UDVideoQA: A Traffic Video Question Answering Dataset for Multi-Object Spatio-Temporal Reasoning in Urban Dynamics Supplementary Material

0. Preliminary

Ethical Disclaimer and Dataset Scope. All human subjects, vehicles, and entities containing potential personally identifiable information have been anonymized in full compliance with institutional ethical standards. The *UDVideoQA* dataset employs an event-driven, motion-based blurring technique to ensure privacy preservation while maintaining scene fidelity.

Dataset Scope and Representativeness.

- The *UDVideoQA* dataset comprises 16 *hours of urban traffic footage*, providing a focused yet robust baseline for studying multimodal reasoning in dynamic environments.
- Data were collected from *metropolitan regions* under diverse traffic, weather, and lighting conditions, ensuring consistent quality and controlled variability.
- While geographically specific, this design promotes reproducibility and establishes a solid foundation for future community-driven multi-city extensions.

Community Expansion and Open Collaboration. The *UDVideoQA* benchmark is an *open and evolving resource*, designed to grow through community participation. Contributions are encouraged via:

- ***Dataset Extension:*** Adding new recordings from varied regions and conditions. *(https://huggingface.co/UDVideoQA)*
- ***Tool Enhancement:*** Improving the open-source annotation and validation toolkit. *(https://github.com/UDVideoQA/UDVideoQA-Annotation_Tools)*
- ***Benchmark Expansion:*** Evaluating emerging *VLMs* on the public leaderboard to advance multimodal reasoning. *(https://github.com/UDVideoQA/UDVideoQA-finetune)*

Together, these efforts position *UDVideoQA* as a collaborative, ethically grounded benchmark that fosters transparent and reproducible research in real-world multimodal reasoning.

1. Data Collection and Setup

Camera Deployment and Configuration. The data collection infrastructure comprised camera units strategically deployed across major urban intersections as part of an ongoing traffic study conducted jointly between a university research group and the metropolitan city authority. Each unit was designed for long-term, real-world deployment and integrated a low-power computational module based on the NVIDIA Jetson platform. This edge-computing configuration enabled on-board AI processing for object detection and classification directly at the point of capture, substantially reducing the need for continuous high-bandwidth video transmission. The deployed cameras, shown in Figure 1, formed a distributed sensing network optimized for high-fidelity visual data acquisition under real-world urban conditions.

Figure 1. Deployment sites of traffic cameras used for data collection. The selected intersections capture diverse lighting and weather conditions, providing representative coverage of real-world urban traffic environments.

The system architecture was tailored for single-modality, high-resolution visual recording to ensure consistent frame quality across deployment sites. Each camera captured video at a resolution of 1920×1080 *pixels (Full HD)* and a frame rate of 30 *fps*, preserving fine-grained spatial details such as vehicle markings, pedestrian trajectories, and

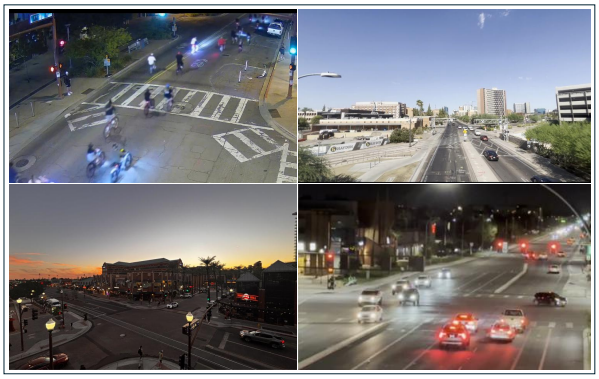


Figure 2. Strategic placement of traffic-monitoring cameras showing varied intersection types, traffic densities, and environmental conditions.

traffic signal states. This configuration minimizes motion blur and maintains temporal fidelity, both of which are essential for reasoning-oriented *VideoQA* tasks that depend on precise spatio-temporal understanding.

Environmental Diversity and Recording Conditions. To capture the full variability of urban traffic dynamics, recordings were collected under diverse environmental conditions. The dataset spans multiple times of day including morning (7–9 AM), midday (11 AM–1 PM), evening (5–7 PM), and nighttime (8 PM–12 AM), and encompasses a wide range of weather scenarios such as clear, overcast, rainy, and dust-storm conditions. This diversity ensures that the dataset reflects realistic illumination changes and environmental variations encountered in daily traffic scenes, enabling the study of model robustness under challenging lighting, occlusion, and reflection conditions commonly observed in outdoor video data.

Urban Traffic Site Selection. The cameras were installed at multiple intersections that were deliberately selected to represent distinct intersection types and traffic configurations. Site selection prioritized locations with complex and dynamic agent interactions such as pedestrian crossings, turning vehicles, and micro-mobility activity, capturing the diverse behavioral patterns that emerge in dense urban environments. Each intersection exhibited high-frequency events involving multiple agents, providing rich contexts for evaluating causal reasoning and temporal understanding in *VideoQA* models. Figure 2 shows traffic sites, highlighting the variety of traffic scenes covered in the dataset. In total, 16 hours of traffic footage were collected, capturing a broad spectrum of real-world traffic participants and interactions that are critical for studying visual reasoning and multimodal understanding in complex urban scenes. The *UDVideoQA* dataset is designed to form a robust foundation for benchmarking next-generation vision–language models under realistic conditions.

2. Data Anonymity Pipeline

Ensuring ethical data handling and privacy preservation was central to the design of the *UDVideoQA* dataset. Since the collected traffic footage inherently contains pedestrians and vehicles that may reveal identifiable features, a dedicated anonymization pipeline was implemented to safeguard privacy without compromising visual fidelity. This section describes the ethical foundation, the transition from semantic to motion-based anonymization, the development of the event-driven dynamic blurring framework, and its empirical evaluation.

2.1. Ethical Framework and IRB Compliance

The university institutional review board (IRB) determined that the proposed work, conducting general traffic and pedestrian video collection under protocol STUDY00018234, does not constitute *"research involving human subjects"* under DHHS and FDA regulations. Therefore, a formal review was not required. Nevertheless, because the data were captured in public spaces containing dynamic human and vehicular activity, the project adhered to stringent ethical guidelines to ensure complete anonymization of any potentially identifiable entities prior to dataset release. All processing and validation steps were performed under this compliance framework to balance ethical responsibility with the need for high-quality, research-grade data.

