Geometry-Guided Camera Motion Understanding in VideoLLMs

Haoan Feng[1], Sri Harsha Musunuri[2], Guan-Ming Su[2],
[1]University of Maryland, College Park, [2]Dolby Laboratories Inc.

hfengac@umd.edu, harsha.musu@gmail.com, guanmingsu@ieee.org

Abstract

*Camera motion is a fundamental geometric signal that shapes visual perception and cinematic style, yet current video-capable vision-language models (VideoLLMs) rarely represent it explicitly and often fail on fine-grained motion primitives. We address this gap with a framework of **benchmarking**, **diagnosis**, and **injection**. We derive **CameraMotionDataset**, a VQA benchmark built on an existing synthetic dataset (MultiCamVideo Dataset) with explicit camera control, formulate camera motion as constraint-aware multi-label recognition, and construct a multiple-choice evaluation protocol–**CameraMotionVQA**. Across diverse off-the-shelf VideoLLMs, we observe substantial errors in recognizing camera motion primitives. Probing experiments on a Qwen2.5-VL vision encoder suggest that camera motion cues are weakly represented, especially in deeper ViT blocks, helping explain the observed failure modes. To bridge this gap without costly training or fine-tuning, we propose a lightweight, model-agnostic pipeline that extracts geometric camera cues from 3D foundation models (3DFMs), predicts constrained motion primitives with a temporal classifier, and injects them into downstream VideoLLM inference via structured prompting. Experiments demonstrate improved motion recognition and more camera-aware model responses, highlighting geometry-driven cue extraction and structured prompting as practical steps toward a camera-aware VideoLLM and VLA system. Dataset and benchmark are available at [1].*

1. Introduction

Video-capable vision-language models (VideoLLMs) have improved substantially on high-level video semantics, including recognition of objects, actions, and narrative events across diverse video lengths [33]. However, an essential component of video meaning, especially in edited content such as films, TV series, and online compilations, lies not only in *what* appears in the frames, but also in *how* it is captured. Camera motion (*e.g.*, `pan`, `tilt`, and `dolly`) is a core cinematographic device that guides attention, reveals spatial layout, and communicates the author's intent [23]. Accordingly, camera motion is both central to film grammar and useful for camera-aware description, layout-oriented retrieval, and spatial reasoning.

Despite its importance, current VideoLLMs remain unreliable in recognizing fine-grained camera-motion primitives. Camera motion is a spatiotemporal geometric signal that is not localized to any single frame, and it is easily confounded by object motion, cuts, and motion blur. Consequently, models with strong frame-level perception may still fail to model the camera as the source of the visual stream. Moreover, as we later show, common VideoLLM pipelines compress visual tokens with network depth, which can attenuate motion-sensitive cues. This gap is unsurprising, as most large-scale video captioning and VQA corpora lack explicit supervision for camera motion.

To study this gap under controlled conditions, we introduce **CameraMotionDataset**, a synthetic dataset of 12k within-shot segments annotated with fine-grained camera-motion labels. A *shot* is a temporally contiguous sequence of frames captured by a single camera without cuts; we focus on non-overlapping 1-second segments within each shot. Each segment is annotated with a set of camera-motion labels drawn from a fixed taxonomy of 15 atomic motions. Built on this dataset, **CameraMotionVQA** provides a multiple-choice benchmark that evaluates open-source VideoLLMs under a standardized VQA protocol.

A central hypothesis of this work is that reliable camera-motion cues can be derived from models with strong geometric and 3D reasoning capabilities *without modifying* the VideoLLM backbone. Specifically, we use a frozen 3D foundation model (VGGT [35]) to extract camera cues, train a lightweight temporal classifier to predict constrained motion primitives, and inject the predictions into downstream VideoLLMs via structured prompting.

Beyond performance gains, we seek to understand *where*

[1] https://huggingface.co/datasets/fengyee/camera-motion-dataset-and-benchmark

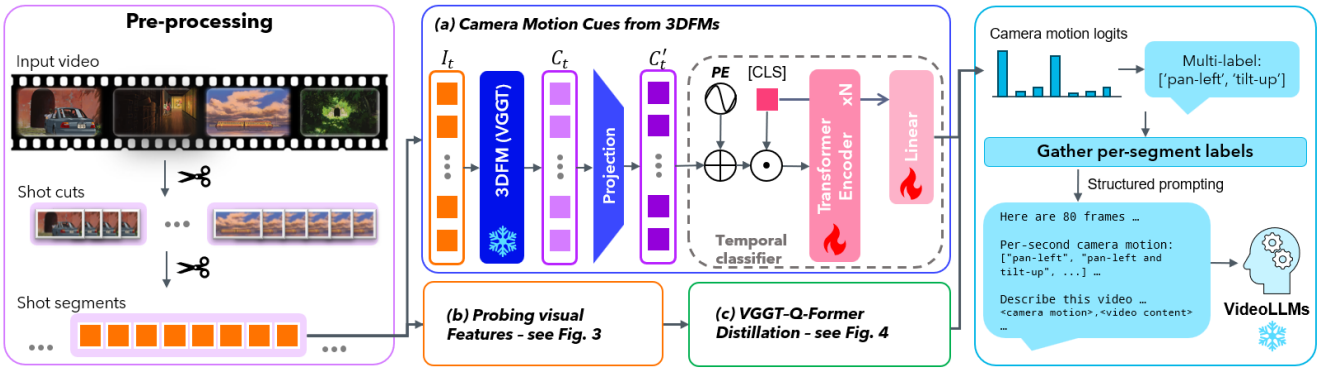

Figure 1. **Overall pipeline.** Camera cues are extracted from a frozen 3DFM (VGGT) and passed to a Transformer-based temporal classifier to predict camera-motion primitives, and per-second motions are injected as a structured prompt field for VideoLLMs without modifying VideoLLM weights. For clarity, the probing and distillation pipelines are shown separately in Fig. 3 and Fig. 4.

