ICCV #*****

Towards Understanding Camera Motions in Any Video

Anonymous ICCV submission

Paper ID *****

Abstract

We introduce CameraBench, a large-scale dataset and benchmark designed to assess and improve camera motion understanding. CameraBench consists of ∼3,000 diverse internet videos, annotated by experts through a rigorous multi-stage quality control process. One of our core contributions is a taxonomy or "language" of camera motion primitives, designed in collaboration with cinematographers. We find, for example, that some primitives like "follow" (or `tracking`*) require understanding scene content like moving subjects. We conduct a large-scale human study to quantify human annotation performance, revealing that domain expertise and tutorial-based training can significantly enhance accuracy. For example, a novice may confuse* `zoom-in` *(a change of intrinsics) with translating forward (a change of extrinsics), but can be trained to differentiate the two. Using CameraBench, we evaluate Structure-from-Motion (SfM) and Video-Language Models (VLMs), finding that SfM models struggle to capture semantic primitives that depend on scene content, while VLMs struggle to capture geometric primitives that require precise estimation of trajectories. We then fine-tune a generative VLM on CameraBench to achieve the best of both worlds and showcase its applications, including motion-augmented captioning, video question answering, and video-text retrieval. We hope our taxonomy, benchmark, and tutorials will drive future efforts towards the ultimate goal of understanding camera motions in any video. Project page:* `https://linzhiqiu.github.io/papers/camerabench`

Figure 1. **Examples of camera movements.** We show videos with their camera trajectories: a `tracking shot` of a toddler (row 1, left), Hitchcock's `dolly zoom` effect (row 2, left), Spielberg's dramatic `pan` and `tilt` in *Jurassic Park* (row 3, left), Nolan's `roll` shot in *Inception* (row 1, right), a `pedestal-up` shot from *The Legend of Zelda* (row 2, right), and a selfie by an amateur photographer, `arcing` to showcase the scenery while centering themselves (row 3, right). Please watch the videos at our website.

1. Introduction

> *We must perceive in order to move, but we must also move in order to perceive.*
>
> — J. J. Gibson, *The Ecological Approach to Visual Perception* [19]

Humans perceive the visual world through movement. Motion parallax [50], for instance, enables precise depth perception essential for navigating the physical world [18]. Similarly, camera motion is crucial for modern vision techniques that process videos of dynamic scenes. For example, Structure-from-Motion (SfM) [51, 59, 73] and Simultaneous Localization and Mapping (SLAM) [12, 16, 55] methods must first estimate camera motion (pose trajectory) to reconstruct the scenes in 4D. Likewise, without understanding camera motion, video-language models (VLMs) [57, 67, 70] would not fully perceive, reason about, or generate video dynamics.

Human perception of camera motion. Understanding camera motion comes naturally to humans because we intuitively grasp the "*invisible subject*" – the camera operator who shapes the video's viewpoint, framing, and narrative. For example, in a video tracking a child's first steps, one can sense a parent's joy through their handheld, shaky movement. Professional cinematographers and filmmakers even use camera motion as a tool [13, 54] to enhance visual storytelling and amplify the emotional impact of their shots. Hitchcock's iconic `dolly zoom` moves the camera forward while zooming out, maintaining the subject's framing while altering the background to create the impression of vertigo. In *Jurassic Park* (1993), Spielberg uses a slow `upward tilt` and `rightward pan` to evoke a sense of awe as the protagonists (and the audience) first see the dinosaurs. In *Inception* (2010), Nolan uses a camera `roll` to mirror shifting gravity, blurring the line of reality. Similarly, game developers use camera movement to enhance player immersion. In *Legend of Zelda: Breath of the Wild* (2017), a smooth `pedestal-up` shot transitions from the character's viewpoint to a breathtaking aerial view, hinting at the journey ahead. Even amateur photographers use camera motion as a tool; for example, selfie videos allow one to play the role of both the cinematographer and the subject. See Figure 1 for examples.

