. This cross-decade temporal coverage and genre diversity capture the evolution of cinematographic styles and technical innovations, from the early days of monochrome filmmaking to modern high-definition digital cinematography. By encompassing films across eras and genres, the dataset avoids bias toward specific stylistic trends, providing a robust foundation for evaluating MLLMs' ability to generalize across diverse visual and narrative contexts.

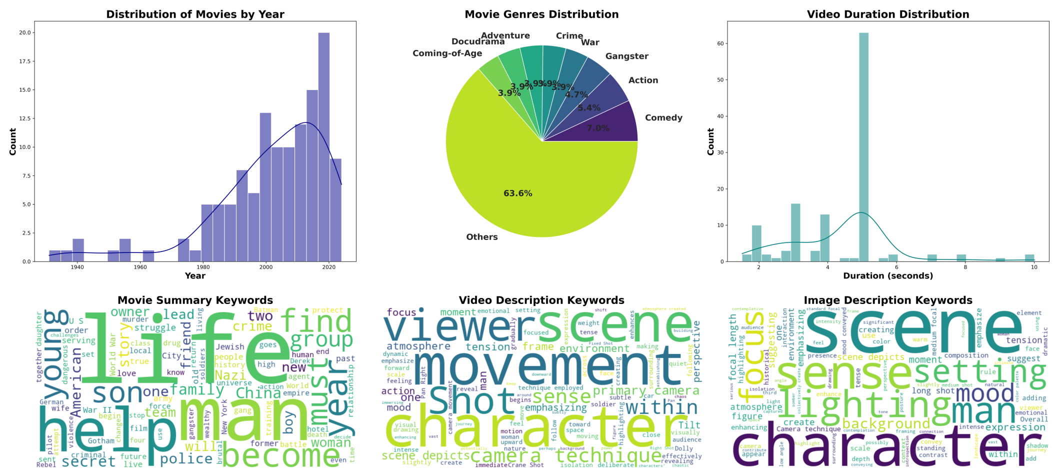


Figure 7: Statistical and semantic overview of the CineTechBench.

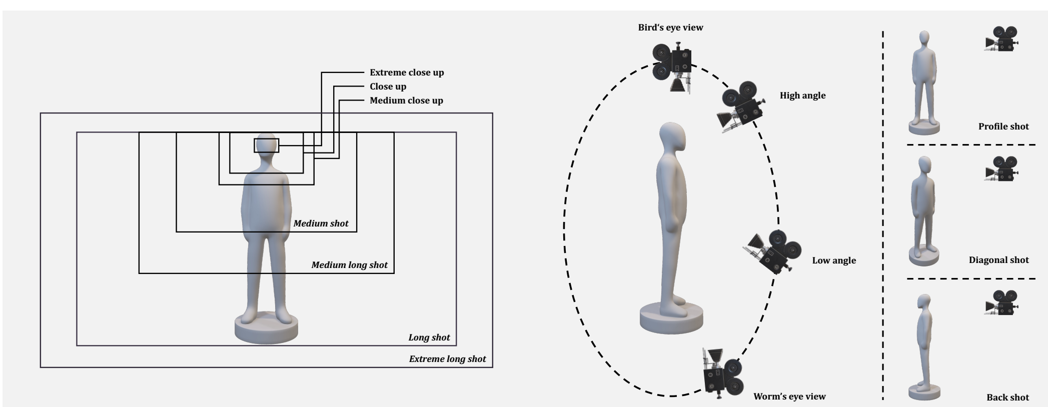


Figure 8: Illustration of categories in the angle and scale dimension.

B Annotation Process Detail

B.1 Annotation Instruction for Description Refinement

Overall Workflow In this refine task, annotators are required to refine the descriptions generated by a large model for images or videos. The descriptions are initially generated based on specific keywords representing various cinematographic techniques. The purpose of this instruction is to ensure consistency, accuracy, and clarity in the annotation process.

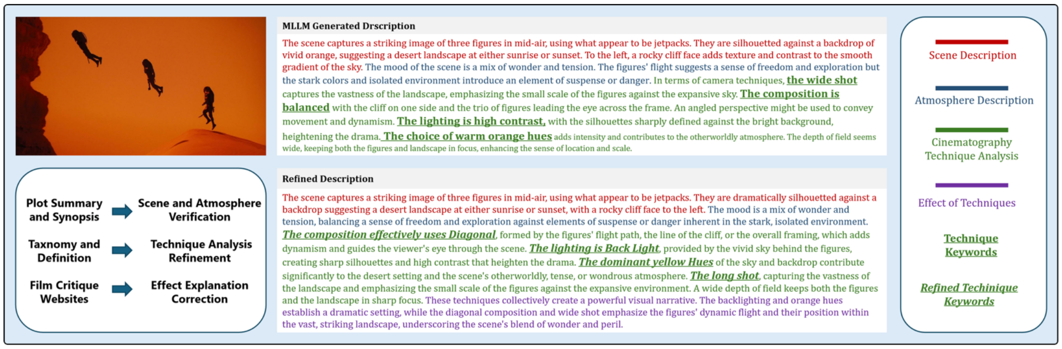


Figure 9: An annotation refine example for MLLM generated description.

1. **Description Structure:** These generated descriptions generally follow a standard structure:
    - **Scene Description:** A general depiction of the visual scene.
    - **Atmosphere Description:** A brief description of the mood or feeling conveyed.
    - **Cinematographic Technique Analysis:** An analysis of the specific cinematographic techniques identified in the scene.
    - **Effect of Techniques:** An explanation of the impact of these techniques on the visual experience. Depending on the context, the effect may be integrated within the technique analysis or provided as a separate section at the end.
2. **Scene and Atmosphere Verification:**
    - Review the scene and atmosphere descriptions.
    - Cross-reference with the context or plot summary of the film to ensure accuracy.
    - Make necessary corrections for clarity, factual accuracy, and alignment with the scene.
3. **Technique Analysis Refinement:**
    - Verify that the description covers all relevant cinematographic techniques.
    - Remove any unnecessary or inaccurate techniques.
    - Ensure that all technical terms align with the predefined standardized taxonomy.
4. **Effect Explanation Correction:**
    - Refine the explanation of the effects generated by the identified techniques.
    - Cross-check with film critique websites to ensure the effects are consistent with expert interpretations.
5. **Final Review:**
    - Ensure the description is coherent, grammatically correct, and accurately represents the visual content.
    - Submit the refined description.

Quality Control

- Each refined description will be reviewed by a senior annotator for quality assurance.
- Descriptions failing to meet the specified standards will be sent back for correction.

An example refine process for MLLM generated description is shown in Figure 9.

B.2 Annotation Interface

Figure 10 illustrates an example of our labeling interface, The tags displayed beneath the image represent accurate dimension labels refined by experts. Annotators can reference these tags to quickly identify the cinematographic technique keywords and refine the corresponding descriptions.

B.3 Crowdsourcing Compensation

Our annotation process was conducted by a team of project authors, skilled students, and professional experts, with all external contributors receiving fair compensation. Three students handled the primary annotation tasks at competitive per-item rates (ranging from 5 to 20 CNY) scaled by task difficulty and set above typical student wages. Additionally, two professional cinematographers provided expert oversight, refined our taxonomy, and served as final arbiters, each receiving a 2,000 CNY consultancy fee for their significant contribution.

