CrossView: Can Vision-Language Models Reason Across Cameras?

Sahil Shah*[1], S P Sharan*[1], Harsh Goel*[1], Manvik Pasula[2], Adithya Hebbalae[1], Minkyu Choi[1], and Sandeep P. Chinchali[1]

[1] The University of Texas at Austin, Austin TX, USA
[2] Independent Researcher, USA

Abstract. Video understanding benchmarks have long centered on single-camera settings, where modern multi-modal language models achieve strong performance across image and video tasks. Yet, the real world runs on multi-camera networks: autonomous vehicles, security systems, and robots all gather data across many simultaneous views. We argue that this is not simply "more" of the single-camera problem; it is fundamentally different. Multi-camera reasoning requires handling context that scales with the number of views, resolving occlusions visible from only a subset of cameras, judging which views matter, and integrating evidence across perspectives that may overlap or diverge. Current models struggle with exactly these challenges, yet no benchmark systematically targets them. We introduce `CrossView`, a multi-camera video question-answering benchmark spanning autonomous driving, security surveillance, egocentric/exocentric video, and robotics. Evaluation of proprietary models, such as GPT-5.2, and open-source models, like Qwen3-VL, reveals consistently low accuracy, with open-source models trailing by a wide margin. Performance scales strongly with a model's ability to jointly process multiple viewpoints, positioning `CrossView` as a rigorous benchmark for multi-camera video. We open-source our code and dataset at `https://utaustin-swarmlab.github.io/CrossView`.

Keywords: Multi-Camera Reasoning · Multi-Modal Language Models· Video Understanding Benchmark

1 Introduction

The "single-camera assumption" has long dominated the landscape of computer vision. From image classification to recent Large Vision-Language Models (LVLMs), benchmarks have primarily evaluated the ability to reason over a single camera. Most state-of-the-art (SOTA) models now excel at this, reaching near-human performance on visual benchmarks such as VQAv2 [15], and video benchmarks like Video-MME [13] and LongVideoBench [53].

However, real-world systems rarely rely on a single perspective. In autonomous driving (AD), robotics, and wide-area surveillance, perception is inherently multi-camera: a self-driving car fuses surround-view cameras to navigate intersections,

* Equal contribution.

Fig. 1. `CrossView` data example illustrating multi-view reasoning. Multiple synchronized camera streams observe the *same* scene from *different* viewpoints (egocentric and exocentric). Answering the question requires integrating information across views (e.g., to identify the event that occurs after the saucepan is placed on the stovetop). `CrossView` evaluates such cross-view reasoning tasks across diverse domains, including robotics, surveillance, autonomous driving, and ego–exo interaction.

a robotic arm coordinates wrist-mounted and overhead views to manipulate objects, and security networks track subjects across disjoint fields of view. Despite this, current multi-modal language models are still evaluated almost exclusively on single-camera inputs.

The gap between single- and multi-camera video understanding is not merely one of data volume. Multi-camera reasoning introduces two fundamental *peculiarities* that current models find uniquely challenging: ❶ Context Scaling, where processing N simultaneous streams expands the required context window by an order of magnitude and often exceeds the effective window of long-context models, and ❷ Cross-View Spatial Reasoning, where models must go beyond single-camera temporal and visual reasoning to perform spatio-temporal stitching (*i.e.*, synthesizing discrete viewpoints into a coherent scene). This entails distinguishing overlapping *vs.* non-overlapping fields of view, selecting the most informative camera for a specific sub-task (camera identification), aggregating evidence across cameras (*e.g.*, counting unique objects), and reasoning about events across disjoint streams.

Existing datasets supply raw multi-camera data, but they primarily only address narrow facets of cross-view reasoning. nuScenes-QA [41,4] targets only perception, motivating methods that merge views into a unified Bird's-Eye-View (BEV) representation [11], while Ego-Exo4D [17] is restricted to activity understanding. There is a distinct lack of benchmarks that force a model to reason spatio-temporally across raw, simultaneous, and heterogeneous camera feeds.

We address this with `CrossView`, a comprehensive multi-camera video question answering (VQA) benchmark designed to stress-test the cross-view reasoning capabilities of modern VLMs. `CrossView` comprises 6,000 questions across four domains (Autonomous Driving, Surveillance, Egocentric-Exocentric interaction, and Robotics) with tasks such as camera identification, cross-view object counting, and spatio-temporal event ordering that cannot be solved from any

single camera in isolation, a property we verify empirically by restricting inputs to a single view (Table 5).

Our contributions can be summarized as follows:

1. Novel Multi-Camera Benchmark: We introduce `CrossView`, the first VQA dataset specifically curated to evaluate joint reasoning over 2 to 8 simultaneous camera streams across four real-world domains.
2. Targeted Multi-View Reasoning Tasks: We design natural-language questions with five distinct categories (Temporal, Event Ordering, Spatial, Counting, and Summarization) alongside a "Camera-ID" task that requires models to understand spatially across multiple viewpoints and temporally across the video to answer questions.
3. Comprehensive Model Evaluation: We provide an extensive study of both proprietary (GPT-5.2) and open-source (Qwen, InternVL, and Gemma families) models, revealing a significant "multi-camera gap" where performance degrades in multi-camera settings.
4. Multi-view Visual Aggregation: We investigate how different architectural choices of stacking frames from multiple videos as visual input impact a model's ability to maintain spatial-temporal understanding across views.

2 Related Works

2.1 Visual Question Answering Benchmarks and Models

Image and Video QA. Visual question answering began at the image level: VQAv2 [15] paired open-ended questions with natural images and established accuracy as the standard metric. GQA [24] added compositional scene-graph-grounded questions for spatial and relational reasoning, and OK-VQA [37] and A-OKVQA [44] emphasized questions requiring external knowledge. Video QA followed with early benchmarks such as TGIF-QA [27] and ActivityNet-QA [57] targeting spatio-temporal reasoning over short video clips. Recent long-video methods [56,45,50] and benchmarks have also emerged: LongVideoBench [53] assesses interleaved video-language understanding, Video-MME [13] spans short to hour-long videos, MLVU [61] focuses on multi-task long video understanding, EgoSchema [36] examines egocentric temporal reasoning, and MVBench [30] introduces evaluation for temporal perception. Despite this progress, all primarily still rely on a *single* camera stream, leaving cross-view evidence aggregation, view selection, and context scaling with the number of cameras entirely unexplored.