camera-motion information is lost in typical VideoLLM pipelines. By probing intermediate vision-encoder features with Q-Former-style query tokens [18], we observe that camera-motion cues become progressively less recoverable from shallow to deeper blocks, consistent with our hypothesis on token compression and the lack of explicit motion supervision. These findings motivate the use of geometry-derived camera cues from 3DFMs as an external, plug-in signal. Our contributions are threefold:

- **CameraMotionDataset and CameraMotionVQA.** We derive a shot-consistent dataset with fine-grained camera-motion labels from an existing synthetic corpus with precise camera parameters, and construct a multiple-choice VQA benchmark for evaluating open-source VideoLLMs.
- **Camera-cue assisted motion recognition and structured prompting.** We propose a model-agnostic pipeline that extracts camera cues from a frozen 3DFM (VGGT), predicts motion primitives with a lightweight temporal classifier, and injects them into frozen VideoLLM inference via structured prompting (Fig. 1).
- **Probing-based diagnosis.** We analyze where and why camera-motion cues are lost in VideoLLM vision encoders, helping explain failure modes on CameraMotionVQA and motivating our geometry-aware cue injection.

2. Related Work

2.1. Camera motion in cinematic video understanding

Cinematography depends on both *what* is shown and *how* it is filmed (camera, framing, composition). Recent benchmarks evaluate film-grammar attributes beyond generic video QA. CameraBench [23] defines a cinematographer-informed motion taxonomy, revealing confusions between extrinsic motion (*e.g.*, `dolly`) and intrinsic change (*e.g.*, `zoom`). CineTechBench [37], ShotBench [24], and VidComposition [34] extend evaluation to multi-attribute cinematography, while Shot-by-Shot [40] shows shot-level cues can steer description generation via prompting.

Existing datasets range from per-frame cinematic attributes [16, 20, 28–30] to geometry-grounded supervision: SpatialVID [36] provides per-frame depth and pose-derived instructions, and OmniTr [43] encodes motion as parameterized programs linking language to trajectories. However, none of them explicitly offer geometry-consistent motion primitives at short-segment granularity.

Camera motion also serves as a control signal in video generation. Recent models [1, 2, 14, 15, 25, 26, 39, 42, 45**?**] condition generation on explicit trajectories, camera embeddings, or 6-DoF poses; MotionCtrl [39] and Direct-a-Video [42] decouple camera and object motion control, while others manipulate object poses [22, 27] or survey multimodal-guided editing [32]. ReCamMaster [2] releases *MultiCamVideo Dataset*, from which we derive our benchmark. These generation-oriented parameterizations also motivate reliable *recognition* of fine-grained primitives under compositional and exclusivity constraints [15].

At the *object* level, MeViS [7, 8] benchmarks referring segmentation with motion expressions, MOSEv2 [9] addresses complex-scene video object segmentation (VOS), and MOVE [44] adds motion-guided few-shot VOS. These works target object rather than camera motion, but share the motivation to move beyond appearance-based features.

2.2. Vision-language models and VideoLLMs

VideoLLMs extend image VLMs to video by encoding sampled frames, compressing visual tokens, and conditioning an LLM for instruction following, captioning, and QA. This "frame aggregation + LLM reasoning" paradigm captures high-level semantics, but token compression can attenuate subtle temporal cues, including camera motion.

Open-source VideoLLMs can be organized by their temporal design and videoLLM interface. Strong-backbone models (*e.g.*, Qwen2.5-VL [3] and InternVL [4]) strengthen the vision tower and time/position encoding to better support long-context video. Connector-based models [19, 41,

49] keep the LLM mostly fixed and add lightweight temporal modules with aggressive token reduction to fit multi-frame inputs. Instruction-tuned assistants [5, 47] emphasize curated video-instruction data and post-training to improve temporal reasoning without major architectural changes. Video-native systems [17, 38, 46] use stronger video encoders and/or video-centric pretraining to improve spatiotemporal representations.

Bridging modules offer a parameter-efficient way to integrate video into LLMs. Query bottlenecks (*e.g.*, Q-Former) summarize frame features into a small set of query tokens, improving efficiency over dense patch tokens [18, 48]. However, cinematography-centric benchmarks still show weak fine-grained camera-motion recognition, suggesting recent progress is driven more by sampling, grounding, and alignment than explicit camera-motion representation.

2.3. 3D foundation models and geometric cues for VideoLLMs

3D foundation models (3DFMs) learn transferable 3D representations and infer geometric attributes from visual input. VGGT [35] is a feed-forward geometry transformer that jointly predicts camera parameters, depth/point maps, and tracking cues from one or multiple views, exposing camera-aware tokens as geometric cues. Follow-up work improves efficiency [31], and learning-based SfM systems amortize reconstruction robustness [6, 10], expanding practical sources of camera geometry.

Recent efforts connect geometric priors to vision-language reasoning via explicit geometry branches [50] or distillation from frozen 3DFMs into compact tokens [13, 21]. VLM-3R [12] takes a tranable approach, augmenting VLMs with instruction-aligned 3D reconstruction that fuses per-view camera tokens with appearance features through end-to-end training. In contrast, our pipeline injects camera cues via structured prompting without modifying VideoLLM weights, making it model-agnostic and complementary to trainable approaches like VLM-3R.

Summary. Benchmarks position camera motion at the intersection of geometry and semantics: although datasets provide explicit trajectories, VideoLLMs struggle to recover fine-grained motion primitives. We therefore leverage off-the-shelf 3DFMs to extract camera cues, predict constraint-consistent motion labels, and inject them into frozen VideoLLMs via structured prompting.

3. Method

3.1. Problem overview

Given an input video, we split it into shots via an off-the-shelf detector (*e.g.*, Shot-by-Shot [40]) and divide each shot into non-overlapping 1-second segments. Each segment is assigned a multi-hot label $\mathbf{y} \in \{0,1\}^K$ over K=15 atomic primitives; co-occurrence is allowed (*e.g.*, `pan-left` + `tilt-up`) but mutually exclusive pairs (*e.g.*, `pan-left` vs. `pan-right`) are forbidden. Our module is lightweight and plug-and-play: a frozen 3DFM extracts camera cues, a temporal classifier predicts motion primitives, and the per-second predictions are injected into a structured prompt for downstream VideoLLMs (see Fig. 1). We adopt VGGT as the 3DFM because it produces per-frame *camera tokens* in a single forward pass, encoding pose and dynamics in a unified coordinate system—cues that we find are not well-preserved by VideoLLM vision encoders.

