Natural Language Camera Movement Understanding

Yuwen Tan[1], Joey Huang[1], Jin Huang[2], Haoxiang Li[2], and Boqing Gong[1]

[1] Boston University, Boston, MA 02215, USA
[2] Pixocial Technology, Bellevue, WA 98004, USA

https://1yuwen.github.io/ACaM-Project-Page/

Abstract. Understanding camera movement in natural language is critical for training and evaluating video generation models, among other applications. However, we demonstrate that existing vision-language models (VLMs) fail this task in surprising ways, frequently confusing translation with rotation, left with right, and object movement with camera movement. To address these limitations, we establish natural language camera movement understanding as a standalone research task. We introduce a two-level cinematographic taxonomy and an extensive, atomic benchmark featuring both real and synthetic videos. Furthermore, we curate a large-scale, multi-source training set enhanced by targeted camera movement augmentation. Our fine-tuned VLM-8B outperforms Gemini 3.1 Pro by 10% and 11% on our benchmark's real and synthetic videos, respectively. Despite these gains, a significant gap remains relative to human performance, underscoring the need to promote and facilitate future research on natural language camera movement understanding.

Keywords: Camera movement understanding · Vision language models · Video understanding

1 Introduction

Camera movement is one of the most powerful tools in filmmaking [4,6,15,25,32], influencing how audiences perceive and engage with the narrative. Filmmakers deliberately use specific camera movements to serve distinct storytelling purposes. For instance, a 'dolly in' shot can emphasize a character's internal realization, while a pan shot can reveal new spatial details (see more examples of camera movement in Fig. 1). In recent years, modern video generation models [5, 9, 16, 17, 23, 37, 48] have begun to explicitly incorporate camera movement into the generation process; users are now describing camera movement using cinematographic text prompts (e.g., 'A slow pan') to guide video generation.

Natural language understanding of camera movement is more important than ever. As modern text-to-video generation models are increasingly being used to generate dynamic scenes with varied camera movements, enhancing the controllability of these specific movements heavily depends on massive, high-quality

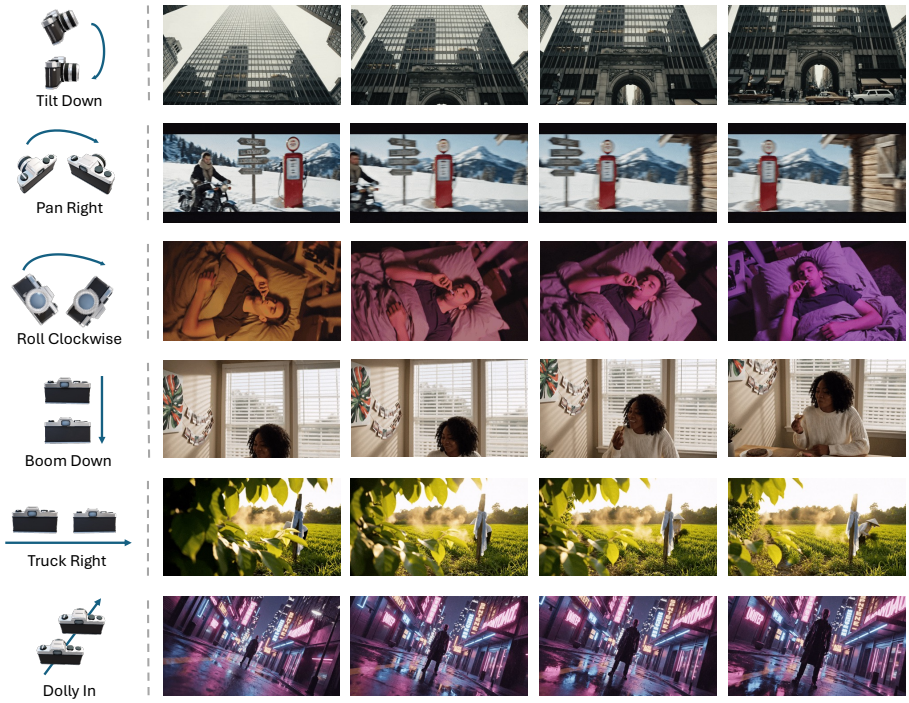

Fig. 1: Some examples of camera movements in cinematographic practice: Camera rotations (pan, tilt, roll) and translations (boom, dolly, truck).

video-text pairs. Therefore, robust models capable of automatically filtering, curating, and captioning cinematographic camera movements in large-scale, unannotated videos are becoming increasingly relied upon to eliminate the bottleneck of expensive manual annotation. Simultaneously, in terms of model evaluation [14, 24, 33, 40], the ability to automatically and reliably verify whether generated videos faithfully execute specific camera prompts has become indispensable for benchmarking modern generative systems.

Unlike traditional camera pose estimation tasks [12,20,27,39,41], which estimate per-frame poses in coordinate space, natural language understanding operates at the semantic level. It maps spatio-temporal visual cues directly to human-interpretable cinematographic languages. While geometry-based methods provide accurate trajectories, they are not as accessible to common users as language prompts and often fail to infer object-centric camera movements [21, 47], such as 'tracking' and 'arc'. Furthermore, their reliance on depth estimation and calibrated intrinsics makes them computationally heavy and unnecessarily complex for natural language camera movement understanding. We believe vision-language models (VLMs) [2, 3, 18, 26, 35, 42, 44, 50] offer a more suitable and scalable framework for this task.

Table 1: Taxonomy of camera movements with cinematographic terminology.