Driving-domain QA. Driving benchmarks work with multi-camera data but still lack true multi-camera reasoning. NuScenes-QA [41] builds QA pairs on nuScenes from 3D scene graphs but stays tied to spatial perception, while NuPlanQA [39] and OmniDrive [48] add planning-oriented temporal reasoning that remains egocentric. The TLV dataset [9] further focuses on search problems in autonomous driving scenarios. Each moves evaluation closer to real-world driving, however

they still abstract away core multi-camera challenges such as allocentric spatio-temporal reasoning. `CrossView` directly tests this by forcing integration across concurrent video streams.

Vision-language models. VLMs are advancing in step with these benchmarks. Proprietary models like GPT-4o [25] and Gemini [47] support long visual contexts and excel on single-camera inputs. Open-source models such as Qwen-VL [2,1] and InternVL [7,6] handle variable-length multi-image and video inputs, while Gemma [14] offers lightweight options. All, however, are designed and evaluated for single-view inputs, leaving multi-feed reasoning unevaluated. `CrossView` addresses this gap.

2.2 Multi-camera Systems and Multi-View Understanding

Multi-camera data is ubiquitous in deployed systems, and a rich body of work addresses perception and understanding over it. We survey the key areas below.

Multi-camera perception and tracking. Multi-target multi-camera (MTMC) tracking is a central problem, with graph-based methods [43,52] jointly optimizing detection and association across views, and benchmarks such as CityFlow [46] and WILDTRACK [5] establishing cross-camera video understanding performance in urban and surveillance settings. Approaches such as MVDet [23] improve cross-view understanding on these datasets via convolutional neural networks. In autonomous driving, nuScenes [4] provides synchronized six-camera feeds used for 3D detection [49,33], BEV segmentation [32,40], and tracking [8]. Cross-camera person re-identification reasons across views to match identities, driven by large-scale datasets such as Market-1501 [60], MSMT17 [51], CUHK03 [31], and MARS [59] and diverse methods [21,34]. The MEVA dataset [10], used in `CrossView`, was originally designed for activity detection in large surveillance networks. These methods perform genuine cross-view reasoning, but they target low-level perception outputs, such as bounding boxes, trajectories, and identity embeddings, rather than the language-grounded semantic understanding, event ordering, and spatial summarization `CrossView` requires.

Multi-view 3D understanding. A related line of work uses multiple viewpoints for 3D scene understanding. Multi-view stereo [55,18] reconstructs dense geometry from calibrated images, bird's-eye-view methods [32,40] project surround-view features into a unified spatial representation for driving. Neural radiance fields [38], and 3D Gaussian splatting [29] synthesize novel views from multi-view inputs, with recent work [22,62] coupling 3D representations to language models for spatial QA. All, however, depend on a reconstructed or unified 3D representation rather than reasoning directly over raw multi-camera inputs.

Egocentric-exocentric understanding. Ego-Exo4D [17] provides synchronized egocentric and exocentric views of skilled activities with keystep, proficiency, and

Table 1. Comparison of `CrossView` with existing VQA benchmarks. **#Cam**: number of simultaneous camera views. **#Q**: number of questions. **Cross-View**: whether questions require integrating evidence across camera perspectives. **Cam-ID**: whether the benchmark includes camera identification/selection questions. **Multi-domain**: whether the benchmark spans multiple domains. **Questions**: Temp (temporal), Evt (event ordering), Spat (spatial), Count (counting), Summ (summarization).

Benchmark	#Cam	#Q	Duration	Domains	Cross-View	Cam-ID	Multi-domain	Temp	Evt	Spat	Count	Summ
Image & Short-Video QA												
VQAv2 [15]	1	1.4M	Image	Natural	✗	✗	✗	✗	✗	✓	✓	✗
GQA [24]	1	22M	Image	Natural	✗	✗	✗	✗	✗	✓	✗	✗
GazeVQA [26]	3	25K	∼2.5 min	Gaze (ego-exo)	✓	✗	✗	✗	✗	✓	✗	✗
TGIF-QA [27]	1	165K	<10 s	GIFs	✗	✗	✗	✓	✗	✗	✓	✗
ActivityNet-QA [57]	1	58K	∼3 min	Activities	✗	✗	✗	✓	✗	✓	✓	✗
General Video QA												
EgoSchema [36]	1	5K	∼3 min	Egocentric	✗	✗	✗	✓	✓	✗	✗	✓
MVBench [30]	1	4K	∼16 s	Multi-source	✗	✗	✓	✓	✓	✓	✓	✗
LongVideoBench [53]	1	6.7K	∼8 min	Multi-source	✗	✗	✓	✓	✓	✓	✗	✗
VideoMME [13]	1	2.7K	∼17 min	Multi-source	✗	✗	✓	✓	✓	✓	✓	✗
MLVU [61]	1	3.1K	∼15.5 min	Multi-source	✗	✗	✓	✓	✓	✗	✓	✓
Driving-Domain QA												
NuScenes-QA [41]	6†	460K	Seconds	Driving	✗	✗	✗	✓	✗	✓	✓	✗
NuPlanQA [39]	8	1M	Seconds	Driving	✗	✗	✗	✓	✗	✓	✗	✗
OmniDrive [48]	6†	300K	Seconds	Driving	✗	✗	✗	✓	✗	✓	✗	✗
Ego-Exo Benchmarks												
EgoExoBench [20]	2–5	7.3K	Minutes	Ego-exo	Partial	✗	✗	✓	✓	✓	✗	✗
LangView [35]	2–5	–	Minutes	Ego-exo	Partial	✓	✗	✗	✗	✗	✗	✗
PDB-Eval [54]	2	44.9K	Minutes	Ego-exo (driving)	Partial	✗	✗	✓	✗	✗	✗	✗
`CrossView` (Ours)	**2–8**	**6K**	**Sec–min**	**AD, Surv., Ego, Robot.**	✓	✓	✓	✓	✓	✓	✓	✓