Computational approaches to camera motion. In contrast, classic computer vision methods learn camera motion from what is "visible" in the frame, relying on techniques like SfM and SLAM to estimate camera poses from video sequences. While these geometry-based approaches perform well on simple, static scenes, it is unclear how well they generalize to *dynamic, real-world videos* due to the difficulty of separating camera motion from scene dynamics [38, 61]. Moreover, these approaches do not capture the *high-level semantics* of camera motion [54], such as the intent behind a shot (e.g., tracking a subject or revealing a scene) or the context in which the motion occurs (e.g., handheld, gimbal-stabilized, or vehicle-mounted). On the other hand, recent multimodal vision systems like GPT-4o and Gemini [45, 48, 57] show strong human-like perceptual capabilities through large-scale training, yet their ability to understand camera motion remains largely untested. Inspired by these end-to-end approaches, we propose a *data-driven* framework for benchmarking and developing models that can perceive camera motion as humans do. However, this seemingly straightforward task poses challenges overlooked by prior work, as we detail next.

Challenges and our approach. We find major issues in widely-used datasets with camera motion annotations, such as MovieNet [27], AVE [1], and DREAM-1K [60]. First, many **lack a clear or correct specification of motion types**, often conflating fundamental concepts like translation with rotation or zoom. Second, these datasets often assign **contradictory labels** to the same video (e.g., labeling a video as both static and moving, which are mutually exclusive). Third, they **lack careful oversight**, resulting in significant annotation errors. To address these issues, we collaborate with professional cinematographers to develop a comprehensive taxonomy, a robust label-then-caption framework, and a training program backed by a large-scale human study to improve annotation quality. These efforts allow us to scale over 150K high-quality annotations across 3,381 videos.

CameraBench. We introduce **CameraBench** to benchmark and develop models for human-like understanding of camera motion, using our initial set of videos (each reviewed by at least one author during the quality control phase). Our comprehensive annotations, which include both labels and captions, allow us to evaluate models on a wide range of tasks, including binary classification of motion primitives, video-text retrieval, video captioning, and video question-answering (VQA). We evaluate a diverse set of 20 models, including discriminative [34, 35, 39, 48, 63] and generative VLMs [4, 33, 40, 45, 57, 72], and SfM/SLAM [38, 59, 61] methods. Although not all models can perform every task (e.g., SfM/SLAM cannot perform VQA tasks or reason about object-centric motion), we ensure fair comparisons by carefully designing the benchmarking protocol.

Findings. We find that classic SfM/SLAM methods [51] often fail to handle dynamic or low-parallax scenes (e.g, when the camera is stationary or only rotating), thus struggling with even classifying basic motion primitives (e.g., "*Is the camera moving up or not?*"). We also observe that recent learning-based SfM/SLAM methods like MegaSAM [38, 61] handle dynamic scenes much better and outperform the classic COLMAP [51] by 1-2x. However, they may still confuse camera motion with object or scene motion in complex scenarios. We argue that our benchmark serves as a *reality check* for future SfM/SLAM methods, helping identify areas for improvement. On the other hand, we find that generative VLMs show promise in understanding camera motion, particularly in tasks requiring semantic reasoning (e.g., tracking shot). This motivates us to use our dataset to post-train VLMs for better camera motion understanding. With our small-scale yet high-quality fine-tuning data, we show that VLMs can achieve 1-2x improvements across both discriminative and generative tasks.

Contributions. We (1) introduce a taxonomy of camera motion primitives, developed in collaboration with domain experts; (2) design a robust annotation framework and training program to improve data quality; (3) collect a benchmark featuring real-world videos of dynamic scenes across diverse genres and motions; and (4) analyze the strengths and limitations of existing models to guide future research. We hope our data, taxonomy, and models can improve understanding of camera motions in any video.