Translation	**Dolly In** (Move in, Push in, Move forward)	**Dolly Out** (Move out, Pull out, Move backward)
	Truck Left (Move to the left)	**Truck Right** (Move to the right)
	Boom Up (Pedestal up, Crane up, Move upwards)	**Boom Down** (Pedestal down, Crane down, Move downwards)
Rotation	**Pan Left**	**Pan Right**
	Tilt Up	**Tilt Down**
	Roll Counterclockwise	**Roll Clockwise**
Focal Length Change	**Zoom In**	**Zoom Out**
Static	**Static** (Stationary, Fixed, No movement)	
Object-centric Movement	**Tracking** (Follow shot)	**Arc** (Orbit)

Drawing upon established cinematography literature [15, 25], we categorize various camera movements into a two-level taxonomy of 17 distinct classes (summarized in Tab. 1). This taxonomy comprehensively covers camera translations, rotations, focal-length changes, object-centric movements, and static shots. Crucially, it provides a channel to systematically investigate the capability of camera movement understanding in VLMs. Although recent works [7, 19, 21, 22, 43, 46] have observed that VLMs struggle significantly with this task, the underlying reasons for their poor performance remain largely unexplored. In this work, we conduct an in-depth analysis to investigate *why* camera movement understanding is so challenging for VLMs. Through preliminary evaluations, we find that their inherent deficiencies fall under five failure modes: VLMs are insensitive to subtle inter-frame changes, confuse physical movement with angular change, and frequently misinterpret movement direction. Furthermore, they struggle to distinguish between optical zoom and physical dolly, and also conflate object movement with global camera movement.

These challenges are nontrivial. We tackle them by first promoting natural language camera movement understanding as a standalone research problem rather than allowing it to become a minor part of fine-grained motion [7, 10, 36], cinematography [13, 19, 22, 28, 29, 34, 43, 46], and camera motion primitives [21]. We then instantiate it with a comprehensive and atomic benchmark containing real-world cinematography and synthetically generated videos, in response to the needs of curating real-world video-text training data and automatic rating of modern video generation models. We purposely choose atomic camera movements in the benchmark—one target camera movement per video clip for evaluation—leaving the combination of multiple movements for future work. Finally, we strive to deliver the best-performing VLM model for camera movement understanding by constructing a large-scale training dataset annotated with camera movement labels from five data sources and applying targeted data augmentation to explicitly enhance motion perception. We fine-tune Qwen3-VL-4B and 8B [2] on the curated dataset, achieving relative improvements of 10%

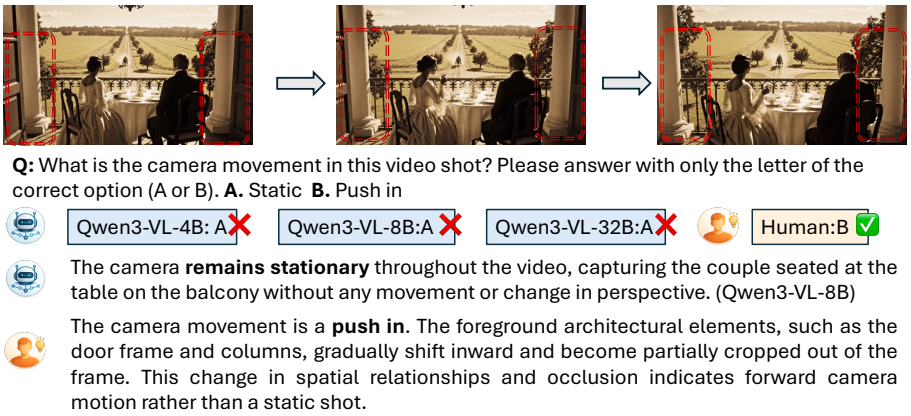

Fig. 2: An example illustrating VLMs' limited sensitivity to small camera movement.

on real-world videos and 11% on synthetic videos over the strong proprietary model Gemini-3.1-Pro. However, a human study still exposes a substantial gap in performance between our best VLM and human, thus calling for future and more extensive research on natural language camera movement understanding.

2 Why is camera movement understanding challenging for current VLMs?

To provide an initial understanding of why this task is so challenging for VLMs, we conduct preliminary evaluations and present several illustrative examples. We find that their deficiencies fall under five failure modes. We elaborate on each of these failure modes in the following subsections. Detailed experimental setups are provided in the supplementary material.

2.1 Difficulty in Perceiving Small Camera Movement

We observe that VLMs frequently misclassify dynamic camera movement as 'static', particularly when the movement intensity is moderate, suggesting limited sensitivity to inter-frame differences. In contrast, such differences are readily perceptible to humans. As shown in Fig. 2, humans can easily infer that the depicted camera movement is 'push in' by perceiving and localizing changes in the foreground, whereas Qwen3-VL models [2] consistently predict these videos as 'static'. To further quantify this issue, we focus on the 'push in' category and manually divide the corresponding videos into three levels of camera movement: low, medium, and high. We then construct a binary-choice question-answering task to evaluate whether models can detect the camera movement. As shown in Tab. 2, Qwen3-VL-32B and Gemini-3-Pro perform near or below random chance under low intensity motions, revealing a substantial gap between VLMs and human perception.

Table 2: VLM accuracy on 'push in' *vs.* three levels of camera movement intensity and the number of frames fed to VLMs.

	Frames=8/Fps=1			Frames=16/Fps=2			Frames=32/Fps=4		
Model	Low	Medium	High	Low	Medium	High	Low	Medium	High
Random	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
Qwen3-VL-32B [2]	41.67	81.90	100.00	**52.08**	86.67	100.00	45.83	**88.57**	**100.00**
Gemini-3-Pro [35]	29.17	47.62	85.37	**31.25**	56.19	**100.00**	29.17	**58.10**	92.68

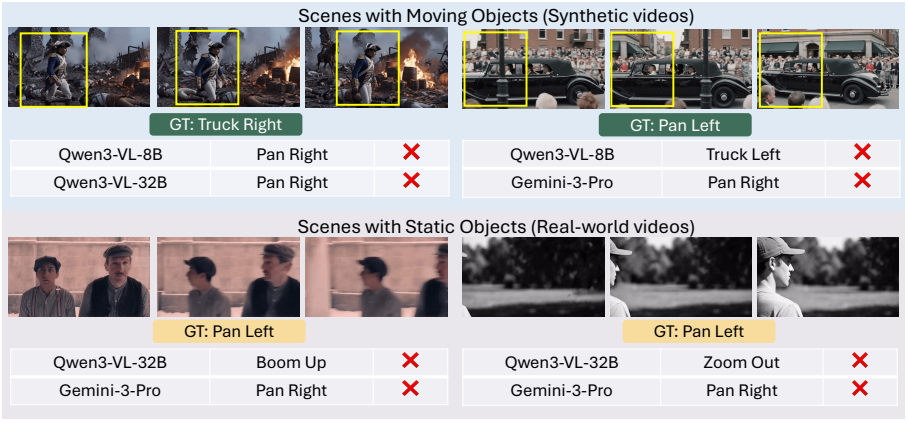

Fig. 3: Examples of translation-rotation and left-right confusion.