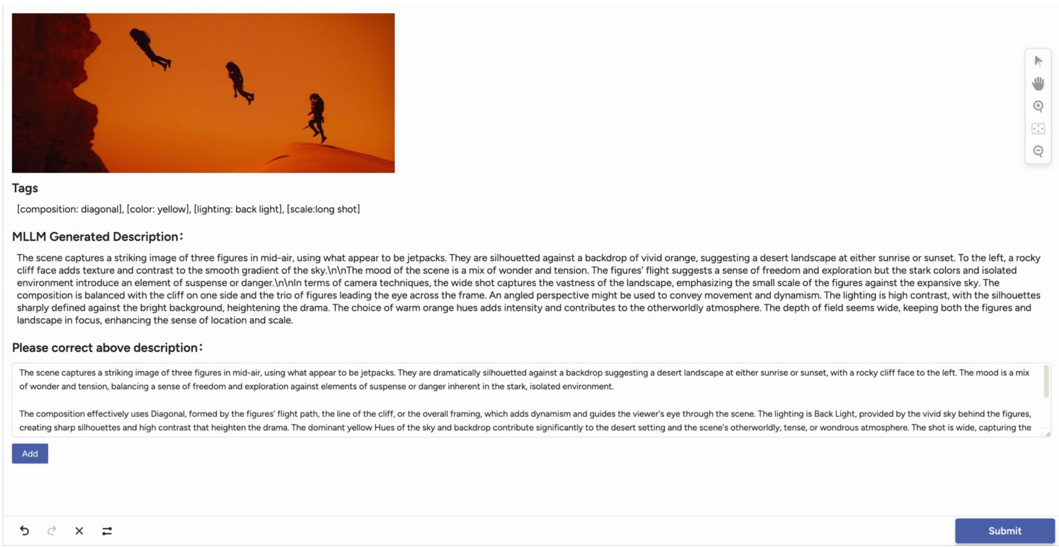


Figure 10: An example label interface.

C Evaluation Metrics

C.1 Description Evaluation Metrics

Inspired by the CAPability benchmark [32], which proposes a comprehensive framework to evaluate the correctness and thoroughness of visual captions, we adopt a similar metric design to assess the descriptive quality of cinematographic techniques in our dataset.

To determine whether a caption correctly addresses a specific dimension, we follow the classification scheme proposed by CAPability [32]. Each caption is categorized into one of the following three cases:

- **Miss**: The caption does not mention any information relevant to the dimension;
- **Positive**: The caption includes information related to the dimension, and the content is consistent with the human annotation;
- **Negative**: The caption mentions the dimension, but the content is incorrect compared to the annotation.

Based on this categorization, we compute four quantitative metrics to evaluate model performance:

- **Hit Rate (HR)**: Measures whether a caption mentions a particular dimension, regardless of correctness. It reflects the referential completeness:
$$\mathrm{HR} = \frac{|\mathcal{S}_{\mathrm{All}} - \mathcal{S}_{\mathrm{Miss}}|}{|\mathcal{S}_{\mathrm{All}}|}$$
- **Precision (AP)**: The proportion of correctly described dimensions among all mentioned:
$$\mathrm{Precision} = \frac{|\mathcal{S}_{\mathrm{Pos}}|}{|\mathcal{S}_{\mathrm{All}} - \mathcal{S}_{\mathrm{Miss}}|}$$

```
You are a cinematography technique analysis expert specializing in evaluating the accuracy of image captions. Please carefully analyze the user-provided caption and complete the task according to the metric specified.

Given an image caption, your task is to determine which kind of {task} is included in the caption.
Image Caption:
"{caption}"
Please analyze the image caption and classify the descriptions of {task} into the following categories: {category1, category2, ...}
Here are the explanations of each category: {definition}

If the caption explicitly mentions one or some of the above {task} categories, write the result of the categories with a python list format into the 'pred' value of the json string. You should only search the descriptions about the {task}. If there is no description of the {task} in the image caption or the description does not belong to any of the above categories, write 'N/A' into the 'pred' value of the json string.

Output a JSON formed as:
{"pred": "put your predicted category as a python list here", "reason": "give your reason here"}
DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only output the JSON. Do not add Markdown syntax. Output:
```

```
You are a video analysis expert specializing in evaluating movement in video captions.

Given a video description and a specified camera movement, your task is to evaluate whether the movement is accurately reflected in the description, and explain why.
Video description:
"{caption}"
Proper camera movement: "{annotation}"

Here are the explanations of each category: {definition}

Please provide a justification for your judgment, with particular attention to the sequence and types of camera movements involved.
Give score of 0 if there is no mention of the movement in the caption.
Give score of 1 if the description describes the movement correctly.
Give score of -1 if the caption describes the movement incorrectly.

Output a JSON formed as:
{"score": put your score here, "reason": "give your reason here"}
DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only output the JSON. Do not add Markdown syntax. Output:
```

Figure 11: Prompt template used for static dimension evaluation (e.g., *Scale*, *Angle*, etc.) and dynamic dimension evaluation (*Camera Movement*).

- **Recall (AR)**: The proportion of correctly described dimensions among all ground-truth annotations:

$$\text{Recall} = \frac{|\mathcal{S}_{\text{Pos}}|}{|\mathcal{S}_{\text{All}}|}$$

- **F1-score (F1)**: The harmonic mean of precision and recall, used as the main metric for overall capability:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Note: When the hit rate (HR) reaches 100%, i.e., every caption mentions the target dimension, the average precision (AP), average recall (AR), and F1-score become mathematically identical.

While CAPability originally defines 12 static and dynamic visual dimensions, we adapt this metric suite to assess the understanding and generation of **cinematographic technique descriptions**. Specifically, we evaluate performance across 7 tailored dimensions: six static dimensions—*Scale, Angle, Composition, Colors, Lighting, Focal Lengths*—and one dynamic dimension—*Camera Movement*. The generated descriptions are compared to human-annotated references to compute the metrics, thereby providing an objective measurement of a model's expressive capacity in film-oriented tasks.

To automate this evaluation process, we use `GPT-4.1-nano` to assess each generated caption with respect to the ground-truth annotations. Specifically, we design one prompt template for evaluating static dimensions (*Scale*, *Angle*, *Composition*, *Colors*, *Lighting*, *Focal Lengths*), and another distinct template for the dynamic dimension (*Camera Movement*). These prompt templates guide the `GPT-4.1-nano` to determine whether the relevant dimension in the caption should be categorized as **Positive**, **Negative**, or **Miss**. Detailed prompt formats are provided in Figure 11.