ICCV
#*****

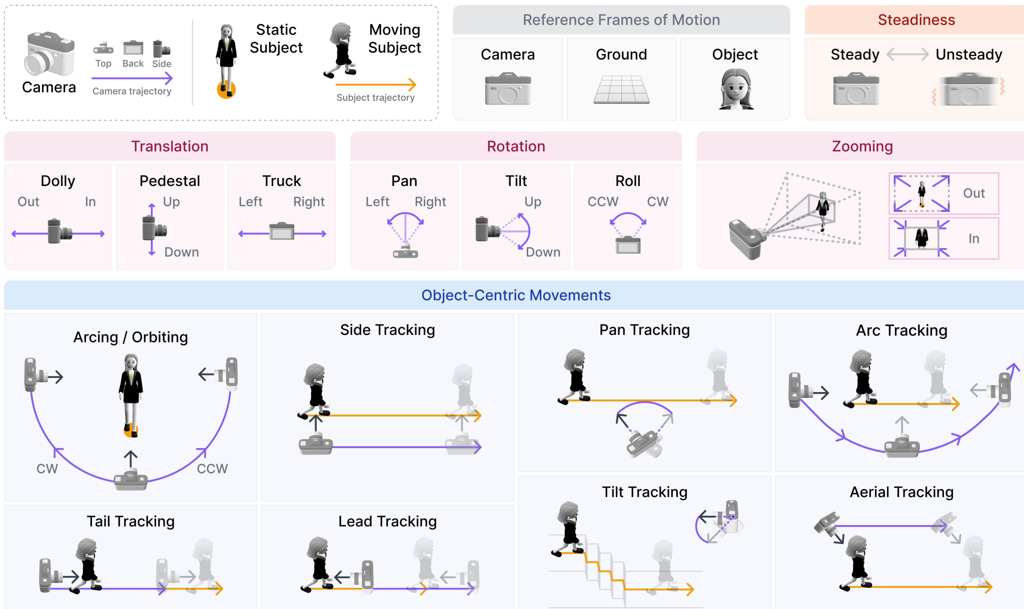

Figure 2. **Taxonomy of camera motion primitives.** Our taxonomy, developed in collaboration with cinematographers and vision researchers, is the first to comprehensively capture camera motion across object-, ground-, and camera-centric reference frames, using precise cinematography terms [13] to eliminate ambiguity. It covers camera steadiness, translation, rotation, intrinsic changes, and common object-centric movements, all detailed in this paper. We refine the taxonomy iteratively over three months by annotating real-world videos and incorporating feedback from researchers and cinematographers to ensure both accuracy and completeness.

2. CameraBench for Motion Understanding

We repurpose our motion primitive labels and captions for both **discriminative** (classification, retrieval) and **generative** (VQA, captioning) tasks.

Baselines. We evaluate a diverse set of **20 models**, including **6 SfM/SLAM** methods: COLMAP [51] and learning-based variants such as MegaSAM [38], CUT3R [61], and others [14, 59, 62]. We also report **3 discriminative VLMs** [35, 77] like InternVideo2 [63] and **11 generative VLMs** including Qwen2.5-VL [4], GPT-4o [45], and LLaVA-Video [72], among others [33, 57, 63, 70, 71].

Classification of motion primitives. We evaluate models on binary classification of motion primitives, restricted to those defined in the camera-centric frame to align with SfM/SLAM outputs. For SfM/SLAM, we compute the seven degrees of translation, rotation, and focal change from estimated camera extrinsics and intrinsics (if available) between the first and last frame. For discriminative VLMs, we use textual definitions of each primitive (“*The camera pans to the left.*”) to compute matching scores. For generative VLMs, we compute VQAScore [41], i.e., the probability of “Yes” to a binary question (“*Does the camera pan to the left?*”). Appendix K details prompts for VLMs.

Results. Table 1 shows that (1) learning-based SfM/SLAM methods like MegaSAM significantly outperform COLMAP and set the state-of-the-art. Nonetheless, no methods fully solve this task, as the best overall AP remains ∼50%. Figure 7 shows failure cases, e.g., SfM/SLAM struggles with low-parallax (rotation only) scenes. (2) While weaker than SfM/SLAM, generative VLMs like GPT-4o show promising results, significantly outperforming discriminative VLMs. This motivates us to fine-tune Qwen2.5-VL using supervised fine-tuning (SFT) on a separate set of ∼1400 videos (with no overlap with the testset). Despite the small dataset size, our SFT model achieves ∼2x performance, matching that of MegaSAM. We note that certain motions like `roll` remain particularly challenging for VLMs, likely due to their long-tailed nature [46] in internet videos.

Beyond camera-centric motion primitives. We collect ∼10K VQA samples across 9 top-level skills and 81 sub-tasks. Crucially, these tasks go beyond camera-centric frame reasoning to evaluate more aspects such as object-centric motion, scene dynamics, steadiness, and more. Some tasks also require logical (e.g., verifying if *only one* motion type exists or if a motion is *absent*) and linguistic reasoning (e.g., checking if a motion description is accurate). We follow community best practices [20, 32], pairing each question with two videos with opposite answers so that models cannot answer blindly without seeing the video (see Figure 6).