3.2. CameraMotionDataset construction and CameraMotionVQA benchmark

We build a labeled camera-motion dataset from the *MultiCamVideo* dataset introduced in ReCamMaster [2]. MultiCamVideo contains photorealistic dynamic scenes rendered in Unreal Engine 5 [11], with dense frame-wise camera calibration (intrinsics) and camera poses (extrinsics). It covers 13.6K dynamic scenes and includes 136K videos with 112K distinct camera trajectories; videos are rendered at 15 fps, and trajectories are explicit camera extrinsic sequences.

As shown in Fig. 2, each video is divided into non-overlapping 1-second segments to ensure camera motion labels are accurately aligned with actual camera changes. For each segment, we uniformly sample T=8 frames and resize to 336×336, which provides a favorable accuracy–efficiency trade-off for camera motion recognition.

Camera motion label from extrinsic params. We assign camera motion labels by analyzing frame-wise camera extrinsic matrices, following the taxonomy and operational definitions in CameraBench [23]. Concretely, we compute per-segment translation and rotation changes (*e.g.*, yaw/pitch/roll deltas and forward/backward translation) and map them to motion primitives (*e.g.*, `pan-left`, `tilt-down`, `dolly-in`) using thresholded pattern matching in pose dynamics. Compound motions are produced when multiple primitives are detected simultaneously (*e.g.*, `arc-clockwise`, `dolly-in`). With precisely controlled camera motion, our labeling approach provides sufficiently accurate annotation results for each segment. To validate label quality, we conduct a human verification study on 720 randomly selected segments and observe 93% agreement.

We further observe substantial class imbalance. Thus, we apply stratified sampling to construct a balanced subset while preserving diversity across atomic and compound-motion classes, yielding 12,274 segments in total. For *CameraMotionDataset*, a standard train/val/test split is created by 80%/10%/10% folds. Camera motion taxonomy and labeling procedure, and dataset re-balancing details are provided in the supplementary.

Atomic primitives, constraints, and canonicalization. We define 15 atomic primitives and represent each segment

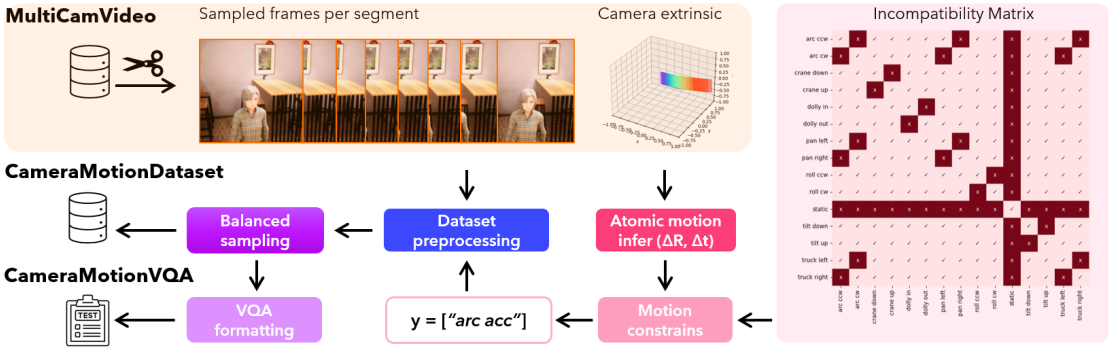

Figure 2. **Flow chart for dataset/benchmark construction.** From MultiCamVideo [2], video clips and camera extrinsics are preprocessed (split, resized, and normalized) and labeled with several motion constraints (*e.g.*, incompatibility). A subset of shot segments is sampled based on motion primitives to balance classes, before storing as CameraMotionDataset and formulated into CameraMotionVQA records.

label as a multi-hot vector. To enforce physically and semantically valid combinations, we define a symmetric incompatibility matrix $\mathbf{M} \in \{0,1\}^{K\times K}$, where $\mathbf{M}_{ij} = 1$ indicates that primitives i and j cannot co-occur (see Fig. 2). Label-set canonicalization limits the maximum number of primitives, uses deterministic ordering and de-duplication, for consistent evaluation and prompt serialization.

CameraMotionVQA benchmark. To benchmark off-the-shelf VideoLLMs under a unified protocol, we construct a multiple-choice VQA benchmark, named CameraMotionVQA. Each question corresponds to a 1-second clip and provides four candidate answers: one ground-truth motion label set and three *distractors*. Distractors are sampled to have similar label complexity as the ground-truth (number of primitives), avoiding trivial answer-length bias. In addition, all candidates are valid under the incompatibility constraints. VQA templates are included in the supplementary.

Tab. 1 positions our dataset and benchmark against existing cinematic benchmarks and video datasets. Based on this comparison, we argue that it is necessary to create a camera motion-focused dataset with trustworthy labels derived deterministically from extrinsic params and sufficient scale to train a robust modern classifier.

3.3. Camera motion cues from 3DFMs

Camera motion is fundamentally geometric, arising from coherent changes in camera pose over time. Although modern VideoLLMs are trained for semantic alignment and temporal coherence, geometric camera-motion cues are not reliably preserved in their intermediate vision representations (Sec. 3.6). To compensate for this missing geometric signal, we use a 3DFM as an external camera motion cue extractor.

Our framework is agnostic to the specific 3DFM: any model that can provide per-frame camera descriptors (*e.g.*, pose-related tokens or camera parameter estimates) can be plugged into our pipeline. VGGT [35] produces camera tokens in a single forward pass and has been shown to encode camera pose and motion dynamics. Given a 1-second segment with T sampled frames, we run VGGT on each frame and obtain a camera token sequence $\{\mathbf{c}_t\}_{t=1}^{T}$, where $\mathbf{c}_t \in \mathbb{R}^{2048}$. These camera tokens serve as the input to our temporal classifier, as shown in Fig. 1.