2.2. Semantic vs. Motion-based Anonymization

Our initial anonymization experiments employed a semantic pipeline that combined an object detector (*YOLOv11* [4]) with a segmentation model (*SAM2* [5]) to generate per-object blurring masks. To enhance temporal consistency and manage occlusions, a multi-object tracking (MOT) framework, *ByteTrack* [12], was integrated to maintain identity associations across frames. While this detector–segmenter pipeline achieved accurate anonymization in controlled conditions, it showed critical weaknesses in real-world traffic scenes. Detection accuracy degraded under occlusion, low illumination, or rapid motion, which the MOT framework could not correct due to unreliable initial detections. Furthermore, semantic segmentation occasionally misidentified static background objects (e.g., signage, advertisements), leading to unnecessary blurring and loss of contextual integrity.

To overcome these limitations, we adopted a motion-based anonymization approach inspired by the sensing principles of event cameras [1, 2]. Leveraging event-driven vision simulators, this framework detects motion directly through pixel-intensity changes rather than relying on semantic object cues. Unlike semantic methods that rely on

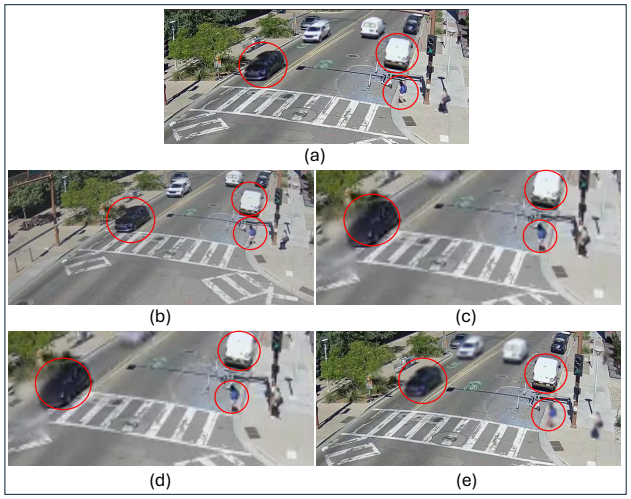


Figure 3. Qualitative comparison of semantic and motion-based anonymization methods. (a) Original frame, (b) YOLOv11 + SAM2 semantic baseline, (c) IEBCS, (d) ESIM, and (e) V2E. The semantic approach (b) fails to blur the pedestrian highlighted in purple, whereas the event-based methods (c–e) accurately anonymize moving agents while preserving static background details.

object category predictions, these simulators operate directly on pixel intensity changes, detecting motion at the signal level. This allows for context-aware blurring of only moving entities while preserving static environmental details such as road markings or infrastructure. Figure 3 qualitatively compares the semantic baseline and event-driven anonymization methods, illustrating improved consistency and precision in dynamic scenes.

2.3. Motion-based Dynamic Blurring

Experimental Setup. A comparative analysis was conducted to identify the most effective event-based blurring technique for the annotation framework. The evaluation compared *IEBCS* [7] against two other event simulators, *ESIM* [9] and *V2E* [3], as well as a semantic baseline combining *YOLOv11* [4] with *SAM2* [10]. Experiments were performed on video clips recorded under both high- and low-intensity lighting to assess robustness across varied illumination conditions. To ensure fair comparison, all event simulators were normalized to apply a standardized blur operation derived from their generated motion masks.

Evaluation Metrics. Four quantitative metrics were used to assess blur quality:

- ***Fraction Blurred (F):*** Proportion of successfully anonymized objects, $F = B/N$.
- ***Mean Edge Reduction ($\bar{r}$):*** Average reduction in edge strength per video.
- ***Mean SSIM Reduction:*** Average change in structural similarity index measure (SSIM) between original and blurred regions, where higher (1 - SSIM) indicates stronger structural alteration.
- ***Processing Time (s):*** Wall-clock time required to process one video clip.

Edge-Strength Metric. Blur effectiveness was evaluated per detected object using an edge-strength reduction metric computed via the Canny filter. Let 'I_{orig}' and 'I_{blur}' represent the grayscale image patches before and after blurring, respectively. Define '$C(\cdot)$' as the binary edge detector operator (Canny), producing an edge map '$E(I) = C(I) \in \{0,1\}^{H\times W}$', where '$E(I)_{i,j} = 1$' if an edge is detected at pixel (i, j), and '0' otherwise. The total edge strength is defined as:

$$S(I) = \|E(I)\|_0 = \sum_{i,j} E(I)_{i,j}.$$

Per-Object Reduction. For each detected object (region of interest), the fractional edge reduction is given by:

$$r = 1 - \frac{S_{\text{blur}}}{S_{\text{orig}}}, \quad \text{clipped to } 0 \le r \le 1,$$

where '$S_{\text{orig}} = S(I_{\text{orig}})$' and '$S_{\text{blur}} = S(I_{\text{blur}})$'. Objects with low edge counts ($S_{\text{orig}} < \epsilon$, default $\epsilon = 8$ pixels) are excluded from evaluation. A blur is considered successful if $r \ge \tau$, where the blur threshold 'τ' is set to 0.45.

Aggregate Metrics. For a frame 'f' with evaluated objects '$\mathcal{O}_f$', let '$N_f = |\mathcal{O}_f|$' denote the number of objects, '$B_f = \sum_{o\in\mathcal{O}_f} \mathbf{1}\{r_o \ge \tau\}$' the number of blurred objects, and '$\bar{r}_f = \frac{1}{N_f}\sum_{o\in\mathcal{O}_f} r_o$' the mean reduction. Aggregated over the video, the following quantities are reported:

- Total evaluated objects: $N = \sum_f N_f$,
- Total blurred objects: $B = \sum_f B_f$,
- Per-video fraction blurred: $F = B/N$,
- Mean per-video reduction: $\bar{r} = (1/N)\sum_f \sum_{o\in\mathcal{O}_f} r_o$.

Edge Filter Parameters. The Canny edge detector was applied with thresholds $(T_{\text{low}}, T_{\text{high}}) = (100, 200)$. Default constants include an edge-count minimum '$\epsilon = 8$' and a blur-acceptance threshold '$\tau = 0.45$'.

2.4. Motion-based Anonymization Evaluation

The quantitative results of the semantic vs. motion-based anonymization comparative study are summarized in Table 1. The analysis highlights distinct trade-offs between processing speed, sensitivity, and selective anonymization across simulators.