2.2 Confusion Between Translation and Rotation

Beyond fine-grained sensitivity, VLMs also frequently struggle to distinguish between camera translation and rotation. As illustrated in Fig. 3, when presented with a tracking shot of a walking man (top left), which is a purely lateral camera translation ('truck right'), both Qwen3-VL-8B and 32B models misclassify the movement as a horizontal rotation ('pan right'). Conversely, in the top right, where the camera rotates horizontally to follow a moving vintage car ('pan left'), Qwen3-VL-8B exhibits the exact opposite error, mispredicting the rotation as a translation ('truck left'). These reciprocal failures strongly indicate that current VLMs predominantly rely on superficial 2D shifts rather than developing a genuine understanding of 3D camera mechanics and spatial geometry.

2.3 Left-Right Confusion

VLMs also often mistake camera's moving left as moving right, and vice versa. A representative example is in the bottom row of Fig. 3. Although the camera pans left to reveal more details on the left side, Gemini-3-Pro predicts the opposite direction (pan right). We further validate this observation using the confusion matrices in Fig. 4. Off-diagonal mass appears between opposite directions within the same movement, particularly for Gemini-3-Pro.

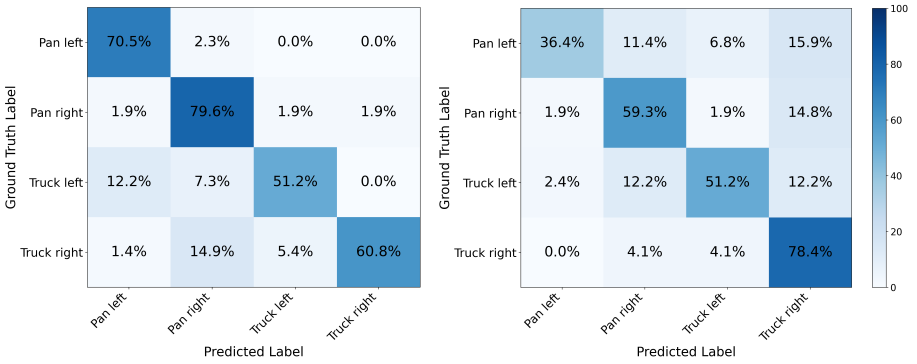

Fig. 4: Partial confusion matrices for Qwen3-VL-32B (left) and Gemini-3-Pro (right), illustrating translation-rotation and left-right confusion. (Full matrix in supplement)

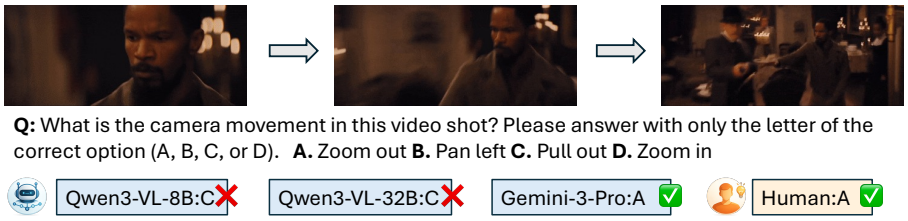

Fig. 5: A failure case in distinguishing optical 'zoom out' from physical 'dolly out'.

2.4 Confusion Between Optical Zoom and Physical Dolly

A prominent blind spot for current VLMs is in their limited ability to distinguish optical scaling from physical movement. Although both produce a centered expansion or contraction, their underlying mechanics differ. Camera translation introduces depth-dependent parallax that shifts the relative positions of foreground and background elements, whereas focal-length changes uniformly magnify the scene while preserving perspective. As illustrated in Fig. 5, models fail to exploit these cues, often misclassifying optical zoom as physical translation. This suggests that VLMs rely on superficial 2D scaling cues that are insufficient in differentiating camera intrinsics (focal changes) from extrinsic movement.

2.5 Confusing Object Movement with Camera Movement

We identify another critical failure mode: VLMs rely on salient foreground dynamics rather than global background cues to infer camera movement. When objects move independently of the camera, they induce local 2D size or position changes, which VLMs often misinterpret as global camera movement, illustrated in Fig. 6. Although the camera translates backward, the racing cars move forward faster than the camera, causing them to occupy more of the frame over

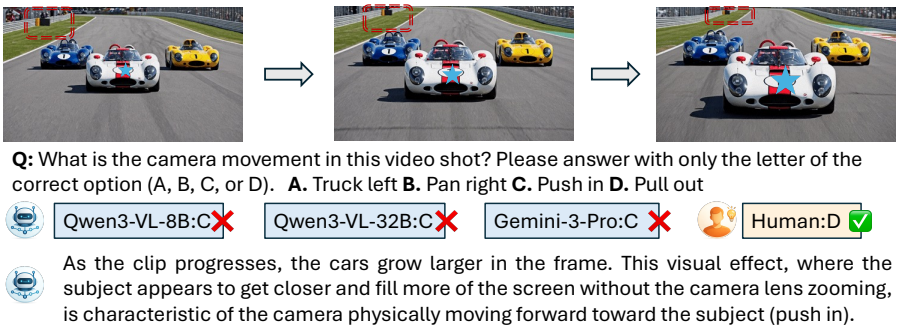

Fig. 6: Object motion challenges current VLMs' inference about camera movement.

Table 3: Comparison with camera motion understanding benchmarks. For benchmarks not designed for this task, statistics are reported only for clips involving camera motion.