3.4. Constraint-regularized motion classifier

We train a lightweight classifier to map camera tokens $\mathbf{c}_t$ to constrained multi-label camera-motion predictions. A linear projection $\mathbf{W}_p$ is applied to form an information bottleneck: VGGT tokens may encode rich camera-related factors, and the projection stabilizes training by distilling the subset most relevant to motion prediction. Then, add a sinusoidal positional encoding and prepend a learnable `[CLS]` token. The resulting sequence is processed by an L-layer Transformer encoder, and logits are predicted from the final `[CLS]` token (*i.e.*, $\mathbf{Z}_L[0,:]$) by a linear projection $\mathbf{W}_o$:

$$\mathbf{z}_t = \mathrm{PE}(\mathbf{W}_p\mathbf{c}_t), \quad t = 1,\dots,T, \tag{1}$$

$$\mathbf{Z}_0 = [\mathbf{z}_{\mathrm{cls}}, \mathbf{z}_1, \dots, \mathbf{z}_T], \tag{2}$$

$$\mathbf{Z}_L = \mathrm{TransformerEnc}(\mathbf{Z}_0), \tag{3}$$

$$\mathbf{s} = \mathbf{W}_o\mathbf{Z}_L[0,:] + \mathbf{b}_o, \tag{4}$$

where $\mathbf{s} \in \mathbb{R}^K$ are the output logits and $p_k = \mathrm{sigmoid}(s_k)$ denote per-class probability. Following the constrained multi-label formulation in Sec. 3.1, we optimize a binary cross-entropy loss with two regularizations:

$$\mathcal{L}_{\mathrm{bce}} = -\sum_{k=1}^{K}\Big(y_k \log p_k + (1-y_k)\log(1-p_k)\Big), \tag{5}$$

$$\mathcal{L}_{\mathrm{inc}} = \sum_{1\le i<j\le K} \mathbf{M}_{ij} p_i p_j, \tag{6}$$

$$\mathcal{L}_{\mathrm{card}} = \max\Big(0, 1-\sum_{k=1}^{K} p_k\Big)^2 + \max\Big(0, \sum_{k=1}^{K} p_k - 3\Big)^2 \tag{7}$$

where λ_{inc} and λ_{card} control the strengths of incompatibility and cardinality regularization. In our implementation, we use $\lambda_{\mathrm{inc}} = 1.0$ and $\lambda_{\mathrm{card}} = 1.0$ by default.

Table 1. **Positioning of our datasets against prior benchmarks and datasets.** "Motion-level" indicates the granularity of camera supervision (primitive vs. composition of motions). Our CameraMotionDataset provides constrained primitive-level supervision with explicit synthetic camera parameters, while CameraMotionVQA converts the same 1s within-shot segments into a multiple-choice evaluation protocol. A full comparison with detailed annotation structure and intended usage is provided in the supplementary material.

	Motion-level	Granularity	Temporal unit	QA protocol	Intended use	Camera params
CameraBench [23]	Primitive	Multi-label	Short clips	Caption, Yes/No QA	Benchmark + fine-tune	No
CineTechBench [37]	Composition	Coarse	Frames / short clips	Caption, MCQ	Benchmark	No
VidComposition [34]	Composition	Coarse	Compiled videos	MCQ	Benchmark	No
CineScale2 [30]	-	-	Frames	No	Training dataset	No
CameraMotionDataset	**Primitive**	**Multi-label**	**1s within-shot**	No	**Training dataset**	**Yes**
CameraMotionVQA	**Primitive**	**Multi-label**	**1s within-shot**	**MCQ**	**Benchmark**	No

At inference, we threshold probabilities at τ=0.5 and enforce constraints by removing mutually exclusive primitives within each primitive group and applying canonicalization.

3.5. Motion injection via structured prompting

As a training and fine-tuning free approach, camera motion cues are injected via structured prompting, leaving VideoLLM weights unchanged. For a shot consisting of S one-second segments, each motion label is predicted and serialized as a short string (*e.g.*, `static` or `pan-left and tilt-up`), and these strings are concatenated into a per-shot list: `Per-second camera motion:` $[m_1, m_2, \ldots, m_S]$. We prepend this motion list to the user instruction to provide an explicit temporal scaffold. Empirically, this conditioning might improve the temporal grounding of generated descriptions and reduce camera-motion hallucinations, as the model can align content changes with the provided motion sequence (Fig. 7). Our final prompt template is:

```
Here are [N] consecutive video frames.
They are evenly sampled at a frame rate
of [r] FPS.
Per-second camera motion: [m1,m2,...].
Describe this video using the filmmaker's
language, highlighting lighting, framing,
composition, and especially camera usage
that connects different frames.
For example: "At the beginning, <video
content>; then <camera motion>, <video
content>; ...; finally, <camera motion>,
<video content>."
```

3.6. Probing motion sensitivity via Q-Former

Q-Former [18] style probing provides an effective "readout" mechanism: a small set of learnable query tokens can extract task-relevant information from high-dimensional frozen visual tokens via cross-attention, enabling a parameter-efficient diagnostic of what information is present in the representation. As shown in Fig. 3, we probe on Qwen2.5-VL [3], whose vision encoder is a dynamic-resolution ViT with window attention blocks and periodic full-attention blocks at indices $\{7, 15, 23, 31\}$. Thus, we extract visual tokens from patch embedding and the full-attention blocks. This pipeline consists of: (i) a linear projection that bottlenecks the frozen visual tokens to the query dimension; (ii) 4 learnable query tokens processed by 2 Transformer Encoders; (iii) a 1D temporal convolution over the query tokens to produce a single classification token; and (iv) a linear classifier to predict multi-label logits. The probe is trained with the same loss in Sec. 3.4.