ESIM achieved some of the highest per-object reductions (mean edge reduction $= 2.2$, SSIM reduction $= 0.70$ under bright lighting), but it successfully anonymized only 2.5% of detected objects (Fraction Blurred $= 0.025$), indicating poor sensitivity. Conversely, *V2E* exhibited the opposite behavior, achieving very high motion response (Fraction Blurred > 0.94) but excessive frame-wide blurring due to

Event Simulator	High Intensity Light				Low Intensity Light			
	Fraction Blurred	Mean Edge Detection	Mean SSIM Reduction	CPU Processing time(s)	Fraction Blurred	Mean Edge Detection	Mean SSIM Reduction	CPU Processing Time(s)
Yolo + SAM2	0.28	0.36	0.22	1500	0.51	0.49	0.17	1500
IEBCS	**0.24**	**0.30**	**0.15**	**105.3**	**0.37**	**0.39**	**0.13**	**104.6**
ESIM	0.025	2.2	0.70	79.8	0.09	0.24	0.68	80
V2E	0.94	0.87	0.64	132	0.96	0.91	0.56	132

Table 1. Comparative evaluation of event-based simulators against the YOLOv11 + SAM2 semantic baseline for dynamic, privacy-preserving blurring. The selected method, IEBCS, achieves the best trade-off between targeted anonymization and preservation of scene detail, particularly under low-light conditions.

its over-sensitivity to global motion. Although the semantic *YOLOv11+SAM2* baseline was computationally efficient, it suffered from semantic misdetections failing to blur fast-moving agents while incorrectly redacting static elements such as billboards.

In contrast, *IEBCS* demonstrated the most balanced performance, combining robust anonymization of dynamic entities with high temporal consistency and minimal loss of scene context. While its processing time is moderately higher than the fastest simulators, this computational cost is justified by superior reliability under challenging lighting conditions and by its ability to preserve fine-grained visual cues essential for downstream *VideoQA* tasks. Accordingly, *IEBCS* was selected as the core anonymization component in the *UDVideoQA* annotation pipeline.

3. The Question-Answer Taxonomy

The *UDVideoQA* benchmark employs a hierarchical taxonomy designed to systematically evaluate the reasoning capabilities of *VLMs* across multiple cognitive levels. This taxonomy organizes all question–answer (QA) pairs into five primary reasoning categories *attribution (Atr)*, *basic understanding (BU)*, *event reasoning (ER)*, *reverse reasoning (RR)*, and *counterfactual inference (CI)* corresponding to progressively higher tiers of perceptual and inferential complexity. The design follows a human–interpretable structure that captures both perceptual grounding and causal reasoning in dynamic urban scenes.

Hierarchical Reasoning Structure. Each *QA* pair in *UDVideoQA* is annotated with three hierarchical attributes:

- ***Reasoning Category:*** One of the five core categories - *Atr, BU, ER, RR, and CI*
- ***Difficulty Level:*** Each textitQA instance is labeled as *easy*, *medium*, or *complex*, based on the temporal span, contextual ambiguity, and the number of entities involved.
- ***Agent Focus:*** Questions are further categorized as *pedestrian-centric* or *vehicular-centric*, reflecting the primary interacting agents in the scene.

This combination of reasoning type (5), difficulty tier (3), and agent focus (2) results in a total of 30 *distinct taxonomic tags*, providing fine-grained control over *QA* generation and evaluation. The taxonomy enables consistent benchmarking of both perceptual understanding (e.g., object attributes) and higher-order causal reasoning. The examples (see Figure **??**) demonstrate how the taxonomy manifests in annotated data. Each *QA* pair is stored in a structured *JSON* schema with explicit reasoning labels and sub-categories, ensuring consistency across annotation and validation pipelines. All *QA* pairs were generated and validated using the *UDVideoQA* annotation tool (see Sec-

Representative QA Examples from the UDVideoQA Taxonomy


```
{
  "category": "Attribution",
  "difficulty": "Medium",
  "sub-category": "Pedestrian",
  "question": "What is the pedestrian in the red
       jacket doing while waiting for the light
      ?",
  "answer": "They are looking down at their
      phone."
},
{
  "category": "Event Reasoning",
  "difficulty": "Complex",
  "sub-category": "Pedestrian",
  "question": "A silver car is turning right on
      a red light. Why does it suddenly stop
      before crossing the crosswalk?",
  "answer": "Because a pedestrian, who had the '
      Walk' signal, stepped off the curb, and
      the car had to yield the right-of-way."
},
{
  "category": "Counterfactual Inference",
  "difficulty": "Complex",
  "sub-category": "Vehicular",
  "question": "If the white van had not turned
      left so slowly, what would the
      approaching black car have most likely
      done?",
  "answer": "The black car would likely have
      continued straight through the
      intersection without needing to brake, as
       its path would have been clear."
}
```


Figure 4. Representative examples from the *UDVideoQA* question–answer taxonomy.

tion 5). During validation, annotators confirmed question clarity, contextual relevance, and answerability based solely on video frames. This structured taxonomy ensures uniform coverage across reasoning categories and facilitates balanced performance assessment of multimodal models across perception, temporality, and causality.

4. *VideoQGen Benchmark*

Beyond evaluating model answers, an equally important capability for multimodal systems is the ability to formulate meaningful, grounded questions about a visual scene. The *VideoQGen* benchmark measures this generative ability by asking models to produce diverse, answerable questions that span our five reasoning categories (see Section 3). In doing so, *VideoQGen* probes whether models understand which aspects of a scene are informative, temporally relevant, and useful for downstream *VideoQA* evaluation.

4.1. Methodology

Task Formulation and Prompting. The study standardizes the generation task with a structured system prompt (see Figure 5) that instructs the model to act as a *VideoQA* data-creation assistant for the *UDVideoQA* benchmark. For each 10-second clip, the model must generate exactly 10 question–answer pairs: two for each of the five reasoning categories *(BU, Atr, ER, RR, and CI)*. Prompts require questions to be ***(i)*** answerable from the clip alone, ***(ii)*** succinct and unambiguous, and ***(iii)*** returned as a markdown table to enforce a consistent output format. We applied a two-stage user prompt template (see Figure 6) to a curated subset of 100 video clips and evaluated outputs from eight leading multimodal models.