†Data sourced from 6 cameras, but questions are grounded to individual views and do not require cross-view reasoning.

language annotations, building on Ego4D [16], which established large-scale egocentric benchmarks for episodic memory, forecasting, and social interaction. Recent work increasingly links the two perspectives: He *et al.* [19] survey ego-exo understanding, Exo2Ego [58] examines knowledge transfer across views, Ego2ExoVLM [42] tests the recognition of egocentric content from exocentric views, and EgoExoBench [20] evaluates VLMs on ego-exo relationships. PDB-Eval [54] measures a driver's behavior in an ego-exo setting, and LangView [35] studies viewpoint selection on Ego-Exo4D [17] and LEMMA [28], directly paralleling the camera-identification task in `CrossView`. These existing benchmarks focus on viewpoint correspondence or selection in isolation; none require *joint* integration of evidence across all cameras for counting, spatial reasoning, temporal ordering, and summarization questions—capabilities `CrossView` targets.

Robotic multi-camera systems. Multi-camera setups are standard in robotic manipulation and navigation. Datasets such as AgiBot [3] provide multi-view recordings from wrist-mounted and external cameras. Prior robotic video benchmarks [63,12] target action prediction and planning from single or paired views rather than language-grounded reasoning across simultaneous feeds. `CrossView` draws robotic scenarios from AgiBot to test whether VLMs can reason about manipulation when evidence is distributed across cameras.

2.3 Benchmark Comparison

Table 1 positions `CrossView` against existing VQA benchmarks. While prior work thoroughly covers single-camera settings across varied durations and ques-

tion types, none systematically evaluate multi-camera reasoning. `CrossView` is the first to combine simultaneous camera views, diverse real-world domains, and question categories designed specifically to probe cross-view spatio-temporal reasoning.

3 Dataset Construction

We first construct a dataset via a two-stage pipeline: we build a Spatio-temporal Scene Graph (STSG) from existing robotics, surveillance, and autonomous driving datasets, then apply a programmatic question-generation engine. The STSG is a structured representation consolidating semantic and event-centric metadata into a unified representation which the question-generation engine queries to synthesize complex questions without risk of model hallucination. We first programmatically identify grounding events/objects, target events, correct ground-truth answers, and plausible distractors from the scene graphs, then pass this metadata to GPT-5.2 to phrase natural-language questions, ensuring semantic correctness without human verification.

3.1 Spatio-temporal Scene Graph (STSG) Construction

The STSG construction in Algorithm 1 begins by consolidating metadata from Ego-Exo4D, nuScenes, AgiBot, and MEVA. We use the spatial coordinates of annotated objects across these datasets [Line 3]: for nuScenes we incorporate LiDAR point clouds for 3D localization [Line 5], while for the others we rely on the provided object-centric 3D or bounding-box annotations [Line 7]. To add semantic information (*i.e.*, object activities and descriptions), we use InternVL-3.5 38B to annotate object-centric activities for datasets lacking native labels [Line 9]: for nuScenes and AgiBot we first associate each object with its bounding box before querying InternVL for corresponding activity descriptions. In contrast, Ego-Exo4D and MEVA supply action and event labels natively. We additionally parse per-frame annotated objects and activities with GPT-5.2 to obtain a scene-level caption per timestamp.

Once objects are localized and their activities identified, we compute pairwise spatial relationships between all entities in the scene across camera views. These relationships include directional predicates such as *behind*, *near*, *left*, and *right*, plus orientation deltas and relative distances [Lines 13–14]. We then form event intervals $[t_{start}, t_{end}]$ by grouping consecutive timestamps over which a given object maintains the same activity [Line 15]; we denote the start and end timestamp of an event E's interval by $E.\texttt{start}$ and $E.\texttt{end}$ respectively. These intervals are derived from the object activity labels obtained in Line 9.

Finally, for each timestamp t, we generate a spatio-temporal graph snapshot $(\mathcal{G}_t = \mathcal{N}_t, \mathcal{E}_t)$, where $\mathcal{N}_t$ denotes object nodes containing activity labels, descriptions, and associated event intervals, and $\mathcal{E}_t$ denotes edges encoding spatial relationships between objects. These graph snapshots are appended chronologically to yield the final STSG $\mathcal{G}$ [Line 16].

Algorithm 1 Spatio-temporal Scene Graph (STSG) Construction

Require: Multi-modal datasets $\mathcal{D} \in$ nuScenes, Ego-Exo4D, AgiBot, MEVA, video timestamps $\mathcal{T}$
Ensure: Spatio-temporal Scene Graph $\mathcal{G}$
Phase 1: Multi-modal perception and metadata consolidation
for each scene in $\mathcal{D}$ **do**
  Extract object trajectories and spatial coordinates
  if dataset is nuScenes **then**
    Refine 3D localization using LiDAR point cloud data
  else
    Utilize provided 3D or bounding-box annotations
  end if
  Generate object-centric activity labels and descriptions using VLM-based annotation for nuScenes/AgiBot or native metadata for Ego-Exo4D/MEVA
  Generate timestamp-level scene captions using GPT-5.2 from annotated objects and activities
  Compute event intervals $[t_{start}, t_{end}]$ by grouping consecutive timestamps with consistent object activities
end for
Phase 2: Spatial and temporal graph construction
Initialize $\mathcal{G} = \emptyset$
for each timestamp $t \in \mathcal{T}$ **do**
  Compute pairwise spatial predicates R_t between object pairs, e.g., *behind*, *near*, *left*, and *right*
  Calculate orientation deltas and relative distances between all object pairs
  Obtain scene graph snapshot $\mathcal{G}_t = \mathcal{N}_t, \mathcal{E}_t$, where $\mathcal{N}_t$ contains object nodes with activity labels, descriptions, and event intervals, and $\mathcal{E}_t$ contains spatial relationship edges between objects
  $\mathcal{G} = \mathcal{G} \cup \mathcal{G}_t$
end for
return $\mathcal{G}$

3.2 Programmatic Question Construction

Our question construction follows a Grounding-Target architecture that programmatically queries the STSG. For each question, we identify grounding event(s) or object(s) that serve as anchors, select the corresponding target event or object to be queried, and sample negative candidates that act as distractors. This structured information, consisting of the grounding anchors, target answer, and distractors, is passed to GPT-5.2 which generates a coherent natural-language question. We provide the QA generation prompts in the Appendix.