VQA results. Table 3 shows that all open-source VQA models perform at or below chance on CameraBench. Nonetheless, our SFT model – fine-tuned on our small training set – achieves state-of-the-art results across all skills, especially the most challenging ones

ICCV #*****

Table 1. **Binary classification on motion primitives defined in the camera-centric frame.** We report Average Precision per primitive. We find that (1) recent SfM/SLAM methods like MegaSAM [38] significantly outperform COLMAP [51], but all methods remain far from solving this task with ∼50% AP. (2) Generative VLMs clearly outperform discriminative ones. Motivated by this, we fine-tune Qwen2.5-VL [4] on a separate training set of ∼1400 videos (no overlap with the test set). We show that simple SFT (highlighted in green) significantly boosts performance by 1-2x, making it match the SOTA MegaSAM in overall AP. We **bold** the best and underline the second-best results; finetuned models are ranked separately.

Model	Translation (Dolly/Pedestal/Truck)						Zooming		Rotation (Pan/Tilt/Roll)						Static	Avg
	In	Out	Up	Down	Right	Left	In	Out	Right	Left	Up	Down	CW	CCW		
Random Chance	29.3	9.7	6.7	8.6	15.8	11.5	11.1	10.2	15.0	15.4	12.7	7.7	8.9	10.2	9.7	12.2
SfM/SLAM																
COLMAP	36.2	13.1	11.9	19.7	34.1	30.0	13.9	14.2	43.9	46.4	28.3	19.1	42.1	48.7	7.5	27.3
VGGSFM	56.6	28.9	28.7	38.2	**48.9**	35.3	21.7	17.3	60.9	58.7	46.6	43.3	61.4	55.5	16.7	41.3
DUSt3R	58.9	24.0	**30.7**	18.0	38.3	26.9	18.2	24.6	59.4	63.8	32.9	27.3	61.0	57.9	13.1	37.0
MASt3R	47.5	21.1	23.5	**40.2**	38.7	38.1	**42.2**	**46.6**	66.6	58.0	63.2	40.3	50.4	53.5	15.7	43.1
CUT3R	68.9	**50.4**	24.7	34.2	37.0	27.6	15.9	21.3	59.1	65.0	65.0	47.5	60.7	66.2	15.1	42.7
MegaSAM	73.8	43.9	24.2	29.1	45.3	**44.2**	11.1	10.2	**79.5**	**82.2**	**73.8**	**65.3**	**71.5**	**75.8**	22.0	**50.1**
CLIPScore																
UMT-B16-CLIP	27.0	10.4	9.0	20.0	19.4	11.8	11.8	9.9	11.9	13.5	13.1	8.4	18.8	15.6	10.0	14.0
UMT-L16-CLIP	27.2	9.8	12.3	10.8	18.5	11.5	17.5	8.9	16.0	17.4	21.9	8.3	7.3	10.0	13.0	14.0
LanguageBind-CLIP	32.7	13.2	7.8	11.2	14.2	11.7	14.4	9.4	20.1	16.4	14.1	8.5	13.8	9.5	10.9	13.9
LanguageBindV1.5-CLIP	33.6	14.5	11.0	10.3	15.0	11.8	14.2	10.1	19.9	16.7	16.1	9.2	17.6	10.2	10.4	14.7