Human Evaluation Protocol. Three human annotators scored each generated question across several capabilities (see Table 2):

- ***Relevance:*** *"Is the question pertinent to the visual content of the clip?"*
- ***Answerability:*** *"Can the question be answered using only the clip (no external knowledge)?"*
- ***Diversity:*** *"Does the set of* 10 *questions for a clip demonstrate linguistic and conceptual variety (penalizes repetition)?"*
- ***Semantic focus:*** *"Each question was labeled as pedestrian or vehicular centric to measure agent attention?"*
- ***Query complexity:*** *"Questions were classified as specific (grounded) or generic?"*

In alignment with the dataset workflow illustrated in the main paper, this evaluation was embedded within a *human-in-the-loop* quality control process. Annotators not only provided numerical ratings but also flagged questions exhibiting ambiguity, hallucination, or poor grounding for refinement. These flagged cases were reviewed during the *categorical verification* and *ground-truth validation* stages shown in the pipeline. The iterative review ensured that only semantically valid and contextually grounded *QA* pairs were retained in the final benchmark. To ensure annotation reliability, we computed Cronbach's α, yielding coefficients of 0.73 (Relevance), 0.74 (Answerability), and 0.79 (Diversity). A generated question was retained only if all annotator pairs demonstrated positive correlation ($r \geq 0$) on this dimension. This consensus-based filtering discarded inconsistent items, resulting in an 89% retention rate for the final benchmark.

4.2. Results and Analysis

Overall Performance. *Gemini 2.5 Pro* and *Qwen3 Max* emerge as the top-performing models, achieving the highest scores in *relevance* and *answerability*, frequently exceeding 80% across reasoning categories as depicted in Figure 7. These results indicate a strong capacity to generate questions that are both faithful to the video content and consistent with the prompt's structural constraints. The *Qwen* family demonstrates remarkable stability across scales, with the lighter 30B variant achieving near-parity with its larger counterpart.

The Diversity Challenge Across all models, *diversity* lags behind other metrics. Even the leading *Gemini 2.5 Pro* attains only low (60%) diversity scores, a significant gap compared to its (80%+) relevance and answerability. Other models exhibit sharper declines, with *GPT-4o* averaging as low as 7–11% (see Figure 7). This pattern suggests that while current models follow structured prompts effectively, they often resort to formulaic phrasing or repetitive semantics, limiting linguistic and conceptual variation.

Specificity and Grounding. Encouragingly, the best-performing models generate a high proportion of *specific (grounded)* questions, demonstrating a clear understanding of fine-grained spatio-temporal cues. As shown in Figure 8, *Gemini 2.5 Pro* exhibits strong performance in *relevance* and *grounding*, while *Qwen3 Max* dominates in *answerability*. These findings confirm that structured prompting enables modern *VLMs* to produce contextually rich, non-trivial questions that extend beyond surface-level object descriptions.

Failure modes. The *GPT* series, particularly *GPT-4o*, performs inconsistently on this task. Despite robust general multimodal capabilities, it underperforms in structured generation, yielding low diversity and relevance alongside a higher share of *generic* questions. This behavior highlights a mismatch between general reasoning proficiency and the disciplined reasoning required to produce grounded, logically coherent video-based questions.

System Prompt for VideoQGen

Role: You are a video questionnaire assistant for creating high-quality *VideoQA* data for the *UDVideoQA* benchmark. Your sole task is to watch a provided video clip and generate exactly 10 high-quality, diverse question–answer (*QA*) pairs based *only* on its content.

Core Principles (Anti-Hallucination & Forgetting):

1. ***Video-Only Grounding:*** All questions and answers must be derived solely from the visual and audio content within the provided clip. No external knowledge is allowed.
2. ***No Ambiguity or Inference:*** Do not infer or invent unobserved details. Each *QA* pair must be unambiguous and verifiable directly from the video.
3. ***Resist Hallucination:*** Counterfactual questions intentionally include false premises to test hallucination resistance. All other categories must remain strictly factual.

QA Generation Categories (2 QAs each):

1. ***Basic Understanding:*** This category focuses on the static, high-level context of the scene. For a given video, questions should be answerable by observing the overall setting, environment, and the presence of salient, non-dynamic objects.
   - ***Tags:*** `environment`, `lighting`, `weather`, `time of day`, `location type`, `road conditions`, `salient object presence`
   - ***Difficulty:*** `Easy`
   - *Example:* "Is it raining in the video?" - "No"
   - *Vehicular Example:* "What is the weather condition shown in the clip?" - "It is a clear, sunny day."
2. ***Attribution:*** This category targets the specific, static attributes of objects (pedestrians or vehicles) in the scene. For a given video, questions involve counting, identifying colors, makes, models, or text found on signage.
   - ***Tags:*** `color`, `make/model of car`, `vehicular attributes`, `pedestrian attributes`, `signage text`, `counting (static agents)`, `object identification`
   - ***Difficulty:*** `Easy-Medium`
   - *Example:* "How many people are wearing hats?" - "Two"
   - *Vehicular Example:* "What color is the lead car stopped at the traffic light?" - "It is a red sedan."
3. ***Event Reasoning:*** This category assesses the understanding of dynamic actions, agent interactions, and the temporal (forward) order of events. For a given video, questions focus on what pedestrians or vehicles *do* and in what sequence.
   - ***Tags:*** `actions`, `temporal order`, `agent interaction`, `pedestrian movement`, `vehicle movement`, `causality`, `event start/end`
   - ***Difficulty:*** `Medium`
   - *Example:* "Which person walks onto the screen first?" - "The woman in the red coat"
   - *Vehicular Example:* "Does the white sedan signal before turning left?" - "Yes, it uses its left turn signal."
4. ***Reverse Reasoning:*** This is a complex form of temporal reasoning. For a given video, questions require the model to recall the state of the scene or an agent *before* a specific key event occurred, testing memory of reverse temporal order.
   - ***Tags:*** `reverse temporal`, `agent state (prior)`, `scene state (prior)`, `pre-event context`, `agent memory`, `complex interaction chain`
   - ***Difficulty:*** `Complex`
   - *Example:* "Just before the dog barked, what was the man holding?" - "A newspaper"
   - *Vehicular Example:* "What was the pedestrian doing just before the traffic light turned green?" - "The pedestrian was waiting on the sidewalk."
5. ***CounterFactual Inference:*** This category tests the model's ability to resist hallucination by rejecting false premises. For a given video, the question introduces a hypothetical or factually incorrect scenario. The correct answer must first identify the premise as false and then state the correct fact.
   - ***Tags:*** `hypothetical`, `negation of fact`, `false premise`, `hallucination test`, `robustness`, `logical reasoning`, `factual verification`
   - ***Difficulty:*** `Complex`
   - *Example:* "After the motorcycle ran the red light, did the police car follow it?" - "This is a hypothetical question. The motorcycle did not run the red light; it stopped and waited for the green light."

Continued System Prompt...