Benchmark	#Domain		#Videos		#QA Format		#Human	#Labels
	Real	Syn	Train	Test	Yes/No	MCQ	Performance	Number
Cinematic2K [19]	✓	✗	✗	1,047	✗	✗	✗	11
CameraBench [21]	✓	✗	1,402	1,071	✓	✗	✓	50
ShotBench-Subset [22]	✓	✗	1,058	464	✗	✓	✗	17
CineTechBench-Subset [43]	✓	✗	✗	120	✗	✓	✗	15
MotionBench-Subset [10]	✓	✗	5,000	385	✗	✓	✗	-
FavorBench-Subset [36]	✓	✗	17,000	546	✗	✓	✗	-
ACaM (Ours)	✓	✓	24,321	2,602	✓	✓	✓	17

time. VLMs incorrectly interpret this 'subject growing larger' cue as 'push in', indicating that they behave more like localized object trackers than global spatial observers when tasked with natural language camera movement understanding.

3 ACaM: A Benchmark for Natural Language Understanding of *A*tomic *Ca*mera *M*ovement

Findings in Sec. 2 indicate that camera movement understanding is a nontrivial task that deserves careful and systematic study. Thus, in order to promote and facilitate future research on it, we compile an extensive benchmark (dubbed ACaM for Atomic Camera Movement) following a taxonomy derived from cinematography, comprising both real-world and synthetic videos.

3.1 Camera Movement Taxonomy

We categorize camera movement into 17 distinct classes (see Tab. 1) following the cinematography literature [15, 25]. CameraBench [21] introduces a comprehensive taxonomy of over 50 motion primitives, while our evaluation focuses on 17 atomic, intentional camera movements that can be unambiguously described in

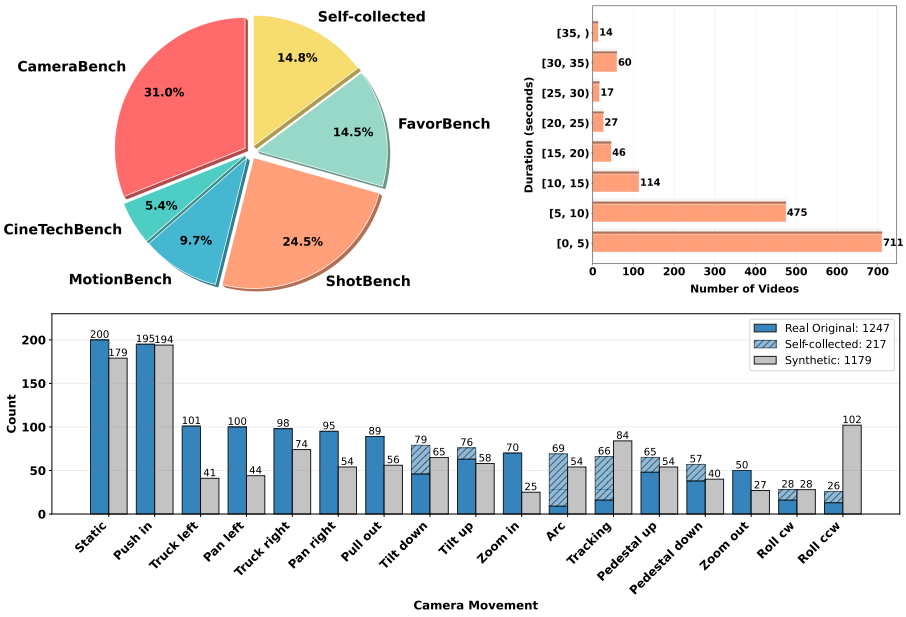

Fig. 7: Dataset statistics of our ACaM evaluation benchmark. Top: Source composition and duration distribution of real-world videos. Bottom: Class-wise camera movement distribution of real-world and synthetic videos.

natural language. Furthermore, our benchmark provides a more comprehensive set of categories than Cinematic2K [19] and CineTechBench [43], spanning five core dimensions: physical translations, angular rotations, focal length changes, static shots, and object-centric movements. Tab. 3 compares our benchmark with existing camera movement and motion understanding benchmarks. Notably, our ACaM is the first to include synthetic videos and human performance as a reference baseline for direct human–model comparison.

3.2 Real-World Videos

Data construction pipeline. To construct a diverse and high-quality set of real-world videos, we unify several existing motion understanding benchmarks [10, 21, 22, 36, 43]. Given the heterogeneous annotation formats and motion categories across these datasets, we apply tailored filtering and refinement strategies to each source. We first use the binary question-answering subset from CameraBench [21], which is designed for camera motion understanding. To focus on atomic movement recognition, we retain only videos with a single affirmative label. We then use Gemini-3-Pro [35] to parse movement descriptions from captions and remove detected secondary movements from the negative answer options, ensuring the QA task remains centered on the primary camera movement.

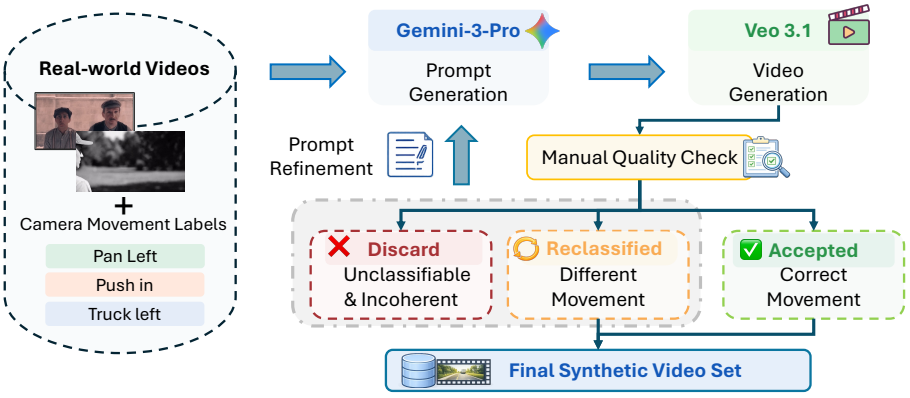

Fig. 8: Pipeline for generating synthetic videos of various camera movement classes.

For the cinematography benchmarks of ShotBench [22] and CineTechBench [43], we extract subsets related to camera movement and retain only evaluation questions involving a single camera movement. Similarly, for fine-grained motion understanding benchmarks (MotionBench [10] and FavorBench [36]), we extract the camera-related QA subsets and remove questions involving object motion. These datasets present an interesting challenge, among others, since they often require identifying camera movement for a specific event or timestamp (see examples in the supplementary material). Finally, to enrich underrepresented camera movement classes (see Fig. 7, bottom), we search for and curate YouTube videos. The assembled dataset is manually verified by graduate students with cinematography training to remove incorrectly labeled or ambiguous samples across all sources (see supplementary materials for details).