C.2 Camera Movement Evaluation Metrics

Formally, we denote the i^{th} frame relative camera-to-world matrix of ground truth as $\left\{R_i^{3\times3}, T_i^{3\times1}\right\}$, and that of generated video as $\left\{\tilde{R}_i^{3\times3}, \tilde{T}_i^{3\times1}\right\}$. We calculate camera rotation errors by the relative angle between generated videos and ground truths in radians for rotation accuracy and we calculated translation error (TransErr) measures the cumulative difference between the predicted and ground truth camera translations across a trajectory:

$$\text{RotErr} = \sum_{i=1}^{n} \arccos \frac{\operatorname{tr}\left(\tilde{R}_i R_i^{\text{T}}\right) - 1}{2}, \quad \text{TransErr} = \sum_{i=1}^{n} \left\| \frac{\tilde{T}_i}{\tilde{s}_i} - \frac{T_i}{s_i} \right\|_2 \tag{1}$$

where $\tilde{T}_i$ and T_i are the predicted and ground truth translations at timestep i, and $\tilde{s}_i$ and s_i are their respective scale factors. For relative TransErr, we perform scene scale normalization on the camera positions of each video clip. The scene scale of generated video $\tilde{s}_i$ and ground truth s_i are individually calculated as the $\mathcal{L}_2$ distance from the first camera to the farthest one for each video clip. For absolute TransErr, we normalize both the video clip to the scene scale of ground truth video, i.e. $\tilde{s}_i = s_i$. CamMC consider camera translation and rotation error at the same time by directly calculating $\mathcal{L}_2$ distance on camera-to-world matrices:

$$\text{CamMC} = \sum_{i=1}^{n} \left\| \left[\tilde{R}_i \,\middle|\, \frac{\tilde{T}_i}{\tilde{s}_i} \right]^{3\times4} - \left[R_i \,\middle|\, \frac{T_i}{s_i} \right]^{3\times4} \right\|_2 . \tag{2}$$

We further use CLIP frame similarity [37] to evaluate the semantic reconstruction performance:

$$\text{CLIP-IS} = \sum_{i=1}^{N} \frac{f_{\text{image}}\left(x_i\right) \cdot f_{\text{image}}\left(\tilde{x_i}\right)}{\|f_{\text{image}}(x_i)\| \cdot \|f_{\text{image}}(\tilde{x_i})\|}, \tag{3}$$

where $\tilde{x_i}$ and x_i are the i^{th} frame of generated video clip and original video clip. Since some commercial video generation models do not allow setting the number of generated frames, we downsample longer videos to match the same frame count before calculating the above metrics.

D Experiment Settings

For commercial MLLMs, we access them via their official APIs. For open-source MLLMs, we deploy them for online inference using SGLang [67], vLLM [23] and LMDeploy [9] frameworks. To evaluate camera movement understanding, we adopt a multi-image input approach for MLLMs that do not support video input. All experiments are conducted on 2 $\times$ Tesla A800 80G GPUs. For all commercial video generation models, we set the generation duration as 5 seconds. For all open-source video generation models, we set the generation frame counts same as the original movie clips.

E Extra Results

Table 1 presents the sub-category accuracy results for question-answering understanding in the angle and lighting dimensions. Unlike other static cinematogrphic technique dimensions, the angle and lighting dimensions are inherently more complex due to their multi-dimensional nature, each encompassing multiple subcategories that introduce significant visual variability. The angle dimension is divided into two main perspectives: vertical and horizontal. The vertical perspective includes four subcategories: high angle, low angle, bird's eye view, and worm's eye view. The horizontal perspective comprises three subcategories: diagonal shot, profile shot, and back shot. The lighting dimension is categorized into three aspects: intensity, quality, and direction. Intensity is divided into high key and low key lighting. Quality is represented by hard light and soft light, while direction is further classified into side lighting, back lighting, and top lighting.

Angle Dimension Among all commercial MLLMs, GPT-4o demonstrates a superior performance (83. 15%) in the vertical perspectives, while achieving the second-highest (80. 65%) in the horizontal perspectives. In contrast, Gemini-2.5-Pro outperforms others in the horizontal perspective (87.10%), while maintaining a strong second position in the vertical perspective (82.03%). Regarding open-source MLLMs, Qwen2.5-Omini and InternVL3 demonstrate the highest accuracy (64.04%) in the vertical perspective, with Kimi-VL securing the second-highest (60.67%). Kimi-VL leads in the horizontal perspective (79.97%), while Qwen2.5-VL, InternVL2.5, and InternVL3 share the second-highest performance (74.19%). These results indicate that, among commercial MLLMs, both vertical and horizontal perspectives are recognized with comparable accuracy. In contrast, for open-source MLLMs, vertical perspectives are generally more challenging for the models to accurately identify, indicating a potential area for further optimization in recognizing fine-grained angle differences.

Lighting Dimension Among commercial MLLMs, Gemini-2.0-Flash achieves the highest accuracy (93.75%) in the intensity category, followed closely by GPT-4o (90.62%). In the quality category, Qwen-VL-Plus stands out with the best performance (78.26%), with Gemini-2.0-Flash ranking second (71.74%). However, in the direction category, all models exhibit a significant drop in accuracy, with GPT-4o outperforms others (53.12%), while Qwen-VL-Plus and Gemini-2.5-Pro share the second-best performance (50.00%). In open-source MLLMs, LLaVA-OneVision demonstrates strong performance in the intensity category (81.25%), with InternVL2.5 securing the second position (78.12%). For quality, Qwen2.5-VL achieves the highest accuracy (76.09%), followed by Phi3.5 (65.22%). The direction category again shows a clear performance drop. InternVL3 attains the best performance (46.88%), with Kimi-VL following closely (43.75%). These findings confirm that the direction category in the lighting dimension is consistently the most challenging for both commercial and open-source models. This can be attributed to the complex nature of light direction recognition, where even subtle changes in lighting angles can dramatically alter the visual appearance of a scene.

Table 6 shows the CAPability performance on seven dimensions of cinematographic technique description generation. In the description generation task among commercial models, Gemini-2.5-Pro and GPT-4o stand out significantly, achieving a clear lead over other models. Specifically, Gemini-2.5-Pro secures 14 first-place rankings and 2 second-place rankings, while GPT-4o achieves 10 first-place rankings and 8 second-place rankings, demonstrating their superior descriptive capabilities. Remarkably, InternVL3 emerges as the best-performing model among open-source models, with 12 first-place rankings and 6 second-place rankings, making it the strongest contender in this category. Notably, several of its results are comparable to those of the top commercial models, Gemini-2.5-Pro and GPT-4o. This performance highlights InternVL3's exceptional capability in description generation.

Figure 12 presents the average performance of hit rate (HR) and F1 score on seven dimensions of cinematographic technique description generation. In the hit rate (HR) chart (left), the models exhibit consistently high accuracy across six dimensions, all exceeding 80%. However, a notable decline is observed in the Movement dimension (29.83%), indicating that recognizing and describing dynamic actions remains a significant challenge for these models. In contrast, the F1 Score chart (right) reveals a starkly different trend. While HR values remain high across most dimensions, the F1 scores are significantly lower, ranging about from 30% to 50% across all dimensions. This substantial disparity between HR and F1 score suggests that although models are capable of recognizing certain cinematographic features (as indicated by high HR), they struggle to generate precise and consistent descriptions of these features. Such a gap highlights a critical issue in the models' ability to translate visual recognition into accurate textual descriptions, reflecting limitations in their descriptive generation capabilities.

Error Bars We conducted an error bar test on six models (GLM-4V, Gemini-2.0-Flash, Qwen2.5-VL-7B, InternVL3-8B, LLaVA-OneVision-7B, Wan2.1-FLF2V-14B), testing each model three times on the corresponding tasks to calculate the standard deviation of three trials. The observed average standard deviations were 2.67% (Acc) for video QA, 1.59% (Acc) for image QA, 1.21% (F1) for description, 2.21% (CamMC) for camera movement reconstruction, which reflect the stability and reliability of our evaluation pipeline.