Temporal Reasoning. We use the temporally consolidated events in the STSG to construct temporal and spatio-temporal questions. For temporal questions, we sample a grounding event (E_g) and a target event (E_t), and assign one of four temporal relations: *Before*, *After*, *During*, or *In-between*. *Before* holds when (E_t.`end` $< E_g$.`start`), while *After* holds when (E_t.`start` $> E_g$.`end`). *Dur-*

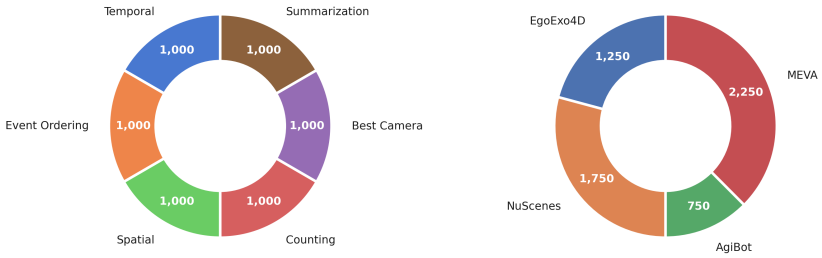

Fig. 2. Dataset question distribution. **Left:** Questions are distributed among six categories, each contributing 1,000 questions. **Right:** Distributions across the four source datasets. MEVA and nuScenes contribute the majority of the questions (2,250 and 1,750 respectively), while Ego-Exo4D and AgiBot provide 1,250 and 750 questions.

ing holds when the temporal overlap between the two events exceeds 50% of the shorter event's duration. *In-between* is defined using two non-overlapping grounding events, (E_{g1}) and (E_{g2}), where the target event lies entirely within the temporal gap between them. For example, this relation can produce questions such as: "What happens in between a blue car parked by the street and the pedestrian with a blue shirt walking across the crosswalk?"

We extend this to spatio-temporal cross-referencing via two categories. In Category I (Spatial-to-Temporal), the grounding anchor is a spatial state, defined by a directional predicate between two objects at a given frame, and the target is an adjacent or intervening temporal activity. For example, a question could be: "When the pedestrian wearing the white shirt approaches and is to the left of the black parked car, what happens immediately after?" In Category II (Temporal-to-Spatial), the grounding anchor is a temporal boundary, and the target is a spatial snapshot, an intervening relation between two events, or a comparative change in object relations. For example, this category can generate questions such as: "What happens in between a woman cycling on the white bike taking a left turn and the yellow car stopping at the intersection?" with the answer: "A delivery truck stops behind the yellow car."

Finally, we convert the sampled tuples into natural-language questions with GPT-5.2. The prompt enforces grounded descriptions by referencing object appearance and activity rather than generic class labels. To increase difficulty, we sample three distractor types: spatial distractors (wrong direction at the correct time), temporal distractors (correct event at the wrong time), and existential distractors (objects absent from the scene). Each sample is stored as a JSON object containing the question, distractors, and reasoning.

Spatial Question. Spatial questions test the model's understanding of 3D environments from multiple viewpoints. For egocentric tasks, the engine extracts directional predicates of objects relative to the ego-vehicle. For frame-of-reference

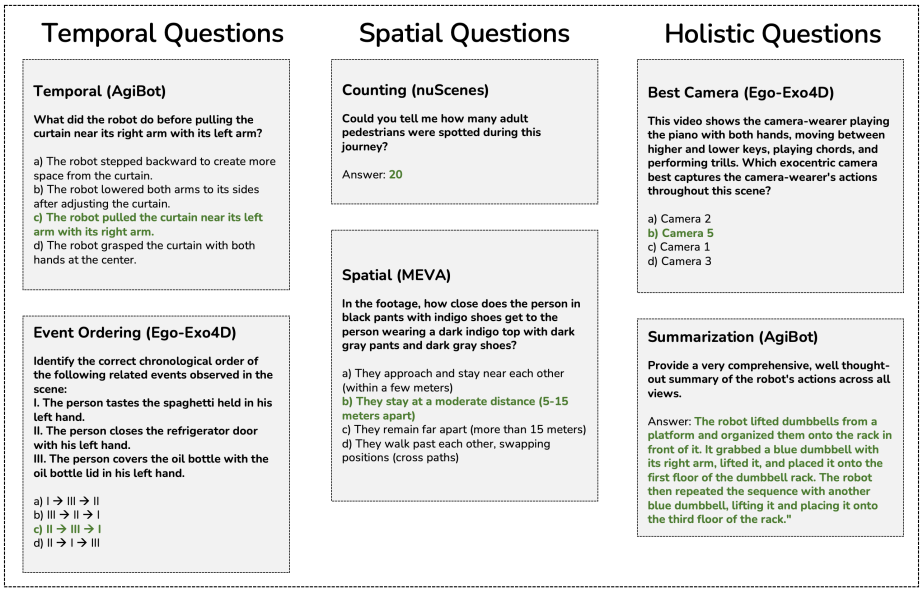

Fig. 3. Question categories in `CrossView`. The benchmark includes six types of multi-camera reasoning tasks grouped into three categories: temporal reasoning (temporal queries and event ordering), spatial reasoning (counting and spatial relationships), and holistic understanding (best-camera selection and scene summarization). Each question requires integrating information across multiple synchronized camera views and spans diverse domains including robotics (AgiBot), autonomous driving (nuScenes), surveillance (MEVA), and egocentric–exocentric interaction (Ego-Exo4D).

tasks, it selects a non-ego reference object, retrieves its orientation quaternion, and computes target positions in that object's local coordinate frame. Multi-hop spatial queries are created by choosing a reference node and at least two target objects at clearly different metric distances, enabling comparative-proximity questions grounded in spatial metadata rather than 2D pixel distance. For example: "Where is the traffic pole relative to the parked blue car that is parked directly adjacent to the main road?"