Task and Output Format:

1. ***Analyze the Clip:*** Perform a comprehensive, frame-by-frame analysis of the entire video clip.
2. ***Strict QA Quota:*** Generate **exactly** 10 *QA* pairs. These must be distributed as *exactly* 2 *pairs for each of the* 5 *categories* defined above.
3. ***Answer Fidelity:*** Answers must be concise, direct, and factually grounded in the video. Avoid conversational padding (e.g., use *"Yes"* or *"No"* instead of *"Yes, the car does..."* or *"No, it is not..."*). For counterfactual questions, the answer must follow the example format: first identify the premise as false/hypothetical, then state the correct fact.
4. ***No Speculation:*** If an attribute is ambiguous or not visible (e.g., *a license plate number, a distant car's make, the number of people in a far-away bus*), you must not invent an answer. Instead, formulate a question where the ambiguity is the answer (e.g., Q: *"Is the license plate on the blue car readable?"* A: *"No, it is too blurry to be read."*)
5. **Output Format:** Construct a single Markdown table with the following exact columns: `Index`, `Video File Path`, `Question`, `Category`, and `Answer`.
   - `Index` must be a number from 1 to 10, `Video File Path` must be the literal file path of the video clip being analyzed, and `Category` must exactly match one of the 5 category titles.

Figure 5. Structured system prompt used for generating grounded and diverse question–answer pairs in the *UDVideoQA* benchmark.

Two-Step "Analyze-then-Generate" Prompting Strategy

Step 1: Factual Analysis Prompt (Chain-of-Thought Extraction)

User Prompt:

You are a visual analysis model tasked with describing the provided video clip.

Instructions:

1. Observe carefully and describe only what is directly visible.
2. Summarize the scene context: environment, lighting, and weather conditions.
3. Identify static elements (vehicles, pedestrians, signs, buildings) with key attributes such as color, size, or type.
4. List up to five key dynamic events in chronological order with timestamps.

Constraints: Use only direct observations. Do not infer off-screen events. End with: [End of Step 1].

Step 2: Grounded QA Generation Prompt (Controlled Synthesis)

User Prompt:

Using the Step 1 factual analysis, generate exactly 10 *question–answer (QA) pairs* grounded exclusively in those observations.

Rules:

1. Produce 2 QA pairs from each of 5 categories:
2. Each QA must be directly verifiable from Step 1.
3. Counterfactuals must explicitly correct false premises using factual details.
4. Answers must be concise, literal, and non-redundant.

Output Format: Present results in a Markdown table. End with: [End of Step 2].

Figure 6. Two-stage prompting pipeline used to generate structured and visually grounded *QA* pairs for the *UDVideoQA* dataset.

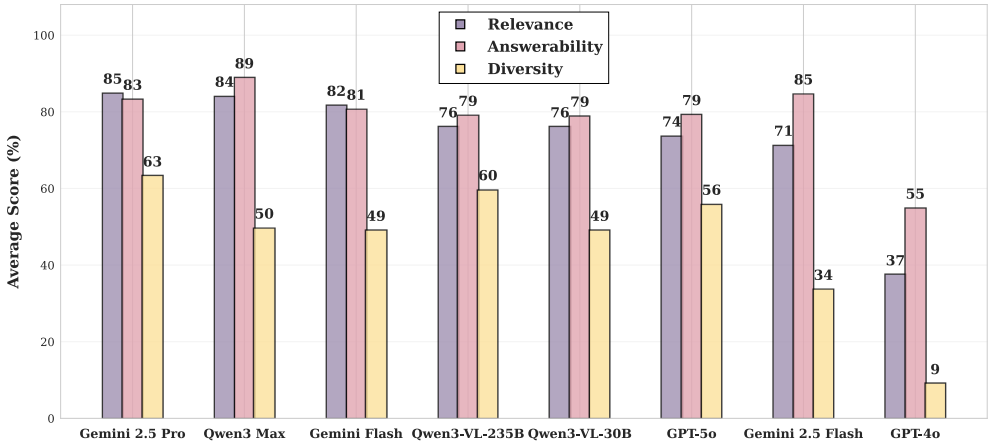


Figure 7. Overall performance of *VideoQGen* models across core metrics - *relevance*, *answerability*, and *diversity*. *Gemini 2.5 Pro* and *Qwen3 Max* achieve the highest scores, while *GPT-4o* performs consistently lower and generates less diverse questions.

Model	Questions		
	Low Intensity Light	Medium Intensity Light	High Intensity Light
Gemini 2.5 Pro	Are there any discernible architectural features on the buildings flanking the street?	What was the pedestrian doing as the dark car was approaching the crosswalk to make its turn?	Before the pedestrian in the dark shirt begins crossing the street at 0:07, what is the state of the intersection?
Qwen3 Max	Are there any temporary construction barriers or signs present on the sidewalk near the right edge of the frame?	If the traffic light for the horizontal street had remained red, would the black sedan have entered the intersection at 0.6 seconds?	If the white van had begun to turn right at 0 : 02, which pedestrian group would it have encountered?
Gemini Flash	Counting all visible vehicles, both moving and stationary, how many are white?	Is it true that a cyclist uses the green-painted bike lane at any point in the video?	A person in a white shirt and dark pants enters from the bottom left. What path do they take?
Qwen3-VL-23B	Was there any person standing on the sidewalk near the orange traffic signs on the bottom right?	If the silver sedan had not entered the intersection at 0.3s, would the white autonomous vehicle have been blocked from turning left?	If the traffic light for the crosswalk had turned red at 5 seconds, would the pedestrians have been in the middle of the street?
Gemini 2.5 Flash	Is it true that any pedestrian attempted to cross the street without using a designated crosswalk?	If the pedestrian at the bottom right had decided to cross at 0:03 , would they have entered the silver car's path?	How many silver/grey passenger cars are moving away from the camera in the upper lanes?
Qwen3-VL-30B	Is the traffic light for the main road green?	What is the color of the traffic light for vehicles going straight through the intersection?	Would the silver car have been able to pass the white van if the van had not stopped?
GPT-5	How many vehicles are visible between 0 and 2 s?	If the signal had stayed red for the entire clip, would any vehicle have entered the intersection?	Is the weather condition clear and dry?
GPT-4o	Is the scene recorded during daylight or nighttime?	Would the vehicle have proceeded if the pedestrian hadn't entered the crosswalk?	How many pedestrians are visible during the clip?

Table 2. Representative examples of generated questions across lighting conditions (low, medium, and high intensity). *Gemini 2.5 Pro* achieves the highest reasoning quality, while *GPT-4o* produces repetitive, low-diversity questions.