F Limitation

Camera Trajectory Estimation Tools One limitation of our benchmark is the lack of ground-truth camera trajectories for the collected movie clips. Acquiring such data is extremely challenging, as professional camera motion metadata is rarely publicly available. To approximate the motion, we employ open-source camera pose estimation tools to reconstruct trajectories from the video clips. However, these methods often introduce inaccuracies due to complex cinematographic factors such as dynamic scenes, motion blur, and non-rigid object motion. This limits the precision of motion-related evaluations, and highlights the need for more accurate and robust trajectory estimation techniques to support fine-grained analysis in future work.

Annotation Process Our annotations rely on trained human experts manually labeling each still image and video clip across seven cinematographic dimensions. While this ensures high semantic fidelity, it also introduces subjectivity and potential inconsistency across annotators. Even with detailed guidelines and cross-checking protocols, subtle distinctions—such as grading "medium" versus "close" shot scales or identifying nuanced lighting contrasts—can vary between annotators.

Table 5: Sub-category accuracy of various MLLMs on angle and lighting question answering understanding. The best and second best results are highlighted by blue and green respectively.

MLLMs	Params	Angle		Lighting		
		Vertical	Horizontal	Intensity	Quality	Direction
Commercial						
GLM-4V-Plus [15]	–	67.42	74.19	71.88	58.70	37.50
Qwen-VL-Plus	–	74.16	70.97	65.62	78.26	50.00
Gemini-2.0-Flash	–	76.40	67.74	93.75	71.74	46.88
Gemini-2.5-Pro	–	82.03	87.10	71.88	65.22	50.00
Doubao-1.5-vision-pro	–	67.42	70.97	84.38	58.70	37.50
GPT-4o [34]	–	83.15	80.65	90.62	69.57	53.12
Open-source						
Kimi-VL [45]	3B	60.67	79.97	62.50	58.70	43.75
Phi3.5 [47]	4B	51.69	41.94	62.50	65.22	37.50
Gemma3-it [44]	4B	41.57	54.84	53.12	63.04	37.50
Qwen2.5-VL [2]	7B	57.30	74.19	65.62	76.09	40.62
Qwen2.5-Omni [55]	7B	64.04	70.97	59.38	52.17	34.38
LLaVA-OneVision [25]	7B	53.93	54.84	81.25	52.17	31.25
LLaVA-NeXT [24]	8B	37.08	58.06	65.62	50.00	15.62
MinCPM-V-2.6 [56]	8B	58.43	54.84	75.00	50.00	28.12
InternVL2.5 [8]	8B	59.55	74.19	78.12	47.83	34.38
InternVL3 [72]	8B	64.04	74.19	68.75	56.52	46.88
Llama-3.2-Vision [46]	11B	43.82	61.29	53.12	47.83	34.38

Table 6: CAPability performance of different MLLMs' on seven dimensions of cinematographic technique description generation.

	Metrics	GLM-4V-Plus	Qwen-VL-Plus	Doubao-v1.5-vision-pro	GPT-4o	Gemini-2.0-Flash	Gemini-2.5-Pro	Kimi-VL	Phi3.5	gemma3-it	Qwen2.5-VL	Qwen2.5-Omni	LLaVA-OneVision	LLaVA-NeXT	LLaVA-NeXT-Video	MinCPM-V-2.6	InternVL2.5	InternVL3	Llama-3.2
AG	HR	85.37	73.17	67.44	82.05	55.81	59.52	97.62	67.50	100.00	90.48	92.68	85.71	90.00	N/A	83.33	95.24	93.02	87.50
	AP	60.00	60.00	62.07	71.88	70.83	88.00	56.10	66.67	60.47	47.37	52.63	52.78	47.22	N/A	51.43	47.50	65.00	54.29
	AR	51.22	43.90	41.86	58.97	39.53	52.38	54.76	45.00	60.47	42.86	48.78	45.24	42.50	N/A	42.86	45.24	60.47	47.50
	F1	55.26	50.70	50.00	64.79	50.75	65.67	55.42	53.73	60.47	45.00	50.63	48.72	44.74	N/A	46.75	46.34	62.65	50.67
SC	HR	94.95	90.91	100.00	90.91	100.00	100.00	90.91	78.79	96.97	98.99	96.97	86.87	85.57	N/A	62.63	96.97	98.99	93.94
	AP	45.74	41.11	44.44	45.56	42.42	38.38	53.33	34.62	34.38	42.86	42.71	40.70	45.78	N/A	41.94	43.75	45.92	39.78
	AR	43.43	37.37	44.44	41.41	42.42	38.38	48.48	27.27	33.33	42.42	41.41	35.35	39.17	N/A	26.26	42.42	45.45	37.37
	F1	44.56	39.15	44.44	43.39	42.42	38.38	50.79	30.51	33.85	42.64	42.05	37.84	42.22	N/A	32.30	43.08	45.69	38.54
CL	HR	97.83	100.00	100.00	100.00	100.00	97.87	100.00	97.83	97.96	100.00	95.35	100.00	93.88	N/A	97.96	100.00	100.00	100.00
	AP	55.56	47.83	43.18	72.00	52.17	50.00	55.32	42.22	39.58	60.00	51.22	50.00	52.17	N/A	39.58	53.19	63.04	47.73
	AR	54.35	47.83	43.18	72.00	52.17	48.94	55.32	41.30	38.78	60.00	48.84	50.00	48.98	N/A	38.78	53.19	63.04	47.73
	F1	54.95	47.83	43.18	72.00	52.17	49.46	55.32	41.76	39.17	60.00	50.00	50.00	50.53	N/A	39.17	53.19	63.04	47.73
CP	HR	100.00	100.00	100.00	100.00	100.00	100.00	100.00	97.67	100.00	100.00	97.75	100.00	98.86	N/A	100.00	100.00	100.00	100.00
	AP	32.58	31.46	32.58	33.71	30.34	46.07	32.58	32.14	23.60	29.21	22.99	32.95	32.18	N/A	32.18	35.95	40.45	15.91
	AR	32.58	31.46	32.58	33.71	30.34	46.07	32.58	31.39	23.60	29.21	22.47	32.95	31.82	N/A	32.18	35.95	40.45	15.91
	F1	32.58	31.46	32.58	33.71	30.34	46.07	32.58	31.77	23.60	29.21	22.73	32.95	32.00	N/A	32.18	35.95	40.45	15.91
LT	HR	85.53	92.11	89.47	98.68	92.11	94.74	96.05	90.67	90.79	94.67	89.33	93.42	94.74	N/A	94.74	94.74	96.05	97.33
	AP	40.00	34.29	42.65	44.00	32.86	38.89	42.47	32.35	30.43	46.48	34.33	35.21	25.00	N/A	34.72	45.83	45.20	41.10
	AR	34.21	31.58	38.16	43.42	30.26	36.84	40.79	29.33	27.63	44.00	30.67	32.90	23.68	N/A	32.90	43.42	43.42	40.00
	F1	36.88	32.88	40.28	43.71	31.51	37.84	41.61	30.77	28.97	45.20	32.39	34.01	24.32	N/A	33.78	44.59	44.30	40.54
FL	HR	100.00	72.22	100.00	97.22	100.00	100.00	94.12	84.93	100.00	98.61	92.65	86.77	90.14	N/A	82.09	92.75	100.00	97.02
	AP	48.57	36.54	52.78	40.00	50.69	60.27	42.19	38.71	41.10	56.34	44.44	38.98	34.38	N/A	32.73	43.75	52.94	63.08
	AR	48.57	26.39	52.78	38.89	50.69	60.27	39.71	32.88	41.10	55.56	41.18	33.82	30.99	N/A	26.87	40.58	52.94	61.19
	F1	48.57	30.64	52.78	39.44	50.69	60.27	40.91	35.56	41.10	55.94	42.75	36.22	32.59	N/A	29.51	42.10	52.94	62.12
CM	HR	27.34	40.62	23.44	34.38	50.00	69.53	N/A	11.72	38.28	24.22	36.72	16.41	16.54	35.16	18.75	20.31	20.31	23.44
	AP	68.57	86.54	93.33	90.91	79.69	83.15	N/A	73.33	83.67	83.87	72.34	76.19	80.95	77.78	87.50	88.46	76.92	56.67
	AR	18.75	35.16	21.88	31.25	39.84	57.81	N/A	8.59	32.03	20.31	26.56	12.50	13.39	27.34	16.41	17.97	15.62	13.28
	F1	29.45	50.00	35.44	46.51	53.12	68.20	N/A	15.38	46.33	32.70	38.86	21.48	22.97	40.46	27.63	29.87	25.97	21.52