Global Scene Summarization. Summarization evaluates the model's ability to comprehensively and holistically understand a long video. Rather than querying a single timestamp, the engine aggregates the entire STSG into a dense temporal narrative by sampling frame-level captions and significant object-centric events at regular intervals. The resulting timeline gives a compressed yet representative overview of the scene's dynamics, which prompts the model to produce a comprehensive summary accounting for the ego-actor's primary interactions and the evolution of the environment across all camera perspectives.

Optimal Viewpoint Selection (Best Camera). The best camera category evaluates whether a model can identify the most informative perspective for a given event. The engine tracks each targeted object or activity's visibility metadata across all camera sensors and identifies the "optimal" camera view using the most informative perspective available in the STSG. Specifically, this is the natively annotated best exocentric view in Ego-Exo4D or the camera that most persistently observes the activity in MEVA. The resulting question presents an activity and asks the model which camera provided the most persistent or clear visual evidence, effectively testing spatial-semantic data indexing.

Object Instance Counting. Counting questions test the model's ability to distinguish and track individual instances over time. To do this, the engine counts unique identifiers (UIDs) across the full temporal span of the STSG, restricted to selected semantic classes. Aggregating UIDs rather than per-frame detections keeps the ground truth accurate under occlusion, reappearance, or movement between camera views, and requires persistent identity tracking across the entire video. An example of this is: "How many pedestrians are located near the red building directly adjacent to the main road?"

Chronological Event Ordering. Event ordering tests the model's ability to reconstruct a scene's temporal flow. The engine scans the STSG for a chain of three to five distinct, diverse events and, via a buffer-frame heuristic, selects events forming a clear chronological sequence with minimal overlap. Once shuffled and presented to the model, they must be restored to their original chronological order using semantic and temporal cues from the video, evaluating how well the model links distinct object-centric actions into a coherent timeline. An example is: "Read the events below and organize them as they appear in the multi-camera video stream: 1. The delivery truck appears and stops behind the yellow car, 2. The yellow car stops at the intersection, 3. A pedestrian crosses the road, 4. A yellow-vested cyclist on a white bike turns right."

4 Results

4.1 Evaluation Strategy

We evaluate eleven VLMs spanning four families (Qwen, InternVL, GPT, and Gemma) on `CrossView`. All results are obtained via uniformly sampling the VLM where frames are drawn independently from each camera. We report accuracy on multiple-choice question types (counting, temporal reasoning, event ordering, spatial reasoning, and camera identification) and evaluate open-ended summarization separately via ROUGE scores. Qwen, GPT, and Gemma models sample 16 frames per camera; InternVL models sample 8 frames per camera on the nuScenes and Ego-Exo4D subsets, and 4 frames on MEVA subset due to context-length constraints.

Table 2. Uniform sampling accuracy (%) on the vehicle-centric and robot-centric subsets of `CrossView`. Results are reported for nuScenes and AgiBot across counting, event ordering, spatial, and temporal reasoning tasks.

		nuScenes				AgiBot	
Family	Model	Counting	Event Ordering	Spatial	Temporal	Temporal	Event Ordering
Qwen	Qwen2.5-3B	32.5	36.8	30.1	32.2	52.0	50.8
	Qwen2.5-7B	27.5	24.4	**45.9**	21.6	58.8	54.4
	Qwen3-4B	46.6	**46.4**	28.9	33.8	66.4	49.2
	Qwen3-8B	43.2	41.2	33.1	36.0	**69.6**	49.2
InternVL	InternVL2-8B	28.8	36.8	37.3	29.2	42.0	40.4
	InternVL2.5-8B	37.1	28.4	32.1	33.4	60.0	46.4
	InternVL3.5-4B	40.2	30.0	26.1	38.4	52.8	42.0
	InternVL3.5-14B	46.2	36.4	43.3	**44.8**	54.8	47.6
GPT	GPT-5.2	**47.8**	29.2	36.5	25.4	66.8	**66.0**
Gemma	Gemma-3-4B	44.7	41.6	18.6	38.2	50.8	36.8
	Gemma-3-12B	43.8	43.2	26.4	35.6	60.4	41.2

Table 3. Uniform sampling accuracy (%) on the human-centric subsets of `CrossView`. Results are shown for Ego-Exo4D and MEVA across temporal reasoning, event ordering, counting, spatial reasoning, and best-camera identification tasks.

		Ego-Exo4D			MEVA				
Family	Model	Temporal	Event Ordering	Best Camera	Counting	Event Ordering	Spatial	Temporal	Best Camera
Qwen	Qwen2.5-3B	50.0	48.0	20.6	22.9	21.9	41.8	32.5	27.3
	Qwen2.5-7B	45.2	49.2	25.8	11.1	**83.2**	17.6	49.2	34.1
	Qwen3-4B	48.8	50.0	33.2	20.4	34.7	**49.0**	48.9	34.8
	Qwen3-8B	51.6	46.4	28.8	23.8	36.7	40.4	52.1	**38.6**
InternVL	InternVL2-8B	28.8	43.6	20.2	18.9	41.4	47.3	52.8	36.6
	InternVL2.5-8B	54.0	41.6	26.0	16.2	41.4	35.0	52.5	30.3
	InternVL3.5-4B	42.8	41.6	23.8	14.3	27.9	46.5	51.9	36.1
	InternVL3.5-14B	**54.4**	43.2	27.6	15.5	26.8	45.5	**60.0**	29.4
GPT	GPT-5.2	49.2	**52.8**	**34.6**	27.8	21.9	31.6	23.3	34.8
Gemma	Gemma-3-4B	39.6	40.8	26.6	**47.8**	7.4	32.5	46.2	27.3
	Gemma-3-12B	47.6	36.8	31.4	36.9	19.2	46.2	51.1	37.0

4.2 Main Results

Multiple-choice accuracy results are reported in Tables 2 and 3 for vehicle/robot-centric and human-centric benchmarks respectively. Summarization ROUGE scores are reported in Table 4. From our evaluations, we have the following insights:

1. **Multi-camera reasoning exposes a fundamental gap that model scale does not close.** Across all four benchmarks, no model family achieves consistently strong performance: notably, GPT-5.2 scores below 50% on nuScenes temporal reasoning and below 35% on camera identification in Ego-Exo4D, barely above the 25% expected from random guessing on these questions (Tables 2 and 3). Crucially, these same models approach saturation on standard single-camera benchmarks, ruling out general visual or language

Table 4. Summarization performance on `CrossView` via ROUGE scores (×100) under uniform frame sampling. R-1 and R-2 denote ROUGE-N overlap of unigrams and bigrams, respectively, and R-L denotes ROUGE-L, based on the longest common subsequence. Frame counts match the evaluation settings used in Tables 2 and 3.