Qualitative Insights. Table 2 presents qualitative samples across models and lighting conditions, illustrating clear differences in reasoning depth. High-performing models such as *Gemini 2.5 Pro* and *Qwen3 Max* (green cells) consistently formulate multi-step reasoning questions, e.g., *"Before the pedestrian begins crossing, what was the state of the intersection?" (reverse reasoning)* or *"If the traffic light had remained red, would the black sedan have entered?" (counterfactual reasoning).* They also generate temporally anchored queries such as *"What was the pedestrian doing as the dark car approached the crosswalk?"*, capturing fine-grained agent interactions even under low-light conditions.

In contrast, weaker models like *GPT-4o* and *GPT-5* (red cells) default to simplistic or generic queries, such as weather descriptions or basic counting (*"Is the weather clear and dry?"*), and (*"How many pedestrians are visible?"*). These outputs reflect a failure to engage with complex, multi-agent dynamics, precisely the reasoning dimension the benchmark is designed to evaluate. Together, these findings confirm that *VideoQGen* effectively differentiates between models that exhibit deep spatio-temporal reasoning and those limited to surface-level perception. Structured prompting enables models to produce semantically rich, logically grounded questions, but genuine linguistic diversity remains a bottleneck. The results underscore that while modern *VLMs* are capable of generating high-quality questions when guided precisely. The figures presented below show a few examples of *VideoQGen* (see Figure 16 and Figure 17 for night and day, respectively).

5. The *UDVideoQA* Annotation Tool

The *UDVideoQA* is a comprehensive, semi-automated application designed to streamline the end-to-end creation of the *VideoQA* dataset. It consolidates the entire workflow from raw video ingestion and anonymization to *QA* validation and model-ready export within a single, unified framework (as in Figure 9). Developed in Python using the `Tkinter` framework and the `ttkbootstrap` library, the tool features a modern, responsive interface optimized for long annotation sessions and cross-platform deployment. The architecture ensures efficient dataset curation while

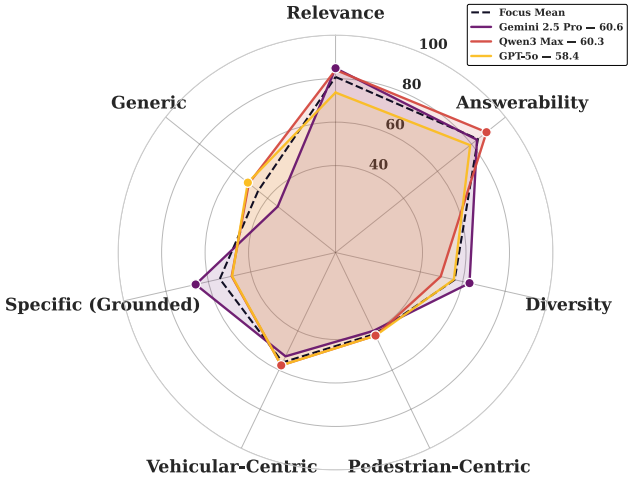


Figure 8. Distinct performance profiles of *VideoQGen* models across evaluation metrics. *Gemini 2.5 Pro* excels in *specific (grounded)* and *relevance*, while *Qwen3 Max* leads in *answerability*. The *focus mean* serves as a baseline for comparison.

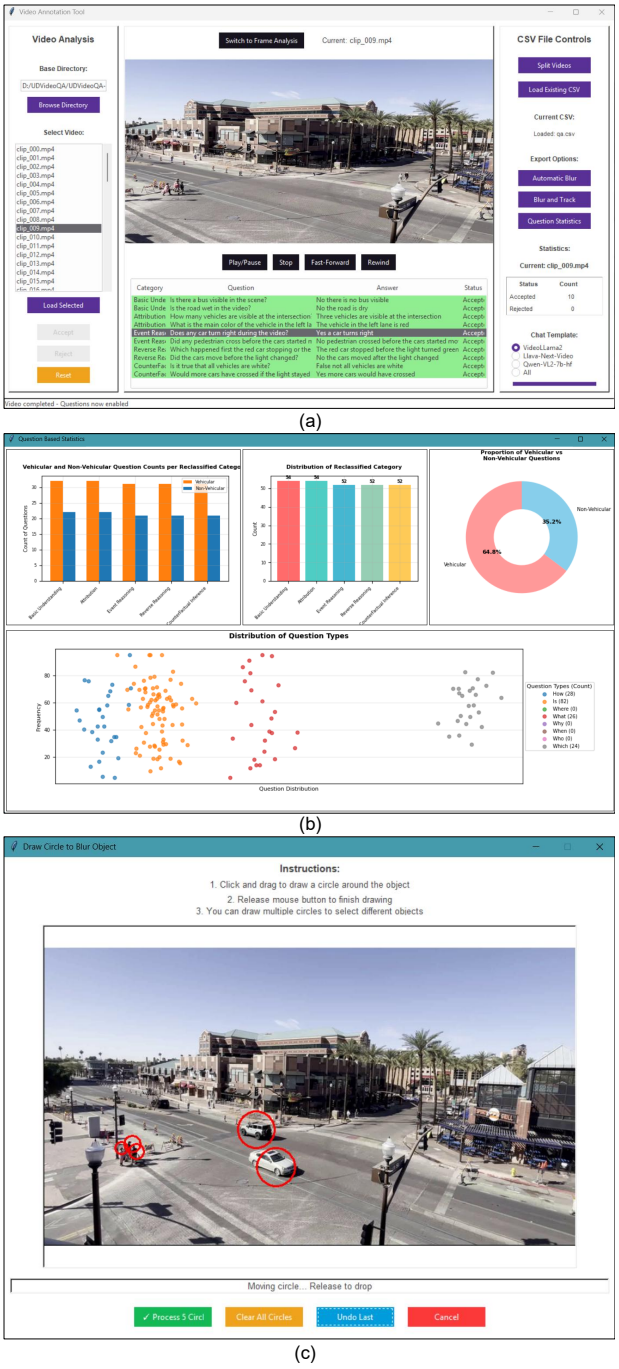
maintaining strict ethical and quality-control standards, as it follows a tri-panel design philosophy for intuitive workflow management: ***(1)*** the *data navigation panel* manages input sources, ***(2)*** the *primary annotation workspace* provides interactive playback and *QA* validation, and ***(3)*** the *processing and export panel* handles anonymization, batch processing, and data serialization. This logical structure guides annotators through clearly defined stages of data ingestion, validation, and export.

5.1. Data Ingestion and Pre-processing

Handling raw, long-form surveillance footage poses significant challenges for annotation scalability. To address this, the tool integrates an automated video segmentation utility that converts source videos into standardized 10s clips. The user specifies input and output directories, and the tool processes all major video formats (e.g., `.mp4`, `.avi`, `.mov`, `.mkv`), preserving resolution and codec fidelity. Each clip is automatically named sequentially (e.g., `clip_001.mp4`) to maintain dataset organization.