Moreover, the intensive manual effort limits the overall scale of our dataset, constraining diversity in film styles, genres and time periods. Future work should explore semi-automated annotation pipelines, active learning, or consensus-driven schemes to improve diversity and scalability.

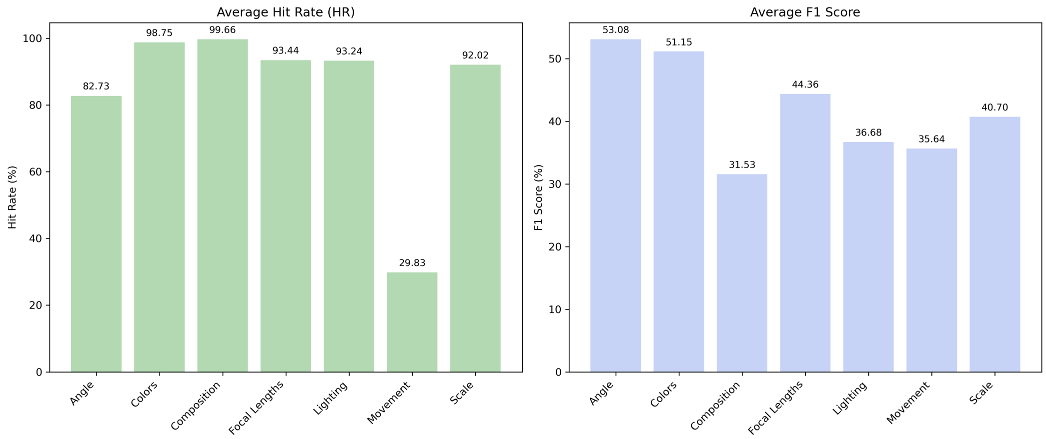


Figure 12: Average hit rate (HR) and F1 score of all MLLMs on seven dimensions of cinematographic technique description generation task.

Connection with Plots While our ultimate motivation is to enable the understanding of visual storytelling, the core contribution of CineTechBench is to provide models with the foundational capability to identify cinematic techniques and analyze their general atmospheric impact. We recognize that our current annotations do not forge the deeper connection between a technique and its specific plot or symbolic meaning, which we frame as an important area for future research.

G Visualization

G.1 Visualization of Cinematographic Technique Understanding

Figure 13 shows more visualization of the answers for the image question-answering task across all dimensions. Through these visualized cases, it is evident that color is the easiest dimension for models to recognize, achieving consistently high accuracy across all models. This result suggests that color information, being a highly distinctive and easily discernible visual feature, is effectively captured and processed by both commercial and open-source MLLMs. In contrast, focal length emerges as the most challenging dimension, where models struggle to achieve high accuracy. This difficulty likely arises from the subtle and complex visual cues associated with focal length, such as depth of field and background blur, which are less visually obvious than color differences. Among all evaluated models, GPT-4o and Gemini-2.5-Pro consistently outperform all other commercial and open-source models across most dimensions, maintaining a significant lead in accuracy. Despite a noticeable performance gap between commercial and open-source models, several open-source models, such as InternVL3 and Qwen2.5-Omini demonstrate impressive results. These models highlighting the potential of open-source MLLMs to close the performance gap with their commercial counterparts.

Also, more visualization of MLLM's answers on video question answering task and descriptions on image and video description generation task are shown in Figure 14. Through these visualized cases, it is evident that the video-based question-answering (QA) task is inherently more complex and challenging compared to the image-based QA task. This increased difficulty can be attributed to the dynamic nature of video content, where temporal information, motion, and scene transitions introduce additional layers of complexity that models must effectively process. Moreover, when comparing QA tasks to description generation tasks, the latter proves to be even more challenging. Generating accurate and comprehensive descriptions of cinematographic techniques in images or videos requires not only recognizing visual elements but also understanding their spatial and temporal relationships. Even models that perform well in perceptual tasks often struggle to generate precise and complete descriptions of cinematographic techniques. This difficulty is particularly pronounced in the context of cinematography, where subtle differences in angle, lighting, and composition can drastically alter the interpretation of a scene. As a result, achieving accurate and contextually appropriate description generation remains a significant challenge, even for models that demonstrate strong performance in other perception-based tasks.