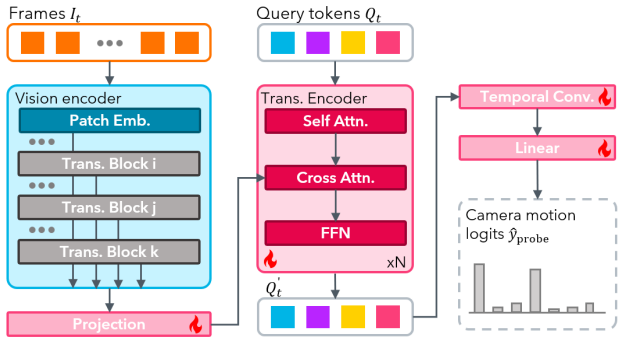

Figure 3. **Probing experiment schematic.** Query tokens Q_t gather camera motion-related information from the projected intermediate visual features of the frozen vision encoder. Camera motion logits are predicted from the temporal convolution output of the transferred vision tokens Q'_t.

3.7. VGGT–Q-Former distillation

VGGT provides high-quality camera tokens but is computationally heavy (1.2B params). As shown in Fig. 4, to reduce inference cost, we propose to distill VGGT's camera perception by a lightweight Q-Former-based student. Specially, the student follows the same *interleaved local- and global-attention* as VGGT for camera reasoning, enabling temporally grounded yet context-aware representations.

Q-Former with interleaved local-frame and global attention. We assign one learnable query token to each (temporally indexed) feature map and have processing blocks that alternate between local (frame) attention and global attention. In local attention, each query attends only to the visual tokens of its assigned frame, encouraging temporally grounded camera cues. In global attention, all queries attend jointly across frames, enabling cross-frame aggregation of camera dynamics. As in VGGT, query tokens Q_t have 256 dimensions, and the outputs from the final local

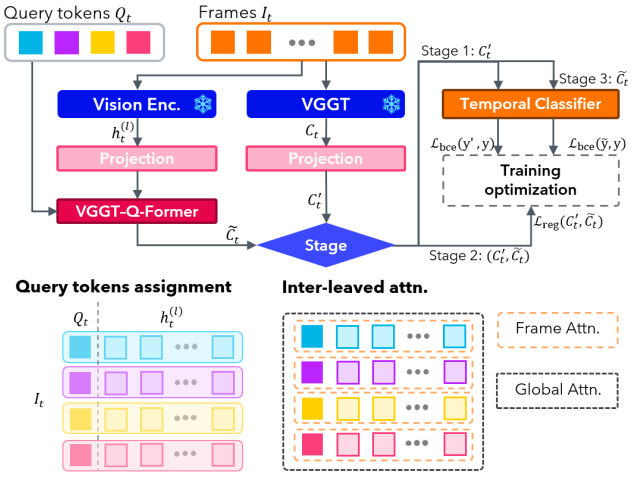

Figure 4. **VGGT–Q-Former schematic.** Camera tokens and visual tokens are both bottlenecked by a projection layer. Query tokens gather camera motion-related information via interleaved local-/global-attention blocks and regress the projected camera tokens to distill the camera perception capability of VGGT. Branch annotated with Stage i indicating different training optimization objectives and flows of tokens.

and global blocks are concatenated to form the final distillation result $\tilde{c}_t$ of dim 512.

Three-stage progressive training. As illustrated in Fig. 4, a progressive training strategy is adopted: (1) train the motion classifier on projected VGGT tokens; (2) train the distiller (VGGT–Q-Former) by regressing the projected VGGT tokens using mean squared error; and (3) jointly fine-tune the VGGT–Q-Former and classifier. The regression loss is

$$\mathcal{L}_{\text{reg}} = \sum_{t=1}^{T} \left| \widetilde{\mathbf{c}}_t - \mathbf{c}'_t \right|_2^2, \tag{8}$$

where $\mathbf{c}'_t$ denotes the projected VGGT token and $\widetilde{\mathbf{c}}_t$ denotes the distilled token.

At inference, VGGT is replaced by VGGT–Q-Former, and the same temporal motion classifier is used to get labels for each segment. In Sec. 4.5, we report the trade-off between efficiency and accuracy of this distillation approach.

4. Experiments

4.1. Experimental setup

All experiments are conducted on a single NVIDIA RTX A6000 GPU. We evaluate camera-motion recognition on **CameraMotionDataset** and **CameraMotionVQA** as detailed in Sec. 3.2. CameraMotionVQA formats each segment as a 4-way multiple-choice question, and models are evaluated using answer accuracy. For evaluating the camera motion multi-label task, we report **instance accuracy** (exact match of all labels), **Macro-F1** across motion primitives, and **Weighted-F1** weighted by label frequency.

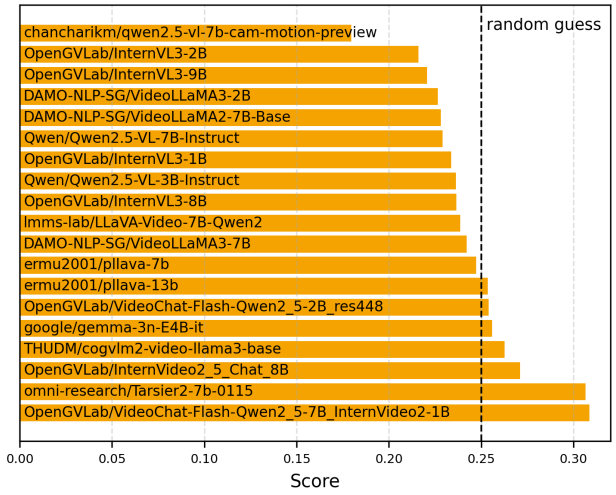

Figure 5. **Off-the-shelf VideoLLM performance on CameraMotionVQA.** Horizontal bars report overall multiple-choice accuracy. All models, labeled by their Hugging Face model names, use an identical frame input and VQA prompt template.

4.2. How well do off-the-shelf VideoLLMs recognize camera motion?

As described in Sec. 3.2, we evaluate diverse off-the-shelf VideoLLMs under the unified CameraMotionVQA protocol. Fig. 5 shows that most models perform near the random-guess rate (25%), revealing a substantial *camera-motion blindness* gap. Notably, a structured fine-tuning baseline from CameraBench [23] performs worse than its off-the-shelf counterpart (Qwen2.5-VL). Qualitatively, models frequently confuse geometrically similar primitives (e.g., `truck` vs. `pan`) and produce inconsistent directional predictions (e.g., `pan-left` vs. `pan-right`) in the presence of salient object motion.