InternVideo2-S2-CLIP	41.7	9.4	5.8	9.7	15.0	12.0	15.0	9.9	20.6	18.8	14.7	9.1	8.3	10.8	11.4	14.2
ITMScore																
UMT-B16-ITM	31.7	11.5	11.4	14.3	16.6	12.8	12.3	9.2	15.1	16.9	16.2	10.0	14.2	12.1	8.9	14.2
UMT-L16-ITM	40.6	10.6	8.5	17.6	21.9	23.6	12.4	9.8	21.3	33.2	31.0	11.2	13.5	12.3	9.4	18.4
InternVideo2-S2-ITM	52.4	12.6	10.5	14.7	15.8	19.7	21.1	16.7	29.4	29.1	24.5	18.4	17.2	13.4	14.0	20.6
VQAScore																
LLaVA-OneVision-7B	46.8	13.5	12.6	16.9	23.7	20.2	10.7	14.4	33.5	33.6	16.9	31.4	19.3	20.8	18.8	22.2
LLaVA-Video-7B	54.7	15.2	16.5	19.3	27.1	23.6	16.2	16.9	33.6	36.8	26.9	37.2	16.1	21.7	22.1	25.6
InternVideo2-Chat-8B	69.9	18.5	19.3	17.6	17.9	23.4	12.2	10.4	22.6	22.7	17.2	22.8	19.6	16.4	20.2	22.0
Tarsier-Recap-7B	59.7	15.1	25.7	23.7	28.8	21.5	14.4	15.0	22.8	27.3	24.6	21.6	15.2	18.7	30.7	21.0
InternLMXComposer2.5-7B	49.0	10.6	11.4	10.4	14.6	10.6	11.8	16.5	14.3	13.9	14.7	17.5	11.7	18.1	21.8	16.5
InternVL2.5-8B	67.9	12.9	28.1	25.9	23.4	23.2	18.6	32.1	37.4	30.9	37.6	36.9	11.5	25.3	23.4	29.5
InternVL2.5-26B	63.6	11.8	21.1	23.6	27.2	19.4	21.8	31.6	42.5	38.3	44.9	43.6	14.3	18.2	25.1	29.8
mPLUG-Owl3-7B	47.6	12.9	13.9	16.9	17.3	18.5	12.9	10.6	31.4	26.6	26.1	37.0	10.4	12.2	17.8	20.8
GPT-4o	66.3	29.2	21.1	38.2	38.0	21.9	41.7	39.3	44.7	42.1	43.6	35.5	24.0	28.7	**32.0**	36.4
InternVL3-8B	61.2	15.5	18.8	29.0	30.5	27.3	29.5	28.1	41.6	49.3	42.0	36.5	21.3	22.3	20.1	31.5
InternVL3-78B	**72.0**	18.2	19.6	32.5	33.8	29.4	26.4	33.4	47.2	53.5	47.8	40.3	27.6	25.0	22.6	36.8
Qwen2.5-VL-7B	63.0	14.1	20.1	22.3	28.5	27.7	23.2	27.2	36.5	44.6	38.4	25.7	26.0	25.5	20.2	29.5
Qwen2.5-VL-32B	66.8	19.1	11.1	31.4	32.1	30.4	27.8	32.6	43.2	50.0	53.2	44.0	26.6	29.0	28.8	35.1
Qwen2.5-VL-72B	67.2	19.1	12.8	26.5	33.3	26.1	27.5	41.2	50.6	46.8	53.4	31.0	33.3	30.9	29.1	35.3
Qwen2.5-VL-7B (Ours SFT)	83.9	38.6	27.8	47.8	67.9	50.0	54.5	75.8	79.2	83.8	76.3	67.6	32.3	41.0	73.6	60.0
Qwen2.5-VL-32B (Ours SFT)	85.6	40.1	29.3	49.4	69.6	51.5	56.0	77.3	80.7	85.4	77.9	69.2	33.9	42.7	75.4	61.6
Qwen2.5-VL-72B (Ours SFT)	**86.8**	**41.3**	**30.5**	**50.6**	**70.7**	**52.6**	**57.1**	**78.5**	**81.9**	**86.6**	**79.1**	**70.4**	**35.0**	**43.8**	**76.6**	**62.8**