Once segmented, clips can be loaded directly into the *primary annotation workspace*, which supports two analysis modes: *video playback mode* for temporal reasoning and causal interpretation, and *frame analysis mode* for fine-grained spatial inspection. Playback controls (play/pause, stop, frame-step, and *fps-aware* 10s skips) ensure efficient navigation. To enforce contextual awareness, the tool automatically locks the *QA* validation list until the entire clip is reviewed, prompting annotators to observe the full temporal context before annotation.

The initial workflow stage addresses the challenge of handling raw, long-form surveillance footage. The tool integrates an automated segmentation utility that allows an annotator to select a source directory and an output directory. The system then processes all supported video formats (e.g., `.mp4`, `.avi`, `.mov`, `.mkv`) and segments them



(a)

(b)

(c)

Figure 9. User interface of the *UDVideoQA* Annotation Tool showing (top) the primary manual annotation workspace, (middle) real-time statistics for *QA* pairs, and (bottom) fine-grained blur tagging using YOLOv11 and SAM2.

into standardized 10-second clips, preserving the original resolution and codec while sequentially naming the output (e.g., `clip_001.mp4`). Once segmented, clips are loaded into the primary annotation workspace. This module supports two distinct analysis modes: a video playback mode for assessing temporal context and causal reasoning, and a *frame analysis mode* for fine-grained, frame-by-frame spatial inspection. Standard media controls (play/pause, stop, and FPS-aware 10-second skips) are provided. A crucial quality control gate is implemented: upon loading a new video, the *QA* validation list remains locked, and a status message indicates "Auto-playing video", guaranteeing that the annotator reviews the entire visual context at least once before beginning annotation.

5.2. Human-in-the-Loop Anonymization Workflow

A key innovation of *UDVideoQA* tool is its two-stage *hybrid anonymization pipeline*, which integrates automated motion-based blurring with AI-assisted manual refinement to ensure both efficiency and ethics are maintained.

Stage 1: Automated Motion-Based Blurring. In the first stage, the *IEBCS*-based motion-blur module anonymizes all dynamic agents in a single operation. Operating on pixel-intensity changes, this event-driven method effectively protects moving subjects while avoiding the semantic misclassifications typical of detector-based approaches. It serves as a high-recall, low-latency first pass for anonymization.

Stage 2: Manual AI-Assisted Track-and-Blur. For fine-grained correction, the *Manual Track-and-Blur* feature combines *YOLOv11* for detection and the *SAM2* for pixel-accurate masking. Annotators can draw a simple region of interest around a target object, prompting the system to generate an initial *SAM* mask and predict its motion trajectory using velocity vectors. A *"predict–update"* cycle then maintains smooth tracking: the predicted position is updated each frame via *YOLO* detections, with masks regenerated only when object shape or trajectory deviates significantly. This approach maintains high temporal precision while minimizing computational overhead, even in cases of occlusion or low light.

5.3. QA Validation and Refinement Module

Beyond anonymization, *UDVideoQA* tool facilitates rigorous manual validation and refinement of the *QA* pairs generated by large language models. *QA* pairs are imported from *CSV* format and displayed in an interactive `ttk.Treeview` table with columns for category, question, answer, and validation *status* (*none*, *accepted*, *rejected*). The validation interface supports keyboard shortcuts (`Enter` to accept, `Backspace` to reject) for rapid annotation throughput. Rejected questions are not discarded outright. Instead, a double-click opens an editable dialog where annotators can modify the question, answer, or category. Upon correction, the record automatically updates to *"accepted"*, turning the validation stage into an active refinement loop. Annotators may save incremental progress for batches of videos before export, ensuring recovery and continuity.

VideoQA Prompt

```
You are a Video Traffic Analyst. Answer the
question using only the provided video.
Observe carefully: focus on subtle details,
timing, misleading cues, and missing elements.
Describe strictly what is visible---no inference,
guessing, or use of external knowledge.
If the answer cannot be visually verified, state
``Cannot be determined from the video.''
If a question assumes something false, correct it
using observed facts.
Provide clear, factual answers grounded solely in
visual evidence.
```

Figure 10. Prompt template used for *VideoQA* model evaluation, instructing models to generate factual, visually grounded answers based solely on the provided video content.

5.4. Integrated Analytics Dashboard

To aid quality monitoring, *UDVideoQA* tool features an integrated *analytics dashboard* that provides live statistics on annotation progress, *QA* category distribution, and validation ratios. The dashboard supports categorical verification, hallucination detection, and timestamp consistency checks, aligning with the *human-in-the-loop* refinement shown in the main pipeline (see Figure 9). This feedback loop ensures annotation consistency and enables transparent reporting of validation coverage during dataset curation.

5.5. Model-Ready Benchmark Export

The final stage of the workflow serializes validated *QA* data into standardized, model-compatible formats. Only entries marked as *accepted* are exported, while rejected or pending items remain archived for traceability. The export module automatically generates *JSONL* chat templates optimized for major *VideoLM* frameworks, including *VideoLLaMA*, *LLaVA-NeXT-Video*, and *Qwen-VL*. Built-in profiles ensure proper formatting of conversation turns and metadata fields (e.g., *timestamps, reasoning type*). The export process includes robust quality checks for missing values and generates timestamped *CSV* backups to preserve annotation states before serialization. This standardized export pipeline allows seamless benchmarking and reproducibility across different multimodal architectures.

Scoring Policy for the LLM Judge as Prompt for Evaluation

Role: You are an impartial, strict grader for *VideoQA* focused on semantic-semantic understanding.

Input Format:

- CSV columns: `index`, `video_file_path`, `question`, `category`, `answer`, `model_answer`.
- Read the CSV
- Normalize Unicode and smart quotes so text compares correctly.

Derived Fields (per row):

- ***clip_id:*** `video_file_path` without extension.
- ***q_idx:*** Question index within each `clip_id`, in pasted order, starting at 1.
- ***subtype:*** Determined from the lowercased question text:
  - ***pedestrian:*** if mentions `pedestrian, pedestrians, person, people, man, woman, child, cyclist, bicyclist, walker`.
  - ***vehicular:*** if mentions `car, cars, vehicle, vehicles, truck, bus, van, sedan, suv, motorcycle, lane, traffic, driving, overtake, approach, foreground`.
  - otherwise, ***background***.
- ***category:*** Use given text (trim whitespace). Canonicalize close variants (e.g., `CounterFActual Inference`) to one of: *Basic Understanding, Attribution, Event Reasoning, Reverse Reasoning, Counterfactual Inference.*
- ***ground_truth:*** Complete answer
- ***model_answer:*** Trimmed text (empty if blank or NA).