We attribute this gap to the lack of explicit camera-motion supervision in VideoLLM training data, where representations are optimized for semantic alignment and temporal reasoning rather than precise 3D geometric change. Our probing results (Sec. 4.4) support this claim and motivate injecting explicit geometric camera cues.

4.3. Camera motion recognition from 3D foundation-model cues

We run a frozen 3D foundation model (VGGT) on $\{I_t\}_{t=1}^{8}$ and extract per-frame camera cues (camera tokens $C_t \in \mathbb{R}^{2048}$). A lightweight temporal classifier (Sec. 3.4) consumes the cue sequence and predicts constrained multi-label motions. Unless otherwise stated, VGGT is frozen, and we train a shallow Transformer classifier with 4 encoder blocks and 8 attention heads. Each camera token is projected to $C'_t \in \mathbb{R}^{512}$; a learnable `[CLS]` token is prepended; the final `[CLS]` embedding is mapped to K logits. In the supplementary, we ablate the number of encoder blocks, attention heads, and hidden size, and find that the current set-

Table 2. **Multi-label camera-motion recognition results on the test split of CameraMotionDataset.** *Inst. Acc.*: exact multi-label matching accuracy; *Macro-F1*: class-averaged F1 score; *Weighted-F1*: sample frequency-weighted F1 score.

Method	Inst. Acc.↑	Macro-F1↑	Weighted-F1↑
VGGT w. constraints	**0.738**	**0.87**	**0.92**
VGGT w/o. constraints	0.572	0.79	0.84
VGGT–Q-Former	0.638	0.83	0.87
Q-Former probing	0.450	0.69	0.74

ting achieves a favorable accuracy–compute trade-off.

Tab. 2 shows that VGGT-derived cues with a lightweight classifier substantially outperform off-the-shelf VideoLLM baselines (Fig. 5). Removing constraint enforcement (*VGGT w/o. constraints*) reduces instance-level accuracy, indicating that modeling axis-wise mutual exclusion improves performance even with strong cues. Per-label prediction results (in the supplementary) show that errors concentrate on rare or ambiguous primitives, consistent with the long-tail distribution. We also observe unreliable predictions for the `static` class, which is likely out-of-distribution for VGGT, whose reconstruction prior assumes camera motion. Static segments may require dedicated handling beyond 3DFM priors.

Despite its effectiveness, VGGT cue extraction is costly: the 1.2B-parameter model performs multi-view 3D reasoning over all frames, dominating latency and memory.

4.4. Probing motion sensitivity in a vision encoder

As described in Sec. 3.6, we freeze the Qwen2.5-VL vision encoder and train a small Q-Former-style probe on intermediate features to predict motion labels. This pathway is used *solely* to diagnose representation bottlenecks. The probe comprises 2 Transformer blocks with 8 heads and 4 learnable query tokens. Query tokens and projected features are 768-dimensional.

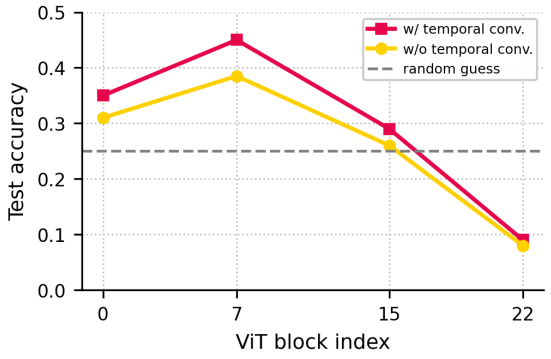

Figure 6. **Results of probing experiment.** Intermediate features from the frozen vision tower at different Transformer block indices are probed by query tokens. Performance peaks at a shallow intermediate block and degrades in later layers, suggesting that camera-motion cues are not reliably preserved.

Table 3. **Efficiency comparison of camera-motion recognition pipelines.** We compare trainable parameter count, peak GPU memory, and inference time throughput for: the full VGGT-based classifier; the distilled VGGT–Q-Former pipeline; and the Q-Former probing experiment. All measurements are conducted on an RTX A6000 with batch size 16 and input resolution 336×336.

Pipeline	Params (M)	Peak mem. (MBs)	Throughput (samples/s)
VGGT classifier	9.47	23649.11	4.39
VGGT–Q-Former	9.15	9202.63	23.36
Q-Former probing	15.18	9232.23	25.12

As shown in Fig. 6, performance peaks at the first full-attention block and declines in later layers. This suggests that camera-motion information is weakly encoded and becomes less accessible as features are optimized for semantic alignment with the language model. Replacing temporal convolution with average pooling further reduces accuracy, indicating that explicit temporal modeling is necessary to derive camera motion from token sequences.

4.5. Distilling 3D Camera Cues for efficiency

VGGT-based cue extraction is computationally expensive at inference: the 1.2B-parameter backbone performs multi-view 3D reasoning over all frames. To reduce cost, we distill teacher camera tokens into a compact student Q-Former (VGGT–Q-Former) that reuses frozen VideoLLM vision features (Sec. 3.7). We use hidden features h_t^l from the 7th block of Qwen2.5-VL, where probing shows higher camera-motion recoverability (Fig. 6). The student attends to visual tokens and predicts embeddings that regress projected teacher tokens. We use 4 learnable queries (one per frame group) and follow VGGT's interleaved local/global attention (Fig. 4): 2 local blocks (query–frame) and 2 global blocks (cross-frame). Training proceeds in three stages: (1) train the motion classifier for 50 epochs; (2) train query tokens for 100 epochs to regress teacher tokens; (3) jointly fine-tune the student and classifier for 30 epochs. We use Adam with a learning rate of 1e-4.

Tab. 3 reports inference overhead. Although instance accuracy drops by 8.13%, distillation substantially reduces end-to-end cost. The distilled model (8.72M *vs.* 1.2B) achieves 5.3× throughput at 39% peak memory, offering a favorable accuracy–latency trade-off. Overall, teacher cues maximize accuracy but are costly, whereas distilled cues improve efficiency at some loss in accuracy.