(e.g., Tracking Shot and Only Motion) that require object-centric and logical reasoning.

Other tasks. We summarize key findings: (1) **Captioning** (Figure 8). We prompt VLMs with "*Describe the camera movements in this video*". Our SFT model generates more accurate captions than state-of-the-art VLMs, both qualitatively and quantitatively, as measured by metrics like SPICE and LLM-as-a-Judge. (2) **Video-text retrieval** (Table 4). We use video pairs in CameraBench's VQA tasks to evaluate retrieval performance and show that generative VLMs (using the discriminative VQAScore [41]), outperform other baselines. (3) **Motion control in image-to-video generation** (Figure 17). While we focus on video understanding, we note that finetuning CogVideoX1.5-I2V [69] using CameraBench can potentially improve its camera motion control.

3. Conclusion

In conclusion, we take the first step toward human-like camera motion understanding by introducing a taxonomy of motion primitives and a robust annotation framework, developed in collaboration with cinematographers. We implement a training program to transform laypeople into proficient annotators of camera movements. We curate a diverse benchmark to analyze existing models and suggest directions for future improvement. Lastly, we show that our high-quality dataset can be used to fine-tune VLMs for improved camera motion understanding.

References

[1] Dawit Mureja Argaw, Fabian Caba Heilbron, Joon-Young Lee, Markus Woodson, and In So Kweon. The anatomy of video editing: A dataset and benchmark suite for ai-assisted video editing. In *European Conference on Computer Vision*, pages 201–218. Springer, 2022. 2, 9, 10, 11, 12

[2] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *arXiv preprint arXiv:2411.18673*, 2024. 9

[3] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024. 9

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 4, 19

[5] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024. 9, 10, 11, 12

[6] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Juncheng Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengchen Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025. 9

[7] Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. *arXiv preprint arXiv:2502.04896*, 2025. 9

[8] Soon Yau Cheong, Duygu Ceylan, Armin Mustafa, Andrew Gilbert, and Chun-Hao Paul Huang. Boosting camera motion control for video diffusion transformers. *arXiv preprint arXiv:2410.10802*, 2024. 9

[9] Alessandro Chiuso, Roger Brockett, and Stefano Soatto. Optimal structure from motion: Local ambiguities and global estimates. *International journal of computer vision*, 39:195–228, 2000. 9

[10] Robin Courant, Nicolas Dufour, Xi Wang, Marc Christie, and Vicky Kalogeiton. Et the exceptional trajectories: Text-to-camera-trajectory generation with character awareness. In *European Conference on Computer Vision*, pages 464–480. Springer, 2024. 9

[11] Kostas Daniilidis and Minas E Spetsakis. Understanding noise sensitivity in structure from motion. In *Visual Navigation*, pages 60–88. Psychology Press, 2013. 9

[12] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1052–1067, 2007. 1

[13] Kyle Deguzman. Types of camera movements in film explained: Definitive guide, 2020. 2, 3, 10

[14] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. *arXiv preprint arXiv:2409.19152*, 2024. 3, 20

[15] Dimitris Eleftheriotis. *Cinematic journeys: Film and movement*. Edinburgh University Press, 2010. 10

[16] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014. 1

[17] Cornelia Fermüller and Yiannis Aloimonos. Ambiguity in structure from motion: Sphere versus plane. *International Journal of Computer Vision*, 28:137–154, 1998. 9

[18] Steven H Ferris. Motion parallax and absolute distance. *Journal of experimental psychology*, 95(2):258, 1972. 1

[19] James J Gibson. The ecological approach to visual perception. 2003. 1

[20] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 3, 13

[21] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 9

[22] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 21

[23] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 9

[24] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 18

[25] Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. *arXiv preprint arXiv:2501.02955*, 2025. 9

[26] Yunzhong Hou, Liang Zheng, and Philip Torr. Learning camera movement control from real-world drone videos. *arXiv preprint arXiv:2412.09620*, 2024. 9

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

Proceedings, Part IV 16, pages 709–727. Springer, 2020. 2, 9, 10, 11, 12

[28] Hongda Jiang, Xi Wang, Marc Christie, Libin Liu, and Baoquan Chen. Cinematographic camera diffusion model. In *Computer Graphics Forum*, page e15055. Wiley Online Library, 2024. 9

[29] Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Holynski. Stereo4d: Learning how things move in 3d from internet stereo videos. *arXiv preprint arXiv:2412.09621*, 2024. 9

[30] Guojun Lei, Chi Wang, Hong Li, Rong Zhang, Yikai Wang, and Weiwei Xu. Animateanything: Consistent and controllable animation for video generation. *arXiv preprint arXiv:2411.10836*, 2024. 9

[31] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *The First Workshop on the Evaluation of Generative Foundation Models at CVPR*, 2024. 19

[32] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 3, 13, 14, 22

[33] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 3

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 19

[35] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19948–19960, 2023. 2, 3

[36] Teng Li, Guangcong Zheng, Rui Jiang, Tao Wu, Yehao Lu, Yining Lin, Xi Li, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. *arXiv preprint arXiv:2502.10059*, 2025. 9

[37] Xiaozhe Li, Kai Wu, Siyi Yang, YiZhan Qu, Guohua Zhang, Zhiyu Chen, Jiayao Li, Jiangchuan Mu, Xiaobin Hu, Wen Fang, et al. Can video generation replace cinematographers? research on the cinematic language of generated video. *arXiv preprint arXiv:2412.12223*, 2024. 9, 10, 11, 12, 15

[38] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast, and robust structure and motion from casual dynamic videos. *arXiv preprint arXiv:2412.04463*, 2024. 2, 3, 4, 14, 21