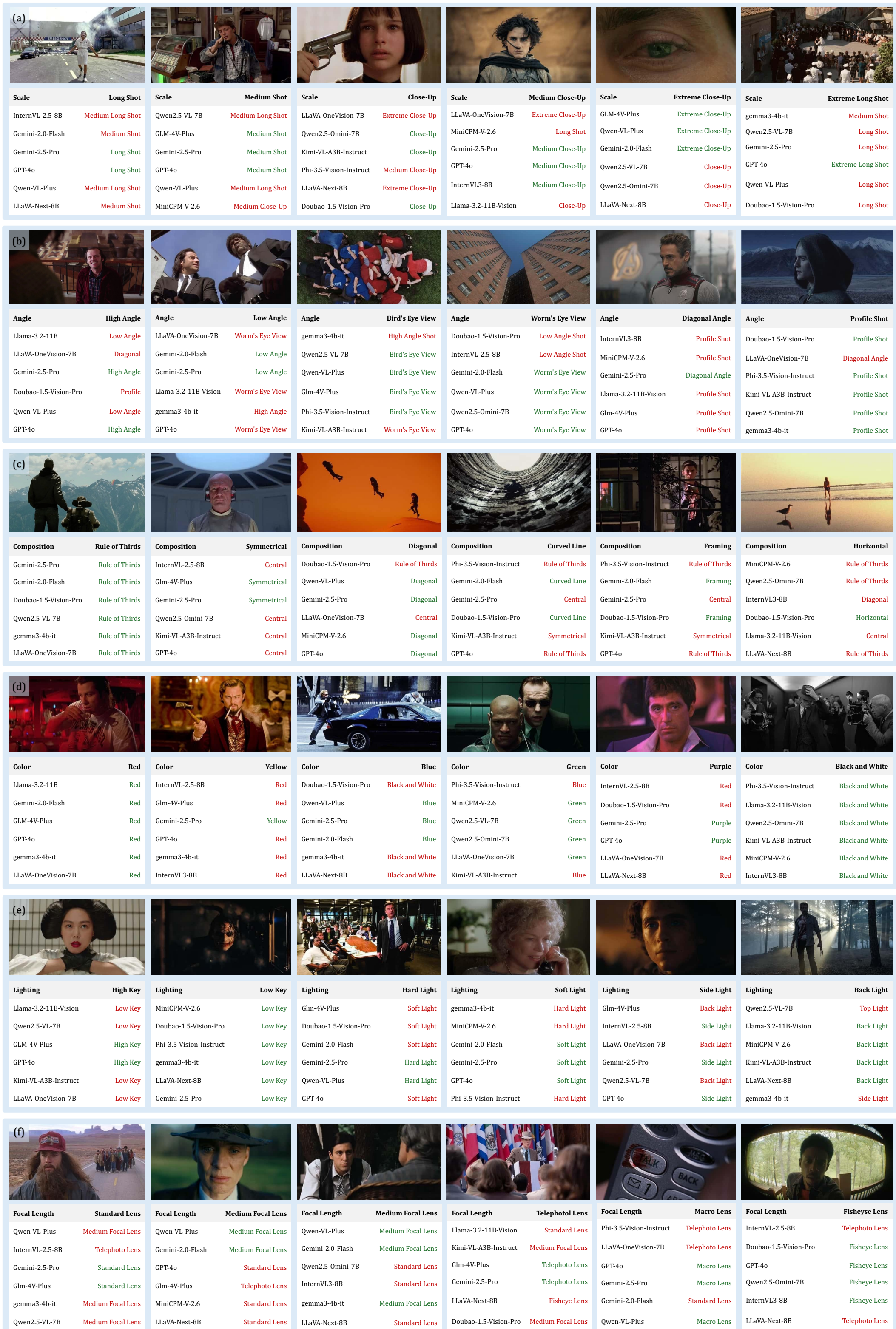


Figure 13: Visualization of MLLMs' answers on image cinematographic technique question answering task. The red text highlights the wrong answers and the green text highlights the correct answers.

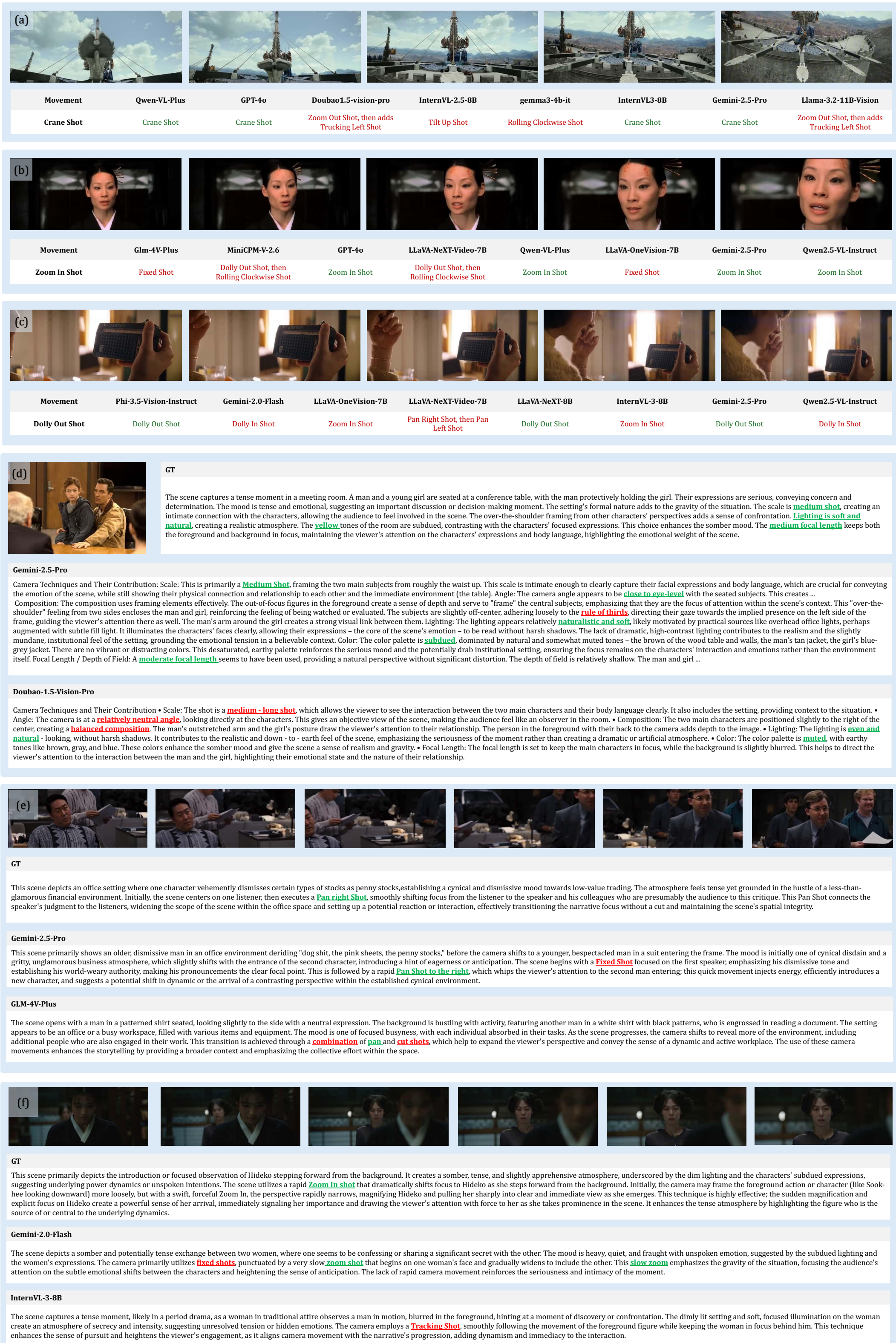


Figure 14: Visualization of MLLMs' answers on video question-answering task and generated descriptions on image and video description task. The red text highlights the wrong answers and the green text highlights the correct answers.

G.2 Visualization of Camera Movement Generation

As shown in Figure 15. The video generation models have a relatively good performance on simple camera movement, e.g., example (a) and a relatively bad performance on camera rotation, e.g., example (c), Gen4turbo and Wan2.1 didn't show qualified rotation sense.