4.6. Qualitative results for structured prompting

Our goal is to enable VideoLLMs to produce camera-aware, filmmaker-style descriptions that capture *what* happens and *how* the shot is filmed and evolves over time. Following Sec. 3.5, we prepend a per-second motion header to the prompt without modifying model weights. Fig. 7 shows

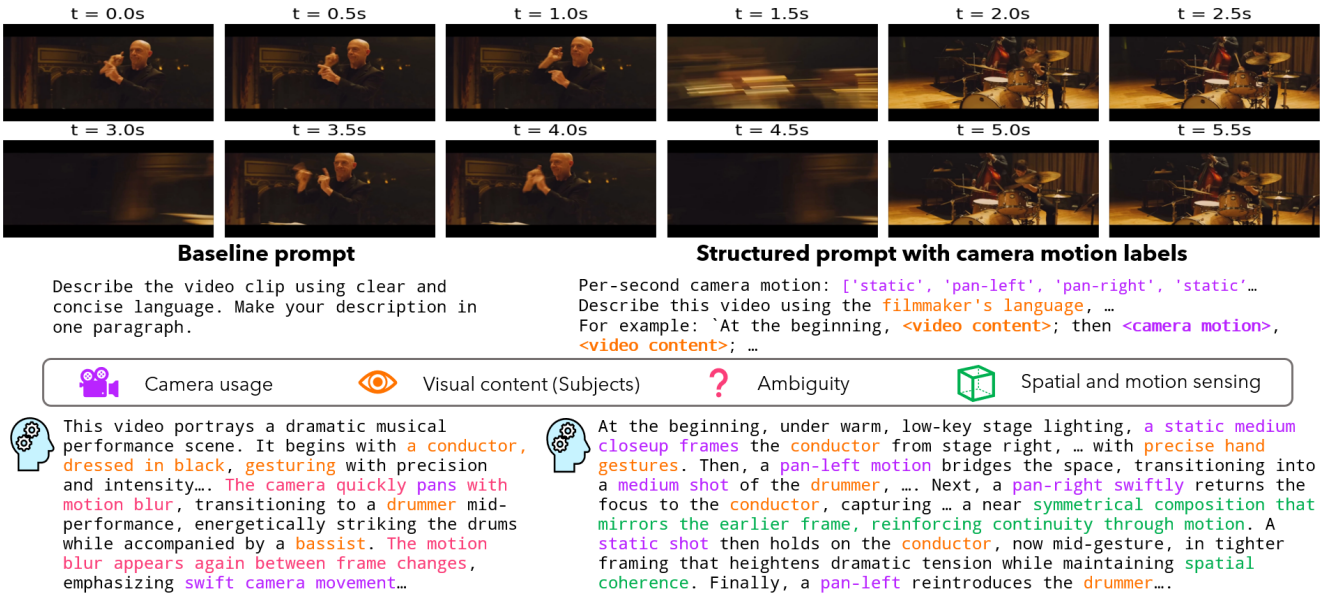

Figure 7. **Structured motion header enables camera-aware and temporally grounded descriptions.** Top: a 5-second clip sampled at 2 fps. Bottom: outputs from the same VideoLLM with different prompts. The baseline description correctly identifies subjects but produces ambiguous motion statements, without specifying direction or temporal structure. With structured motion cues, the model generates explicit camera usage, maintains subject focus, and introduces spatial-temporal reasoning such as continuity and coherence.

a representative example. Under the baseline prompt, the model recognizes the conductor and drummer, but describes motion vaguely (*e.g.*, "camera quickly pans with motion blur") without direction or temporal structure. The description mixes possible cuts and camera movement, leading to ambiguity in how frames are connected.

In contrast, with the provided explicit motion cues, the same VideoLLM produces a temporally structured narrative. It explicitly grounds motion direction (`pan-left` and `pan-right`), describes framing (`static medium close-up`), and adds spatial reasoning (*e.g.*, "mirrors the earlier frame", "spatial coherence"). This suggests motion headers improve motion correctness and encourage geometry-aware, temporally consistent reasoning.

Beyond motion correctness, the header biases the model toward spatial and motion reasoning. The structured cues promote continuity, geometric consistency, and frame-to-frame transitions, reducing generic descriptions. This suggests that camera motion labels act as geometric priors that steer large VideoLLMs toward more temporally grounded and cinematographically aware reasoning. Full outputs and other examples are provided in the supplementary.

5. Conclusion and Discussion

Camera motion is a fundamental geometric signal that shapes video perception, yet it is rarely modeled explicitly in current VideoLLMs. In this work, this gap is addressed through a loop of *benchmarking*, *diagnosis*, and *injection*: we introduce a shot-consistent 1-second dataset and a VQA benchmark that formulates fine-grained motion primitives as a constrained multi-label recognition task; we show via probing that camera motion cues are only weakly recoverable from frozen VideoLLM vision features; and we propose a lightweight pipeline that extracts cues from a frozen 3D foundation model (VGGT). The pipeline predicts constrained motion primitives and injects them via structured prompting *without updating VLM weights*.

Camera cues support applications where *how* a scene is filmed matters in addition to *what*, including descriptive video services (DVS) [40], media recommendation and retrieval by filmmaking metadata [30], camera-aware video attribute or authorship analysis [20], and plagiarism detection. More broadly, our results suggest camera awareness remains a structured supervision problem and there is representation gap in current VideoLLMs. In this work, we propose a plug-and-play solution, in which synthetic camera-controlled data provides scalable supervision, 3DFMs inject geometric priors, and distillation offers a favorable cost–performance trade-off. Limitations include the synthetic-to-real gap, focus on camera *extrinsic* rather than *intrinsic* changes (*e.g.*, zoom), and only a single 3DFM backbone is explored. A detailed discussion is provided in the supplementary, including a data augmentation strategy, and a LLM-as-judge evaluation protocol. Future work will extend to broader camera configurations and 3DFMs, and systematically study how camera cues transfer to downstream video understanding and generation tasks.

Acknowledgments

This work was conducted during the first author's internship at Dolby Laboratories Inc. Special thanks to Xinran Wang [37] for sharing the full CineTechBench dataset.