Dataset Statistics. Fig. 7 summarizes some statistics of our real-world video set. After filtering and refinement, the final benchmark contains 1,423 videos and 1,464 QA pairs, covering 17 camera movement categories aggregated from six different sources. More than 80% of the videos are shorter than 10 seconds, ensuring that the evaluation primarily focuses on the dominant, atomic camera movement behind each video. Regarding the class distribution, 'static' and 'push in' account for a slightly larger portion than the others, while the 'rolling' classes are underrepresented even after we enrich them with YouTube videos.

3.3 Synthetic Videos

In addition to real-world videos, ACaM also incorporates AI-generated videos, mimicking the automatic rating of generative models using VLMs. Fig. 8 shows the pipeline for generating synthetic videos using Veo 3.1 [9]—our lab currently has research credits for no platform but Google Cloud.

To construct the synthetic video set, we start from our realistic videos for each camera movement class. For each clip, we feed the raw video and its camera

Table 4: Sources and statistics of our instruction-tuning set (MCQ: Multi-Choice QA)

Dataset	#Extracted Clips	#Annotation	#Label Source	#Label Num.
CameraBench [21]	723	Motion Caption	Human Annotation	16
ShotBench [22]	375	MCQ Format	Human Annotation	14
SpatialVID [38]	29,733	Movement Labels	Estimated Poses	12
MultiCamVideo [1]	12,048	Camera Pose	Synthetic Trajectory	13
GenDoP [49]	2,438	Movement Caption	Estimated Poses	6
Overall	45,317	MCQ Format	Mixed	17

movement label to Gemini-3-Pro [35] while instructing it to analyze the video and produce a structured video generation prompt. The resulting prompts are then fed to Veo 3.1, which performs video generation to synthesize clips aligned with the camera movement in the prompt. After generation, we conduct a manual quality-control pass as follows. Each clip is assigned to one of three outcomes: *accepted* if it exhibits the intended camera movement, *reclassified* if it depicts a different camera movement label in our taxonomy, or *discarded* if the movement is unclassifiable or temporally incoherent (e.g., abrupt transitions). For discarded or reclassified cases, we let Gemini-3-Pro rewrite the video generation prompt to reduce ambiguity and enhance clarity of the intended camera movement. We then send the refined prompts to Veo 3.1 for a second round of video generation.

In general, 'zoom', 'arc', and 'roll (clockwise)' are more difficult to generate than the other camera movement classes. For these categories, we employ a separate prompt generation script that conditions solely on the designated camera movement label, without any anchor on real-world videos. The supplementary material details the whole pipeline with examples.

Data Statistics. The synthetic videos are four seconds long. Their class-wise distribution is shown on the bottom of Fig. 7. Despite the iterative pipeline, certain 'zoom' and 'roll' classes of videos remain difficult to generate. Interestingly, Veo 3.1 has a systematic bias toward 'roll counterclockwise'; many instructions for 'roll clockwise' actually lead to videos of 'roll counterclockwise'. 'Zoom'-related instructions often result in 'static' shots or 'dolly' movement.

4 Towards VLMs that understand camera movement in natural language

To enhance VLM capabilities in camera movement understanding, we construct an instruction-tuning set consisting of 27K samples by drawing from 45K raw video clips, enhanced with targeted camera movement augmentation.

4.1 Data Processing and Filtering

While CameraBench [21] and ShotBench [22] provide videos with camera movement labels, their training sets are small (approximately 1K video clips; see Tab. 3). We pool them and further incorporate video clips from three datasets

Fig. 9: Targeted data augmentation for distribution balancing. Top: The shift from an unbalanced 45K raw dataset to a refined 27K instruction-tuning set. Bottom: The augmentation operators used to synthesize samples for tail classes.

that contain synthetic or estimated camera trajectories. Specifically, we use a subset of SpatialVID [38], which provides estimated camera poses, despite being highly dominated by the 'push in' movement. MultiCamVideo [1] is a dataset of synthetic videos rendered in Unreal Engine 5 with controlled camera trajectories, and GenDoP [49], which provides natural language descriptions of camera movements. Tab. 4 summarizes the data sources and labels.

Our training corpus prioritizes atomic camera movement, which we believe is the first step towards complete natural language understanding for various, potentially compound, and long-form camera movements. From ShotQA [22], we retain only the subset of single movements. For CameraBench [21] and GenDoP [49], we employ GPT-5-nano [31] to parse camera movement descriptions from their video captions and convert them to multi-choice question-answering (MCQ) tasks. For MultiCamVideo [1], frame-wise camera poses are projected into semantic movement labels using geometric heuristics, retaining only the segments containing a single camera movement. In SpatialVID [38], due to noisy VLM-generated global captions, we instead adopt their frame-wise camera movement descriptions as the groundtruth. Since pose annotations in SpatialVID and GenDoP are derived from geometry-based estimation and prone to artifacts, we further perform manual verification to correct any labeling errors. After filter-