[39] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models, 2023. 2, 21

[40] Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*, 2024. 2, 13

[41] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. 3, 4, 14, 19

[42] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 9

[43] Shihong Liu, Zhiqiu Lin, Samuel Yu, Ryan Lee, Tiffany Ling, Deepak Pathak, and Deva Ramanan. Language models as black-box optimizers for vision-language models. *arXiv preprint arXiv:2309.05950*, 2024. 21

[44] Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Chatcam: Empowering camera control through conversational ai. *Advances in Neural Information Processing Systems*, 37:54483–54506, 2025. 9

[45] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3

[46] Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails of vision-language models. *arXiv preprint arXiv:2401.12425*, 2024. 3

[47] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv e-prints*, pages arXiv–2410, 2024. 9

[48] 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. 2, 18

[49] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 17–34. Springer, 2020. 9, 10, 11, 12

[50] Brian Rogers and Maureen Graham. Motion parallax as an independent cue for depth perception. *Perception*, 8 (2):125–134, 1979. 1, 11

[51] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 2, 3, 4, 9, 14, 20

[52] Xincheng Shuai, Henghui Ding, Zhenyuan Qin, Hao Luo, Xingjun Ma, and Dacheng Tao. Free-form motion control: A synthetic video generation dataset with

controllable camera and object motions. *arXiv preprint arXiv:2501.01425*, 2025. 9

[53] Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11218–11221, 2024. 10

[54] Raymond Spottiswoode. *A grammar of the film: An analysis of film technique*. Univ of California Press, 1969. 2, 10

[55] Takafumi Taketomi, Hideaki Uchiyama, and Sei Ikeda. Visual slam algorithms: A survey from 2010 to 2016. *IPSJ transactions on computer vision and applications*, 9 (1):16, 2017. 1

[56] Yunlong Tang, Junjia Guo, Hang Hua, Susan Liang, Mingqian Feng, Xinyang Li, Rui Mao, Chao Huang, Jing Bi, Zeliang Zhang, et al. Vidcomposition: Can mllms analyze compositions in compiled videos? *arXiv preprint arXiv:2411.10979*, 2024. 9, 10, 11, 12, 15

[57] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 2, 3, 19

[58] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 22

[59] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21686–21697, 2024. 1, 2, 3, 20

[60] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024. 2, 9, 10, 11, 12, 15

[61] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025. 2, 3, 20

[62] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3, 20

[63] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 2, 3

[64] Yuelei Wang, Jian Zhang, Pengtao Jiang, Hao Zhang, Jinwei Chen, and Bo Li. Cpa: Camera-pose-awareness diffusion transformer for video generation. *arXiv preprint arXiv:2412.01429*, 2024. 9

[65] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *Advances in Neural Information Processing Systems*, 37:34322–34348, 2025. 9

[66] Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Trajectory attention for fine-grained video motion control. *arXiv preprint arXiv:2411.19324*, 2024. 9

[67] Jinbo Xing, Long Mai, Cusuh Ham, Jiahui Huang, Aniruddha Mahapatra, Chi-Wing Fu, Tien-Tsin Wong, and Feng Liu. Motioncanvas: Cinematic shot design with controllable image-to-video generation. *arXiv preprint arXiv:2502.04299*, 2025. 1, 9

[68] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 9

[69] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 4

[70] Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888*, 2025. 1, 3, 9

[71] Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, Shuangrui Ding, Qipeng Guo, Haodong Duan, Xin Chen, Han Lv, Zheng Nie, Min Zhang, Bin Wang, Wenwei Zhang, Xinyue Zhang, Jiaye Ge, Wei Li, Jingwen Li, Zhongying Tu, Conghui He, Xingcheng Zhang, Kai Chen, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2.5-omnilive: A comprehensive multimodal system for long-term streaming video and audio interactions. *arXiv preprint arXiv:2412.09596*, 2024. 3

[72] 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. 2, 3

[73] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *European Conference on Computer Vision*, pages 20–37. Springer, 2022. 1

[74] Sixiao Zheng, Zimian Peng, Yanpeng Zhou, Yi Zhu, Hang Xu, Xiangru Huang, and Yanwei Fu. Vidcraft3: Camera, object, and lighting control for image-to-video generation. *arXiv preprint arXiv:2502.07531*, 2025. 9

[75] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 9

[76] Zhenghong Zhou, Jie An, and Jiebo Luo. Latent-reframe:

Enabling camera control for video diffusion model without training. *arXiv preprint arXiv:2412.06029*, 2024. 9

[77] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Wang HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment, 2023. 3