		R-1				R-2				R-L			
Family	Model	nuScenes	Ego-Exo4D	AgiBot	MEVA	nuScenes	Ego-Exo4D	AgiBot	MEVA	nuScenes	Ego-Exo4D	AgiBot	MEVA
Qwen	Qwen2.5-3B	**26.3**	28.8	34.4	9.5	**5.3**	4.3	7.9	0.5	15.0	17.7	21.8	7.6
	Qwen2.5-7B	24.7	18.6	33.2	**12.3**	4.6	3.7	**11.4**	0.3	15.2	13.1	**24.2**	**11.5**
	Qwen3-4B	17.9	25.9	21.8	6.7	4.9	4.3	6.3	**1.0**	11.8	15.5	14.3	4.9
	Qwen3-8B	16.4	28.0	29.1	6.5	4.5	3.8	7.5	0.8	10.9	16.8	18.4	5.0
InternVL	InternVL2-8B	16.1	23.0	31.4	5.7	4.0	2.6	6.2	0.6	10.8	15.8	21.2	4.2
	InternVL2.5-8B	15.9	26.9	31.0	9.8	4.1	3.1	6.1	0.8	10.8	17.3	20.3	6.8
	InternVL3.5-4B	23.7	29.3	31.9	6.2	5.2	4.2	6.6	0.5	15.4	18.3	20.6	4.9
	InternVL3.5-14B	25.3	28.9	31.5	10.1	5.1	4.0	6.6	0.6	**15.7**	17.9	20.5	8.0
GPT	GPT-5.2	11.5	**37.5**	**35.1**	10.1	2.5	**7.2**	9.6	0.6	7.0	**21.2**	22.3	6.3
Gemma	Gemma-3-4B	14.9	23.7	31.6	7.0	3.2	2.5	6.5	0.5	9.2	15.5	20.7	5.5
	Gemma-3-12B	15.0	27.8	34.3	6.7	4.0	4.2	8.1	0.6	9.6	17.7	22.9	5.6

capacity as the bottleneck. The deficit is uniform across architectures and parameter counts, pointing instead to a systematic absence of multi-camera understanding in current pretraining data. A representative failure is shown in Figure 4, where the model collapses a right-side object into a front-of-vehicle judgment because individual camera views are reasoned over in isolation rather than jointly. Restricting inputs to a single camera confirms that many questions genuinely require evidence from multiple viewpoints (Table 5), emphasizing that multi-camera reasoning is an inherently distinct problem.

2. **Tasks that require synthesizing information *across* cameras are consistently the hardest.** Counting objects over multiple cameras and identifying the best camera for a given action, the two categories that require reasoning *jointly* across synchronized feeds, consistently yield the lowest scores. Counting on MEVA and nuScenes falls within the 25–47% range (Table 2), and best-camera identification on Ego-Exo4D spans only 20–35% across all models (Table 3), well below the 40–55% those same models achieve on temporal and event-ordering tasks from identical videos, suggesting that these models are often guessing rather than reasoning about viewpoint utility.
3. **Scene scale and camera density are the primary drivers of benchmark difficulty.** Performance degrades monotonically as scene complexity grows. AgiBot, a dataset that involves a small, controlled environment with fixed-camera robot-manipulation tasks, yields the highest absolute accuracies (up to 69.6% on temporal reasoning) whereas MEVA, a wide-area outdoor deployment with many overlapping, concurrently active cameras, produces the lowest scores in every category, most severely on counting and summarization tasks (Tables 3 and 4). This showcases how complex scene and environment reasoning cannot be decomposed into a single-camera reasoning task.

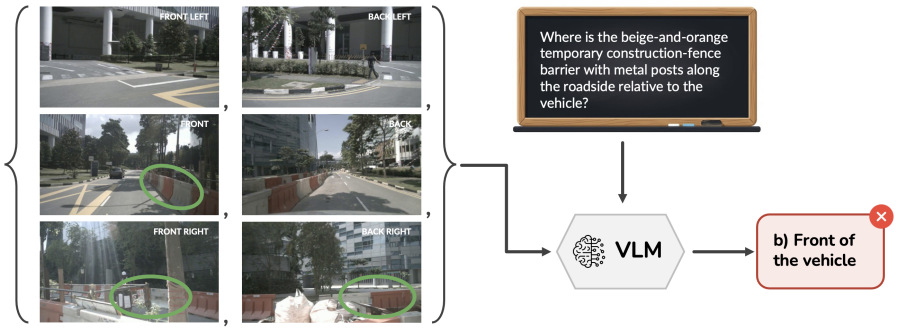

Fig. 4. Qualitative failure case. A driving example from `CrossView` under uniform sampling, with frames drawn independently from each of the six surround-view cameras. The VLM localizes the roadside construction-fence barrier as **b)** *Front of the vehicle*, likely because the front-right and back-right views capture it from a near front-on perspective; the correct answer is **c)** *Right of the vehicle*.

5 Discussion

5.1 Is There a Best Camera?

A natural question is whether a single "best" camera can substitute for the full camera set and yield a similar accuracy. To investigate this, we compare feeding only the best single camera to Qwen2.5-7B-Instruct against the full multi-camera uniform results from Tables 2 and 3, shown in Table 5. For Ego-Exo4D we test two candidates chosen for maximal scene coverage: the best-performing ground-truth exocentric camera (Best Exo) and the egocentric camera (Ego); for nuScenes, we select the front-facing vehicle camera (Front).