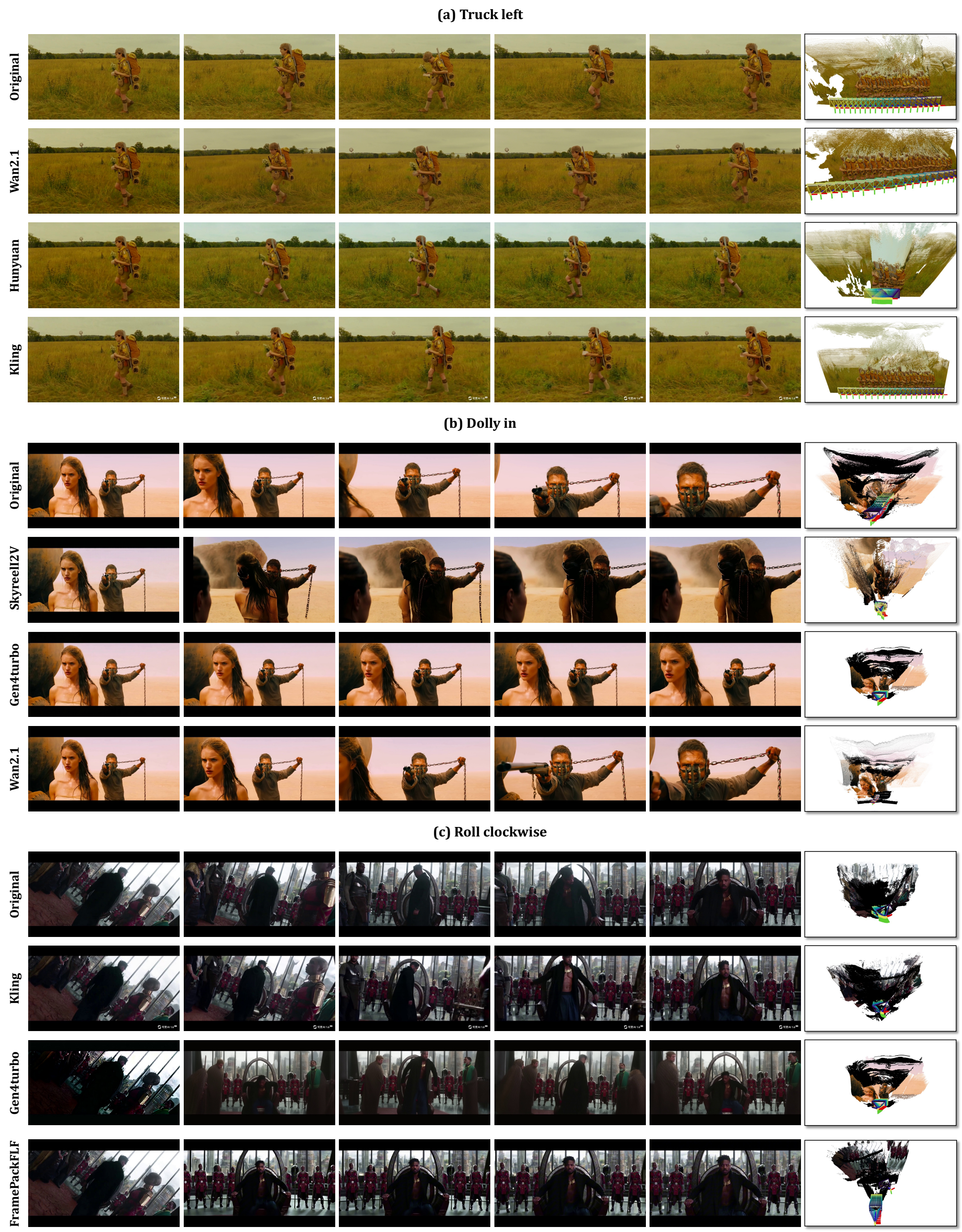


Figure 15: Generated movie clips by different video generation models and the corresponding camera trajectory estimated by Monst3r [61].

H Copyright

We fully respect the copyright of all films and do not use any clips for commercial purposes. Instead of distributing or hosting video content, we only provide links to publicly available, authorized sources (e.g., official studio or distributor channels). This approach ensures that we neither infringe on copyright nor redistribute protected materials. All assets are credited to their original rights holders, and our use of these links falls under fair-use provisions for non-commercial, academic research. Accordingly, all textual assets created for this benchmark, including our annotations, are distributed under the Creative Commons Attribution Non Commercial No Derivatives 4.0 International (CC-BY-NC-ND-4.0 `https://spdx.org/licenses/CC-BY-NC-ND-4.0`) license.

I Taxonomy Definition

In this section, we show each category definition of each dimension in our taxonomy. In detail, we show definition of categories in each dimension in Table 7, and there is an illustration for categories in shot scale and angle in Figure 8.

Table 7: Definition of categories in seven dimensions.

Scale Category	**Definition**
Extreme Close-Up	An extreme close-up (ECU) is a shot that captures a subject in an extremely tight frame, focusing on a specific detail of the subject, such as an eye, a mouth, a ring, or a handwritten letter. This shot excludes most of the surrounding context, drawing the viewer's attention exclusively to the minute details of the subject.
Close-Up	A close-up (CU) is a shot that frames the subject's face, head, or a significant object, filling the screen with detailed visual information. For human subjects, a Close-Up typically shows the head and shoulders, allowing the audience to focus on facial expressions and emotions.
Medium Close-Up	A medium close-up (MCU) is a shot that frames a subject from the chest up, providing a balance between the subject's facial details and body language. This shot maintains the emotional focus of the Close-Up while also including some contextual information.
Medium Shot	A medium shot (MS) frames the subject from the waist up, providing a clear view of both facial expressions and body language. It is a versatile shot that strikes a balance between subject focus and contextual surroundings.
Medium Long Shot	A medium long shot (MLS), also known as a "three-quarters shot", frames the subject from the knees up, providing a broader view of the subject within the setting. It is often used to maintain a sense of the subject's body language while still focusing on the individual.
Long Shot	A long shot (LS) is a wide framing that captures the entire subject from head to toe, along with a significant portion of the surrounding environment. The subject is visible but occupies a relatively smaller portion of the frame.
Extreme Long Shot	An extreme long shot (ELS), also known as a wide shot (WS) or establishing shot, captures a vast expanse of the setting, with the subject appearing very small or even insignificant within the environment. This shot may cover vast landscapes, cityscapes, or wide action scenes.
Angle Category	**Definition**
High Angle	A high angle shot is captured with the camera positioned above the subject, angled downward. This perspective often makes the subject appear smaller, weaker, or vulnerable, depending on the narrative context.
Low Angle Shot	A low angle shot is captured with the camera positioned below the subject, angled upward. This perspective makes the subject appear larger, more dominant, or intimidating.
Bird's Eye View	A bird's eye view (or overhead shot) is an extremely high angle shot taken directly above the subject, providing a top-down perspective. This view emphasizes spatial layout and geometric patterns within the scene.
Worm's Eye View	A worm's eye view is an extreme low-angle shot taken from below the subject, almost directly upwards. This perspective can make subjects appear overwhelmingly large or powerful, or it can capture towering structures from ground level.
Diagonal Angle	A diagonal angle, is a camera angle that captures the subject from a non-frontal or backside, non-profile perspective. The camera is positioned at an intermediate angle between the subject's side and front or back, typically ranging from approximately 30° to 60° off-axis. This versatile angle allows the viewer to perceive multiple dimensions of the subject simultaneously, offering a more dynamic and three-dimensional representation.
Profile Shot	A profile shot is captured with the camera positioned to the side of the subject, showing the subject's profile or side view. This framing emphasizes the subject's silhouette, facial contours, and gestures.