References

[1] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22875–22889, 2025. 2

[2] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 2, 3, 4

[3] 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, 5

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 2

[5] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 3

[6] Andrea Porfiri Dal Cin, Georgi Dikov, Jihong Ju, and Mohsen Ghafoorian. Anymap: Learning a general camera model for structure-from-motion with unknown distortion in dynamic scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16674–16684, 2025. 3

[7] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2694–2703, 2023. 2

[8] Henghui Ding, Chang Liu, Shuting He, Kaining Ying, Xudong Jiang, Chen Change Loy, and Yu-Gang Jiang. Mevis: A multi-modal dataset for referring motion expression video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[9] Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Xudong Jiang, Yu-Gang Jiang, Philip HS Torr, and Song Bai. Mosev2: A more challenging dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2508.05630*, 2025. 2

[10] Sven Elflein, Qunjie Zhou, and Laura Leal-Taixé. Light3r-sfm: Towards feed-forward structure-from-motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16774–16784, 2025. 3

[11] Epic Games. Unreal engine 5. 3

[12] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025. 3

[13] Minghao Guo, Meng Cao, Jiachen Tao, Rongtao Xu, Yan Yan, Xiaodan Liang, Ivan Laptev, and Xiaojun Chang. Glad: Geometric latent distillation for vision-language-action models. *arXiv preprint arXiv:2512.09619*, 2025. 3

[14] 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. 2

[15] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13416–13426, 2025. 2

[16] Daniel Helm, Florian Kleber, and Martin Kampel. Histshot: A shot type dataset based on historical documentation during wwii. In *ICPRAM*, pages 636–643, 2022. 2

[17] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 3

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. 2, 3, 5

[19] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 2

[20] Yuzhi Li, Tianfeng Lu, and Feng Tian. A lightweight weak semantic framework for cinematographic shot classification. *Scientific Reports*, 13(1):16089, 2023. 2, 8

[21] Yifan Li, Zhixin Lai, Wentao Bao, Zhen Tan, Anh Dao, Kewei Sui, Jiayi Shen, Dong Liu, Huan Liu, and Yu Kong. Visual large language models for generalized and specialized applications. *arXiv preprint arXiv:2501.02765*, 2025. 3

[22] Ziye Li, Hao Luo, Xincheng Shuai, and Henghui Ding. Anyi2v: Animating any conditional image with motion control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17302–17311, 2025. 2

[23] Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, et al. Towards understanding camera motions in any video. *arXiv preprint arXiv:2504.15376*, 2025. 1, 2, 3, 5, 6

[24] Hongbo Liu, Jingwen He, Yi Jin, Dian Zheng, Yuhao Dong, Fan Zhang, Ziqi Huang, Yinan He, Yangguang Li, Weichao Chen, et al. Shotbench: Expert-level cinematic understanding in vision-language models. *arXiv preprint arXiv:2506.21356*, 2025. 2

[25] Yiwen Liu, Jianguo Jiang, Min Yu, Boquan Li, Myung Hwan Na, and Gang Li. Posemaster: Editing your pose in a video with a one-shot framework. In *International Conference on Advanced Data Mining and Applications*, pages 239–253. Springer, 2025. 2

[26] Yawen Luo, Xiaoyu Shi, Jianhong Bai, Menghan Xia, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Camclonemaster: Enabling reference-based camera control for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–10, 2025. 2

[27] Zhenyuan Qin, Xincheng Shuai, and Henghui Ding. Scenedesigner: Controllable multi-object image generation with 9-dof pose manipulation. *arXiv preprint arXiv:2511.16666*, 2025. 2

[28] 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 *European Conference on Computer Vision*, pages 17–34. Springer, 2020. 2

[29] Mattia Savardi, András Bálint Kovács, Alberto Signoroni, and Sergio Benini. Cinescale: A dataset of cinematic shot scale in movies. *Data in Brief*, 36:107002, 2021.

[30] Mattia Savardi, András Bálint Kovács, Alberto Signoroni, and Sergio Benini. Cinescale2: a dataset of cinematic camera features in movies. *Data in Brief*, 51:109627, 2023. 2, 5, 8

[31] Zhijian Shu, Cheng Lin, Tao Xie, Wei Yin, Ben Li, Zhiyuan Pu, Weize Li, Yao Yao, Xun Cao, Xiaoyang Guo, et al. Litevggt: Boosting vanilla vggt via geometry-aware cached token merging. *arXiv preprint arXiv:2512.04939*, 2025. 3

[32] Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*, 2024. 2

[33] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *IEEE TCSVT*, 2025. 1

[34] 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? In *CVPR*, pages 8490–8500, 2025. 2, 5

[35] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025. 1, 3, 4

[36] Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025. 2

[37] Xinran Wang, Songyu Xu, Xiangxuan Shan, Yuxuan Zhang, Muxi Diao, Xueyan Duan, Yanhua Huang, Kongming Liang, and Zhanyu Ma. Cinetechbench: A benchmark for cinematographic technique understanding and generation. *arXiv preprint arXiv:2505.15145*, 2025. 2, 5, 9

[38] 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 *ECCV*, pages 396–416. Springer, 2024. 3

[39] 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*, pages 1–11, 2024. 2

[40] Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Eshika Khandelwal, Gül Varol, Weidi Xie, and Andrew Zisserman. Shot-by-shot: Film-grammar-aware training-free audio description generation. *arXiv preprint arXiv:2504.01020*, 2025. 2, 3, 8

[41] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 2

[42] 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. 2

[43] Xiaoda Yang, Jiayang Xu, Kaixuan Luan, Xinyu Zhan, Hongshun Qiu, Shijun Shi, Hao Li, Shuai Yang, Li Zhang, Checheng Yu, et al. Omnicam: Unified multimodal video generation via camera control. *arXiv preprint arXiv:2504.02312*, 2025. 2

[44] Kaining Ying, Hengrui Hu, and Henghui Ding. Move: Motion-guided few-shot video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11632–11642, 2025. 2

[45] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 100–111, 2025. 2

[46] 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. 3

[47] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 3

[48] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 543–553, 2023. 3

[49] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 3

[50] Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *arXiv preprint arXiv:2505.24625*, 2025. 3