Table 5: Results of various models on our ACaM's real-world videos

Model	Static	Rot.	Trans.	Zoom	Arc	Track	Overall	Avg
Random Guess	24.50	24.28	24.00	24.54	24.78	25.00	24.27	24.37
Human Performance	99.00	93.07	90.39	90.83	97.10	96.97	93.44	93.14
Visual Geometry Models								
Mega-SaM [20]	89.55	**74.94**	54.16	–	–	–	–	–
ViPE [12]	61.81	49.71	**73.17**	–	–	–	–	–
Spatial VLMs								
G^2VLM-2B [11]	9.50	4.52	4.52	7.44	5.71	7.58	5.66	5.24
Spatial-MLLM [45]	59.00	16.98	31.72	41.67	24.29	77.27	34.22	29.68
VLM-3R-7B [8]	**94.50**	20.69	28.55	35.83	51.43	89.39	40.22	35.02
General VLMs								
Qwen3-VL-4B [2]	*93.50*	53.47	34.21	61.67	56.52	80.30	53.01	46.96
Qwen2.5-VL-7B [3]	81.00	33.91	34.71	58.33	66.67	92.42	46.86	45.62
Qwen3-VL-8B [2]	85.00	65.84	39.50	61.67	79.71	74.24	58.27	55.25
LLaVa-OV-7B [18]	92.50	28.71	25.79	24.17	62.32	75.76	39.55	34.85
InternVideo2.5-8B [44]	82.50	31.44	28.43	64.17	50.72	92.42	43.51	41.07
InternVL3.5-8B [42]	89.00	48.51	30.58	58.33	40.58	86.36	48.77	46.15
InternVL3.5-14B [42]	81.00	43.56	39.83	63.33	55.07	89.39	51.37	47.21
Qwen3-VL-32B [2]	92.00	64.11	47.93	64.17	57.97	86.36	61.95	58.66
GPT-5 [31]	61.50	66.58	54.55	61.67	68.12	80.30	61.20	61.54
Gemini-3-Pro [35]	83.50	64.36	60.84	65.00	82.61	96.97	68.24	66.83
Gemini-3.1-Pro [35]	83.92	63.34	64.00	62.39	76.81	95.38	68.44	67.73
Camera Movement Specialized VLMs								
CameraModel-7B [21]	55.50	68.81	47.93	45.83	**91.30**	**98.48**	58.88	60.56
ShotVL-7B [22]	91.00	60.40	51.07	32.50	68.12	98.48	60.52	57.09
CamReasoner-7B [47]	76.50	41.58	32.89	59.17	39.13	80.30	45.83	44.99
Our SFT Qwen3-VL-4B	76.00	*71.53*	*64.13*	*76.67*	78.26	93.94	*70.83*	*72.27*
Our SFT Qwen3-VL-8B	85.50	70.05	64.13	**84.17**	*84.06*	*96.97*	**72.75**	**74.28**

ing, the final training corpus contains 45,317 video clips in the MCQ format, as summarized in Tab. 4. More details are provided in the supplementary material.

4.2 Data Sampling and Augmentation

The top-left panel of Fig. 9 shows that the integrated instruction-tuning set exhibits a significant long-tail distribution, which could introduce strong bias to VLMs if we fine-tune models directly on this set. Hence, we re-sample the videos and design targeted data augmentation, striving to balance the videos across camera movement categories. Specifically, we address the scarcity of 'zoom' by progressively cropping static frames and resizing them to the original resolution, yielding approximately 1K augmented videos. We employ temporal reversal to transform a subset of 'dolly in' sequences into their 'dolly out' counterparts, yielding 2K augmented videos. Similarly, horizontal flipping is applied to 'truck' movement, producing about 1.7K additional videos. Finally, to enrich the extremely scarce 'roll' classes, we synthesize rolling dynamics by applying continuous affine rotations to static video shots. By combining the re-sampling and video augmentation strategies, we arrive at a refined, more balanced training corpus of 27K high-quality video-MCQ pairs.

Table 6: Results of various models on our ACaM's synthetic videos

Model	Static	Rot.	Trans.	Zoom	Arc	Track	Overall	Avg
Random Guess	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00
Human Performance	98.88	97.72	98.04	96.15	100.00	97.62	98.05	97.61
Visual Geometry Models								
Mega-SaM [20]	95.53	**76.35**	61.66	–	–	–	–	–
ViPE [12]	68.16	52.42	**78.65**	–	–	–	–	–
Spatial VLMs								
G^2VLM-2B [11]	4.44	2.91	6.99	1.92	5.45	13.10	5.54	4.51
Spatial-MLLM [45]	68.33	29.65	27.07	50.00	43.64	94.05	40.75	36.04
VLM-3R-7B [8]	**99.44**	18.90	43.45	51.92	47.27	89.29	48.68	38.94
General VLMs								
Qwen3-VL-4B [2]	*97.77*	54.42	42.92	73.08	40.74	72.62	58.02	53.68
Qwen2.5-VL-7B [3]	91.06	52.99	50.54	82.69	61.11	94.05	62.43	58.50
Qwen3-VL-8B [2]	90.50	62.39	45.10	*80.77*	55.56	83.33	61.92	60.91
LLaVa-OV-7B [18]	96.09	39.60	37.25	36.54	51.85	86.90	51.06	41.59
InternVideo2.5-8B [44]	94.97	44.73	31.15	75.00	27.78	*91.67*	50.98	46.71
InternVL3.5-8B [42]	96.09	47.29	33.99	67.31	35.19	73.81	51.74	47.73
InternVL3.5-14B [42]	88.27	41.60	47.93	76.92	31.48	86.90	55.47	51.15
Qwen3-VL-32B [2]	91.62	68.95	55.77	82.69	51.85	67.86	67.01	64.17
GPT-5 [31]	64.80	69.80	57.95	82.69	53.70	63.10	63.78	65.09
Gemini-3-Pro [35]	89.94	61.54	68.63	73.08	53.70	88.10	70.65	66.27
Gemini-3.1-Pro [35]	92.07	62.00	74.07	69.23	55.56	85.37	73.01	69.58
Camera Movement Specialized VLMs								
CameraModel-7B [21]	58.10	62.68	51.63	71.15	**88.89**	98.81	61.83	68.45
ShotVL-7B [22]	96.09	69.80	63.62	63.46	29.63	**98.81**	71.33	64.49
CamReasoner-7B [47]	84.92	51.28	42.70	67.31	55.56	77.38	55.81	53.74
Our SFT Qwen3-VL-4B	82.68	*74.36*	66.67	78.85	68.52	84.52	73.28	74.40
Our SFT Qwen3-VL-8B	94.41	72.36	*74.51*	**82.69**	*68.52*	91.67	**78.20**	**77.44**

5 Experiments

We apply supervised fine-tuning (SFT) to Qwen3-VL-4B and 8B over our training set. We also test geometry-based methods [12, 20], spatial VLMs [8, 11, 45], general VLMs [2, 3, 18, 26, 35, 42, 44], and prior camera movement specialized VLMs [21, 22, 47]. The supplementary material reports the detailed training and evaluation setups, and Tab. 5 and Tab. 6 show the quantitative results on our benchmark's real and synthetic videos, respectively. We report both the overall accuracy and the average accuracy across the 17 camera-movement classes (additional results for the binary QA setting are provided in the supplementary material). The main observations are as follows.