Continued on next page

Table 7 – continued from previous page

Category	Definition
Back Shot	A back shot is a camera angle taken from behind the subject, typically showing the subject's back or shoulders while they face away from the camera. This can also include over-the-shoulder shots.
Composition Category	**Definition**
Symmetrical	Symmetrical composition is a technique where elements within the frame are arranged in a balanced and mirror-like manner, creating a sense of harmony and equilibrium. This can be achieved through vertical, horizontal, or radial symmetry.
Central	Central composition is a technique where the main subject is positioned at the exact center of the frame, drawing immediate attention to it. This approach uses the inherent strength of central focus, often resulting in a powerful and direct visual impact.
Diagonal	Diagonal composition is a technique that uses diagonal lines or elements within the frame to guide the viewer's eye and create a sense of movement, depth, and dynamism. These diagonal lines can be naturally present in the scene (such as a leaning tree) or can be intentionally created by tilting the camera (known as a dutch angle). This approach allows for a dramatic and visually engaging effect.
Rule of Thirds	The rule of thirds is a guideline that divides the frame into nine equal sections with two horizontal and two vertical lines. The main subjects are placed along these lines or at their intersections, creating a balanced and naturally pleasing composition.
Framing	Framing is a technique where elements within the scene are used to naturally frame the subject, directing the viewer's focus towards it. These framing elements can include natural objects (such as trees), architectural elements (such as windows), or other elements within the environment.
Curved Line	Curved line composition uses naturally occurring or deliberately arranged curved lines within the frame to guide the viewer's eye, create a sense of flow, or emphasize the softness of the scene. These lines can be literal (such as a winding road) or implied (such as a subject's pose).
Horizontal	Horizontal Composition is a technique where the main visual elements are arranged along a horizontal axis, emphasizing width and creating a sense of stability. This can be achieved using the horizon line, landscapes, or other horizontally aligned subjects.
Colors Category	**Definition**
Red	Red is a warm, highly intense color often associated with strong emotions, including passion, love, anger, danger, and urgency. In cinematography, it is used to draw attention, create tension, or symbolize strong emotional states.
Yellow	Yellow is a bright, warm color that is often associated with happiness, optimism, energy, and warmth. However, it can also represent caution, anxiety, or deceit, depending on the context.
Blue	Blue is a cool, calming color commonly associated with tranquility, stability, melancholy, and introspection. It is widely used to convey a sense of calmness, sadness, or detachment.
Green	Green is a color often associated with nature, growth, freshness, and harmony. However, in certain contexts, it can also represent envy, corruption, or toxicity.
Purple	Purple is a color traditionally associated with royalty, luxury, mystery, and spirituality. It is a color that can evoke both sophistication and fantasy, depending on the context.
Black and White	Black and white is a monochrome color scheme that removes all hues, focusing on contrasts between light and dark. This style emphasizes texture, composition, lighting, and shadow, often creating a timeless, dramatic, or nostalgic aesthetic.
Lighting Category	**Definition**
High Key	High key lighting is a technique characterized by bright, even illumination with minimal shadows and a high level of ambient light. This style is achieved using multiple light sources or a large, soft light source to reduce contrast.
Low Key	Low key lighting is a dramatic lighting technique that emphasizes strong contrast between light and dark areas, with deep shadows and minimal fill light. It is achieved using a primary light source with little to no fill light.
Hard Light	Hard light is a type of lighting that produces sharp, well-defined shadows and high contrast between illuminated and dark areas. It is created using a small, direct light source such as a spotlight or bare bulb.
Soft Light	Soft light is a technique that produces diffused, gentle illumination with gradual transitions between light and shadow. This effect is achieved using large light sources, diffusion panels, softboxes, or indirect lighting.
Back Light	Back light is a technique where the light source is positioned behind the subject, often creating a rim or halo effect around the subject's outline. This light separates the subject from the background and adds depth to the scene.
Side Light	Side light is a technique where the light source is placed at a 90-degree angle to the subject, illuminating one side while leaving the other side in shadow. This creates a strong contrast between light and darkness.

Continued on next page

Table 7 – continued from previous page

Category	Definition
Top Light	Top light is a technique where the light source is placed directly above the subject, casting shadows downward. This creates dramatic shadows on the subject's face and emphasizes the upper contours.
Focal Length Category	**Definition**
Standard Lens	A standard lens, also known as a Normal Lens, is a lens with a focal length that closely matches the human eye's natural field of view. In most cases, this ranges between 35mm to 50mm for full-frame cameras. Standard lenses provide a balanced perspective without significant distortion, making them highly versatile for various types of scenes.
Medium Focal Length	Medium focal length refers to lenses with a focal length slightly longer than standard lenses, typically between 50mm and 85mm for full-frame cameras. These lenses offer moderate compression and a slightly narrowed field of view, making subjects appear closer without the extreme effects of telephoto lenses.
Telephoto Lens	A telephoto lens is a long-focus lens with a focal length greater than 85mm, typically ranging from 85mm to 300mm or beyond for full-frame cameras. These lenses provide a narrow field of view and significant background compression, making distant subjects appear closer.
Fisheye Lens	A fisheye lens is an ultra-wide-angle lens with a focal length typically between 8mm and 16mm, designed to capture an extremely wide field of view, often with a 180° angle. It creates a distinctive curved, distorted image, which can be either circular (full-frame fisheye) or rectangular (rectilinear fisheye).
Macro Lens	A macro lens is a specialized lens designed for extreme close-up photography, capable of achieving a high level of magnification (typically 1:1 or greater). These lenses have a short minimum focusing distance, allowing detailed capture of small subjects.
Movement Category	**Definition**
Fixed Shot	A fixed shot is a static camera setup where the camera remains completely stationary throughout the shot. There is no movement in any direction (pan, tilt, or zoom). The composition and perspective are determined solely by the subject's movement within the frame.
Dolly In Shot	A dolly in shot is achieved by moving the camera towards the subject on a dolly track, creating a sense of gradual approach, increasing subject emphasis, or building tension.
Dolly Out Shot	A dolly out shot is achieved by moving the camera away from the subject on a dolly track, expanding the field of view, creating a sense of distancing, revelation, or release.
Crane Shot	A crane shot is a type of camera movement where the camera is mounted on a crane, allowing it to move vertically, horizontally, or in complex patterns across a scene. This technique provides sweeping, cinematic perspectives.
Trucking Left Shot	A trucking left shot is a lateral camera movement to the left, maintaining a consistent perspective of the subject. This is often used to follow a subject moving horizontally.
Trucking Right Shot	A trucking right shot is a lateral camera movement to the right, maintaining a consistent perspective of the subject. This is also used for tracking horizontal movement.
Pan Left Shot	A pan left shot is achieved by rotating the camera horizontally to the left from a fixed position, allowing a gradual reveal of the scene from right to left.
Pan Right Shot	A pan right shot is achieved by rotating the camera horizontally to the right from a fixed position, allowing a gradual reveal of the scene from left to right.
Tilt Up Shot	A tilt up shot is a vertical camera movement where the camera tilts upward from a fixed position, gradually revealing the upper part of the scene or subject.
Tilt Down Shot	A tilt down shot is a vertical camera movement where the camera tilts downward from a fixed position, gradually revealing the lower part of the scene or subject.
Rolling Clockwise Shot	A rolling clockwise shot is a dynamic camera movement where the camera rotates around its lens axis in a clockwise direction, creating a spiraling effect.
Rolling Counterclockwise Shot	A rolling counterclockwise shot is a dynamic camera movement where the camera rotates around its lens axis in a counterclockwise direction, creating an opposite spiraling effect.
Tracking Shot	A tracking shot is a camera movement that follows a subject along a path, maintaining consistent framing. It can be achieved using a handheld setup.
Zoom In Shot	A zoom in shot is an optical camera technique where the focal length of the lens is adjusted to bring the subject closer without moving the camera physically. This effect magnifies the subject within the frame.
Zoom Out Shot	A zoom out shot is an optical camera technique where the focal length of the lens is adjusted to increase the field of view, making the subject appear smaller within the frame.
Combinational Shot	A combinational shot, is a complex camera movement technique that combines two or more distinct camera movements within a single continuous take. This may include any combination of Dolly, Trucking, Pan, Tilt, Zoom, Crane, Rolling, or Tracking movements executed in sequence or simultaneously.