We observe that no single camera reliably substitutes for the full camera set. Multi-camera inputs yield consistent gains on counting and spatial reasoning, tasks that inherently require integrating evidence across viewpoints, with the front-camera baseline on nuScenes dropping 8.8% on counting and 4.4% on spatial reasoning. Temporal reasoning on Ego-Exo4D similarly favors multi-camera input by roughly 4.5–4.8% regardless of whether the egocentric or best exocentric view is used, indicating that even the most informative single perspective loses temporally salient cues captured by peripheral cameras.

The one reversal (nuScenes event ordering where the front camera leads by 7.6%) does not imply that a single view is superior. In reality, it is consistent with the finding in Section 4 that models often struggle to synthesize evidence *across* cameras: faced with redundant or conflicting views from six feeds, restricting input to the most semantically rich view can reduce noise on tasks with a dominant egocentric axis. Together, these results locate the bottleneck in the absence of principled cross-camera reasoning, a deficit that better camera selection alone cannot overcome.

Table 5. Effect of restricting inputs to a single camera on `CrossView` for Qwen2.5-7B-Instruct. Compared to multi-camera inputs, single-camera views often reduce performance, showing that many questions require evidence from multiple viewpoints. Deltas are relative to the multi-camera baseline for Qwen2.5-7B from Tables 2 and 3 (green = single-camera better, red = multi-camera better, gray = no change).

Dataset	Input	Counting	Temporal	Event Ordering	Spatial
Ego-Exo4D	Best Exo	–	40.7 (−4.5)	47.3 (−1.9)	–
	Ego	–	40.4 (−4.8)	48.8 (−0.4)	–
nuScenes	Front	18.7 (−8.8)	21.6 (±0.0)	32.0 (+7.6)	41.5 (−4.4)

Table 6. Comparison of uniform and stitched frame sampling on `CrossView` for Qwen2.5-7B. Stitched sampling improves performance on several tasks requiring cross-view spatial reasoning (*e.g.*, counting and event ordering on nuScenes), but gains are dataset-dependent. Deltas are relative to the uniform baseline from Tables 2 and 3 (green = stitched better, red = uniform better).

Dataset	Strategy	Counting	Temporal	Event Ordering	Spatial
nuScenes	Uniform	27.5	21.6	24.4	45.9
	Stitched	38.0 (+10.5)	26.2 (+4.6)	39.6 (+15.2)	46.3 (+0.4)
AgiBot	Uniform	–	58.8	54.4	–
	Stitched	–	58.4 (−0.4)	48.4 (−6.0)	–
Ego-Exo4D	Uniform	–	45.2	49.2	–
	Stitched	–	53.6 (+8.4)	48.0 (−1.2)	–
MEVA	Uniform	11.1	49.2	83.2	17.6
	Stitched	13.3 (+2.2)	52.2 (+3.0)	70.9 (−12.3)	34.2 (+16.6)

5.2 Uniform vs. Stitched Sampling

We compare two multi-camera frame sampling strategies: *uniform*, which samples N frames independently per camera, and *stitched*, which tiles all cameras into a single composite image per timestep (as seen in Figure 1). Table 6 reports stitched results.

Stitched sampling helps most on tasks that benefit from simultaneous spatial comparison across views. On nuScenes, where six cameras form a near-contiguous 360° surround, stitching improves results across all categories, with counting and event ordering gaining 10.5 and 15.2 points respectively. The largest single gain occurs on MEVA spatial reasoning (+16.6), where wide-area scenes with many concurrent cameras benefit most from preserving inter-camera layout within a single context window. However, on AgiBot with the most controlled, low-clutter environment, stitching slightly degrades both metrics, suggesting that for simple scenes with few distractors, the added visual complexity of tiled frames introduces noise rather than useful context. Together, these results showcase how

frame sampling is largely task-dependent, motivating work on adaptive frame sampling for complex scenes.

6 Conclusion

We introduced `CrossView`, a benchmark for evaluating multi-camera reasoning in VLMs across autonomous driving, surveillance, egocentric–exocentric interaction, and robotics. Unlike traditional VQA benchmarks that operate on a single visual stream, `CrossView` requires integrating information across synchronized viewpoints to answer questions involving counting, spatial reasoning, event ordering, and camera selection. Our evaluation reveals a consistent *multi-camera reasoning gap*: even state-of-the-art VLMs struggle to combine evidence across views, with particularly low performance on tasks such as cross-view counting and identifying the most informative camera. This deficit persists across model families and scales, suggesting a structural limitation in how current models process and learn from multi-view data.

Real-world perception, from autonomous vehicles to robotic platforms, relies on *networks* of cameras rather than single viewpoints, and `CrossView` exposes how far current models remain from reasoning effectively in these settings. We hope `CrossView` encourages the development of architectures, training objectives, and data pipelines that explicitly support cross-view evidence integration, advancing VLMs toward robust multi-camera understanding.

Acknowledgements

This material is based upon work supported in part by the Office of Naval Research (ONR) under Grant No. N00014-22-1-2254. Additionally, this work was supported by the Defense Advanced Research Projects Agency (DARPA) contract DARPA ANSR: RTX CW2231110. Approved for Public Release, Distribution Unlimited.