Geometry-based models [12, 20] demonstrate complementary strengths to VLMs; Mega-SaM excels in static scenes and camera rotations, whereas ViPE performs the best on translations. However, neither achieves robust performance across all classes, revealing the limitations of geometry-based methods to natural language understanding of camera movement. Indeed, these methods do not explicitly model semantics and often assume fixed intrinsics when estimating camera pose, limiting their applicability to predict 'zoom', 'tracking', and 'arc'.

SpatialVLMs [8,11,45] incorporate geometry-aware learning objectives when they fine-tune base VLMs. However, their performance on our ACaM is surprisingly low, most likely because they were specialized for the spatial understanding of the video content only, rather than its spatial relationship with camera movement. Among the general VLMs of no more than 8B parameters, Qwen3-VL [2] achieves the strongest performance. Furthermore, Qwen3-VL's 32B model significantly outperforms its 4B and 8B models and is nearly on par with Gemini-3-Pro on synthetic videos. Gemini-3/3.1-Pro exhibits a stronger capability to perceive translational movement than other VLMs.

We next compare our SFT Qwen3-VLs to the existing camera movement specialized VLMs. CameraModel-7B [21] is competitive, even on par with Gemini-3-Pro on synthetic videos, but it sacrificed its discrimination ability for static shots. In contrast, ShotVL-7B [22] performs better on translational movement and static shots than other models. Overall, the existing specialized models perform better on object-centric movements than other VLMs, but they all struggle with 'translation' and 'zoom'. Our SFT Qwen3-VL-4B significantly outperforms the larger 7B VLMs that were previously specialized for camera movement, with 9% to 19% gains, and our 8B model boosts the results further, even improving over Gemini-3.1-Pro by 10% and 11% on the real and synthetic videos, respectively. However, even the strongest models fall far behind humans, revealing a noticeable gap between human and VLM performance.

Ablation study. We further analyze the impact of targeted data sampling, augmentation, and LoRA rank in Tab. 7. Notably, the combination of targeted sampling and augmentation significantly improves performance, increasing accuracy from 58.56% to 75.03% on synthetic videos, which validates the effectiveness of the data balancing strategy. The LoRA rank has a marginal impact on the 4B model, whereas a higher rank yields better performance for the 8B model (results for the 8B model are provided in the supplementary material).

Table 7: Ablation of sampling, augmentation and different LoRA rank.

Samp.	Aug.	LoRA	Model	Avg.	
				Real	Syn
×	×	r=64	4B	48.42	58.56
✓	×	r=64	4B	63.78	72.07
✓	✓	r=64	4B	70.88	75.03
✓	✓	r=128	4B	71.88	75.20
✓	✓	r=256	4B	72.27	74.40
✓	✓	r=256	8B	**74.28**	**77.44**

6 Related Work

Motion Understanding Benchmarks. Motion understanding is a fundamental component of video understanding [7,10,21,36] and plays an important role in many downstream applications. MotionBench [10] and FavorBench [36] evaluate fine-grained motion perception, covering both camera and object motion. CameraBench [21] focuses specifically on camera motion, providing around 3K expert-annotated videos and introducing a cinematography-inspired taxonomy for systematic evaluation. CamReasoner [47] further formulates camera motion understanding as a reasoning task, leveraging chain-of-thought supervision and

reinforcement learning to improve performance on CameraBench. In addition, ShotBench [22] (RefineShot [46]), CineTechBench [43], and Cinematic2K [19] construct expert-annotated QA pairs from film content across cinematography-related [13, 28, 29] dimensions, including camera motion understanding.

Camera Pose Estimation through Visual Geometry Learning. Estimating camera parameters and scene geometry from videos is a long-standing problem in computer vision. Classical approaches include SLAM [12, 20] and Structure-from-Motion (SfM) [27, 30], which explicitly estimate camera trajectories and reconstruct 3D structure from multi-view consistency. More recent methods adopt end-to-end learning frameworks to directly predict 3D geometry or camera poses from video inputs, such as CUT3R [41], and VGGT [39]. Although these approaches recover geometric structure, they generally lack high-level semantic reasoning. Recent efforts attempt to bridge geometry and language by incorporating geometric representations into VLMs [8, 11, 45].

7 Conclusion

This work establishes natural language camera movement understanding as a standalone research task. We instantiate it by introducing ACaM, an extensive benchmark grounded in a two-level cinematographic taxonomy and covering both real-world and synthetic videos. To deliver the best-performing VLM model, we construct a large-scale training corpus with targeted sampling and augmentation. Our fine-tuned VLM-8B outperforms Gemini-3.1-Pro by 10% on real-world videos and 11% on synthetic videos. Despite these gains, the persistent gap to human performance underscores the need for further research.

Acknowledgments

This work was supported in part by NSF 2540851 and a Gemini Academic Program Award.

References

1. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14834–14844 (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Bordwell, D., Thompson, K., Smith, J.: Film art: An introduction, vol. 7. McGraw-Hill New York (2008)

5. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), https://openai.com/research/video-generation-models-as-world-simulators, last accessed 2026/06/29
6. Brown, B.: Cinematography: theory and practice: image making for cinematographers and directors. Routledge (2016)
7. Du, Y., Fan, T., Nan, K., Xie, R., Zhou, P., Li, X., Yang, J., Yang, Z., Tai, Y.: Motionsight: Boosting fine-grained motion understanding in multimodal llms. arXiv preprint arXiv:2506.01674 (2025)
8. Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Qu, H., Wang, D., Yan, Z., et al.: Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. arXiv preprint arXiv:2505.20279 (2025)
9. Google: Veo. https://deepmind.google/models/veo/ (2024), last accessed 2026/06/29
10. Hong, W., Cheng, Y., Yang, Z., Wang, W., Wang, L., Gu, X., Huang, S., Dong, Y., Tang, J.: Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8450–8460 (2025)
11. Hu, W., Lin, J., Long, Y., Ran, Y., Jiang, L., Wang, Y., Zhu, C., Xu, R., Wang, T., Pang, J.: G^2 vlm: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. arXiv preprint arXiv:2511.21688 (2025)
12. Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.H., et al.: Vipe: Video pose engine for 3d geometric perception. arXiv preprint arXiv:2508.10934 (2025)
13. Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: Movienet: A holistic dataset for movie understanding. In: European conference on computer vision. pp. 709–727. Springer (2020)
14. Jia, Z., Zhang, Z., Qian, J., Wu, H., Sun, W., Li, C., Liu, X., Lin, W., Zhai, G., Min, X.: Vqa2: visual question answering for video quality assessment. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 6751–6760 (2025)
15. Keating, P.: The dynamic frame: camera movement in classical hollywood. Columbia University Press (2019)
16. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
17. Kuaishou: Kling. https://kling.ai/ (2024), last accessed 2026/06/29
18. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: LLaVA-onevision: Easy visual task transfer. Transactions on Machine Learning Research (2025), https://openreview.net/forum?id=zKv8qULV6n
19. Li, X., Wu, K., Yang, S., Qu, Y., Zhang, G., Chen, Z., Li, J., Mu, J., Hu, X., Fang, W., et al.: Can video generation replace cinematographers? research on the cinematic language of generated video. arXiv preprint arXiv:2412.12223 (2024)
20. Li, Z., Tucker, R., Cole, F., Wang, Q., Jin, L., Ye, V., Kanazawa, A., Holynski, A., Snavely, N.: Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10486–10496 (2025)
21. Lin, Z., Cen, S., Jiang, D., Karhade, J., Wang, H., Mitra, C., Ling, T., Huang, Y., Liu, S., Chen, M., et al.: Towards understanding camera motions in any video. arXiv preprint arXiv:2504.15376 (2025)

22. Liu, H., He, J., Jin, Y., Zheng, D., Dong, Y., Zhang, F., Huang, Z., He, Y., Li, Y., Chen, W., et al.: Shotbench: Expert-level cinematic understanding in vision-language models. arXiv preprint arXiv:2506.21356 (2025)
23. MiniMax: Hailuo. `https://hailuoai.video/` (2024), last accessed 2026/06/29
24. Mou, Z., Xia, B., Huang, Z., Yang, W., Jia, J.: Gradeo: Towards human-like evaluation for text-to-video generation via multi-step reasoning. arXiv preprint arXiv:2503.02341 (2025)
25. Nielsen, J.I., Kau, E., Raskin, R.: Camera movement in narrative cinema: towards a taxonomy of functions. Department of Inf. & Media Studies, University of Aarhus (2007)
26. OpenAI: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
27. Pan, L., Baráth, D., Pollefeys, M., Schönberger, J.L.: Global structure-from-motion revisited. In: European Conference on Computer Vision. pp. 58–77. Springer (2024)
28. Rao, A., Wang, J., Xu, L., Jiang, X., Huang, Q., Zhou, B., Lin, D.: A unified framework for shot type classification based on subject centric lens. In: European Conference on Computer Vision. pp. 17–34. Springer (2020)
29. Savardi, M., Kovács, A.B., Signoroni, A., Benini, S.: Cinescale2: a dataset of cinematic camera features in movies. Data in Brief **51**, 109627 (2023)
30. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
31. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
32. Spottiswoode, R.: A grammar of the film: An analysis of film technique. Univ of California Press (1969)
33. Sun, K., Huang, K., Liu, X., Wu, Y., Xu, Z., Li, Z., Liu, X.: T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8406–8416 (2025)
34. Tang, Y., Guo, J., Hua, H., Liang, S., Feng, M., Li, X., Mao, R., Huang, C., Bi, J., Zhang, Z., et al.: Vidcomposition: Can mllms analyze compositions in compiled videos? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8490–8500 (2025)
35. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
36. Tu, C., Zhang, L., Chen, P., Ye, P., Zeng, X., Cheng, W., Yu, G., Chen, T.: Favorbench: A comprehensive benchmark for fine-grained video motion understanding. arXiv preprint arXiv:2503.14935 (2025)
37. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
38. Wang, J., Yuan, Y., Zheng, R., Lin, Y., Gao, J., Chen, L.Z., Bao, Y., Zhang, Y., Zeng, C., Zhou, Y., et al.: Spatialvid: A large-scale video dataset with spatial annotations. arXiv preprint arXiv:2509.09676 (2025)
39. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)

40. Wang, J., Duan, H., Jia, Z., Zhao, Y., Yang, W.Y., Zhang, Z., Chen, Z., Wang, J., Xing, Y., Zhai, G., et al.: Love: Benchmarking and evaluating text-to-video generation and video-to-text interpretation. arXiv preprint arXiv:2505.12098 (2025)
41. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10510–10522 (2025)
42. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
43. Wang, X., Xu, S., Shan, X., Zhang, Y., Diao, M., Duan, X., Huang, Y., Liang, K., Ma, Z.: Cinetechbench: A benchmark for cinematographic technique understanding and generation. arXiv preprint arXiv:2505.15145 (2025)
44. Wang, Y., Li, X., Yan, Z., He, Y., Yu, J., Zeng, X., Wang, C., Ma, C., Huang, H., Gao, J., et al.: Internvideo2. 5: Empowering video mllms with long and rich context modeling. arXiv preprint arXiv:2501.12386 (2025)
45. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747 (2025)
46. Wu, H., Cai, Y., Ge, H., Chen, H., Yang, M.H., Wang, Y.: Refineshot: Rethinking cinematography understanding with foundational skill evaluation. arXiv preprint arXiv:2510.02423 (2025)
47. Wu, H., Cai, Y., Li, Z., Ge, H., Sun, B., Yuan, J., Wang, Y.: Camreasoner: Reinforcing camera movement understanding via structured spatial reasoning. arXiv preprint arXiv:2602.00181 (2026)
48. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
49. Zhang, M., Wu, T., Tan, J., Liu, Z., Wetzstein, G., Lin, D.: Gendop: Auto-regressive camera trajectory generation as a director of photography. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18229–18239 (2025)
50. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)