References

1. 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)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), `https://arxiv.org/abs/2502.13923`
3. Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., Gao, S., He, X., Hu, X., Huang, X., et al.: Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. arXiv preprint arXiv:2503.06669 (2025)

4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
5. Chavdarova, T., Baqué, P., Bouquet, S., Maksai, A., Jose, C., Bagautdinov, T., Lettry, L., Fua, P., Van Gool, L., Fleuret, F.: Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5030–5039 (2018)
6. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
8. Chiu, H.k., Li, J., Ambruş, R., Bohg, J.: Probabilistic 3d multi-modal, multi-object tracking for autonomous driving. In: 2021 IEEE international conference on robotics and automation (ICRA). pp. 14227–14233. IEEE (2021)
9. Choi, M., Goel, H., Omama, M., Yang, Y., Shah, S., Chinchali, S.: Towards neuro-symbolic video understanding. In: European Conference on Computer Vision. pp. 220–236. Springer (2024)
10. Corona, K., Osterdahl, K., Collins, R., Hoogs, A.: Meva: A large-scale multiview, multimodal video dataset for activity detection pp. 1060–1068 (2021)
11. Ding, X., Han, J., Xu, H., Liang, X., Zhang, W., Li, X.: Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13668–13677 (2024)
12. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
13. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24108–24118 (2025)
14. Gemma Team: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
15. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: 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. pp. 6904–6913 (2017)
16. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video pp. 18995–19012 (2022)
17. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., 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. pp. 19383–19400 (2024)
18. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition. pp. 2495–2504 (2020)
19. He, Y., Huang, Y., Chen, G., Lu, L., Pei, B., Xu, J., Lu, T., Sato, Y.: Bridging perspectives: A survey on cross-view collaborative intelligence with egocentric-exocentric vision. International Journal of Computer Vision **134**(2), 62 (2026)
20. He, Y., Huang, Y., Chen, G., Pei, B., Xu, J., Lu, T., Pang, J.: Egoexobench: A benchmark for first-and third-person view video understanding in mllms. Advances in Neural Information Processing Systems **38** (2026)
21. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
22. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. Advances in Neural Information Processing Systems **36**, 20482–20494 (2023)
23. Hou, Y., Zheng, L., Gould, S.: Multiview detection with feature perspective transformation. In: European Conference on Computer Vision. pp. 1–18. Springer (2020)
24. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
25. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
26. Ilaslan, M., Song, C., Chen, J., Gao, D., Lei, W., Xu, Q., Lim, J., Shou, M.: Gazevqa: A video question answering dataset for multiview eye-gaze task-oriented collaborations. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 10462–10479 (2023)
27. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2758–2766 (2017)
28. Jia, B., Chen, Y., Huang, S., Zhu, Y., Zhu, S.C.: Lemma: A multi-view dataset for le arning m ulti-agent m ulti-task a ctivities. In: European Conference on Computer Vision. pp. 767–786. Springer (2020)
29. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
30. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024)
31. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 152–159 (2014)
32. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. vol. 47, pp. 2020–2036. IEEE (2024)
33. Liu, Y., Yan, J., Jia, F., Li, S., Gao, A., Wang, T., Zhang, X.: Petrv2: A unified framework for 3d perception from multi-camera images pp. 3262–3272 (2023)
34. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
35. Majumder, S., Nagarajan, T., Al-Halah, Z., Pradhan, R., Grauman, K.: Which viewpoint shows it best? language for weakly supervising view selection in multi-

view instructional videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29016–29028 (2025)
36. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023)
37. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019)
38. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. vol. 65, pp. 99–106. ACM New York, NY, USA (2021)
39. Park, S.Y., Cui, C., Ma, Y., Moradipari, A., Gupta, R., Han, K., Wang, Z.: Nuplanqa: A large-scale dataset and benchmark for multi-view driving scene understanding in multi-modal large language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8066–8076 (2025)
40. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: European conference on computer vision. pp. 194–210. Springer (2020)
41. Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G.: Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4542–4550 (2024)
42. Reilly, D., Govind, M.K., Xue, L., Das, S.: From my view to yours: Ego-to-exo transfer in vlms for understanding activities of daily living. arXiv preprint arXiv:2501.05711 (2025)
43. Ristani, E., Tomasi, C.: Features for multi-target multi-camera tracking and re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6036–6046 (2018)
44. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European conference on computer vision. pp. 146–162. Springer (2022)
45. Shah, S., Sharan, S., Goel, H., Choi, M., Munir, M., Pasula, M., Marculescu, R., Chinchali, S.: Neus-qa: Grounding long-form video understanding in temporal logic and neuro-symbolic reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 8805–8813 (2026)
46. Tang, Z., Naphade, M., Liu, M.Y., Yang, X., Birchfield, S., Wang, S., Kumar, R., Anastasiu, D., Hwang, J.N.: Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8797–8806 (2019)
47. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
48. Wang, S., Yu, Z., Jiang, X., Lan, S., Shi, M., Chang, N., Kautz, J., Li, Y., Alvarez, J.M.: Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In: Proceedings of the computer vision and pattern recognition conference. pp. 22442–22452 (2025)
49. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: Conference on robot learning. pp. 180–191. PMLR (2022)

50. Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., Bansal, M.: Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3272–3283 (2025)
51. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 79–88 (2018)
52. Wen, L., Lei, Z., Chang, M.C., Qi, H., Lyu, S.: Multi-camera multi-target tracking with space-time-view hyper-graph. International Journal of Computer Vision **122**(2), 313–333 (2017)
53. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
54. Wu, J., Echterhoff, J., Han, K., Abdelraouf, A., Gupta, R., McAuley, J.: Pdb-eval: An evaluation of large multimodal models for description and explanation of personalized driving behavior. In: 2025 IEEE Intelligent Vehicles Symposium (IV). pp. 242–248. IEEE (2025)
55. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European conference on computer vision (ECCV). pp. 767–783 (2018)
56. Ye, J., Wang, Z., Sun, H., Chandrasegaran, K., Durante, Z., Eyzaguirre, C., Bisk, Y., Niebles, J.C., Adeli, E., Fei-Fei, L., et al.: Re-thinking temporal search for long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8579–8591 (2025)
57. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 9127–9134 (2019)
58. Zhang, H., Chu, Q., Liu, M., Shi, H., Wang, Y., Nie, L.: Exo2ego: Exocentric knowledge guided mllm for egocentric video understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 12502–12510 (2026)
59. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: European conference on computer vision. pp. 868–884. Springer (2016)
60. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision. pp. 1116–1124 (2015)
61. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., et al.: Mlvu: Benchmarking multi-task long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13691–13701 (2025)
62. Zhu, C., Wang, T., Zhang, W., Pang, J., Liu, X.: Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. arXiv preprint arXiv:2409.18125 (2024)
63. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023)