Grading Logic:

- ***correct*** $\in \{1, 0\}$ (no partial credit).
- ***Binary/categorical:*** accept exact or unambiguous semantic equivalence (case-insensitive; e.g., "yes." = "yes").
- ***short textual:*** accept meaning-identical paraphrases; if missing key info or contradicts ground truth $\rightarrow 0$.
- ***If*** `model_answer` is blank, "i don't know," "unsure," or contradictory $\rightarrow 0$.
- Do not infer beyond what is written in the ground truth.

Weights by Category:

Category	Weight
Basic Understanding	1.0
Attribution	1.2
Event Reasoning	1.3
Reverse Reasoning	1.3
Counterfactual Inference	1.5

Scoring:

$$\text{points} = \text{correct} \times \text{weight}$$

Figure 11. Grading prompt used for *VideoQA* benchmark evaluation. The LLM-based grader enforces strict, non-inferential correctness through canonicalized categories, subtype detection, and category-weighted scoring.

6. Experiments

This section describes our experimental protocol, evaluation methodology, quantitative benchmark results, and qualitative error analysis on the *UDVideoQA* dataset. Our goals are ***(i)*** to measure zero-shot reasoning capabilities of SOTA multimodal models, ***(ii)*** to evaluate gains achievable via parameter-efficient fine-tuning, and ***(iii)*** to diagnose systematic failure modes that limit current *VideoQA* performance.

6.1. Experimental Setup

Zero-shot Evaluation. All candidate models are first tested directly on the *UDVideoQA* test set to measure out-

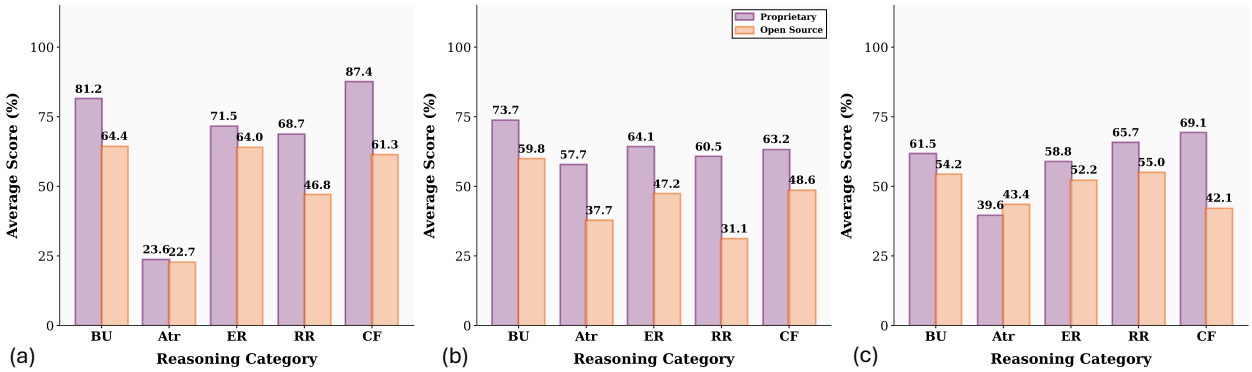


Figure 12. Average *VideoQA* performance of proprietary and open-source models across five reasoning categories under different lighting conditions: (a) day, (b) afternoon, and (c) nighttime scenarios.

of-the-box generalization and reasoning capabilities without any domain-specific tuning. Each model was given the same *VideoQA* prompt (see Figure 10) to ensure consistency

Fine-tuning (parameter-efficient). Selected open-source models were fine-tuned on the *UDVideoQA* training split to quantify adaptation benefits. Fine-tuning configurations include:

- ***InternVL-38B:*** LoRA adapters with rank $r = 16$ and $\alpha = 32$ were applied. The core LLM, vision backbone, and MLP layers were frozen (`--freeze_llm`, `--freeze_backbone`, and `--freeze_mlp`) so only adapter parameters were trained. Jobs were run via SLURM (`srun`) or `torchrun` on multi-GPU nodes.
- ***Qwen-VL (32B):*** LoRA with rank $r = 64$, main components frozen and only LoRA adapters trained for efficiency.
- ***Qwen-VL (7B):*** The vision tower was unfrozen and trained (no LoRA) to better adapt the small model's visual encoder. Training used 4×A100 GPUs.

All fine-tuning runs used standard data augmentation and early stopping tuned on a held-out validation subset. Hyperparameter sweeps and compute budgets were kept consistent across models to ensure fair comparisons.

6.2. Evaluation Methodology

To avoid brittle lexical matches and better assess semantic correctness and reasoning complexity, we adopt a multi-stage, hybrid evaluation:

LLM-based Semantic Scoring (LLM Judge). We use an external LLM judge (Gemini 2.5 Pro) driven by the *Strict Grading System Prompt* (see Figure 11) to perform semantic–semantic matching. The judge assigns a binary score (correct $\in \{0, 1\}$) after normalizing text, canonicalizing categories, and assessing semantic equivalence rather than raw string overlap. The grader rejects blank, non-committal, or contradictory responses [6].

Weighted Complexity-based Scoring. To reflect varying cognitive load across question types, we weight categories by difficulty: *basic understanding* (1.0), *attribution* (1.2), *event reasoning* (1.3), *reverse reasoning* (1.3), and *counterfactual inference* (1.5). The final per-question points equal the binary correctness times the category weight, and model scores are aggregated over the evaluation set.

Human validation. To calibrate and control for potential *LLM* judge biases, we randomly sample a *"gold-standard"* subset (1K predictions) that is independently annotated by domain experts. An agreement between the human labels and the *LLM* judge is reported. Disagreement cases were inspected and used to refine the grading prompt and normalization heuristics.

6.3. Quantitative Results

Table 3 and Figure 12 summarize results across models, reasoning categories, and day/night splits. Figure 13 shows performance across the top 4 performing models. Some key observations are:

Proprietary vs. Open-source. Proprietary models achieve the strongest zero-shot performance. Notably, *Gemini 2.5 Pro* attains the top scores across multiple time windows (morning: 75.78%, evening: 71.13%) and leads in pedestrian-centric (87.90%) and vehicular-centric (68.59%) settings. *Gemini 2.5 Flash* performs best in the afternoon (74.98%).

Fine-tuned Models Competitiveness. Parameter-efficient fine-tuning yields substantial gains. The fine-tuned *Qwen*

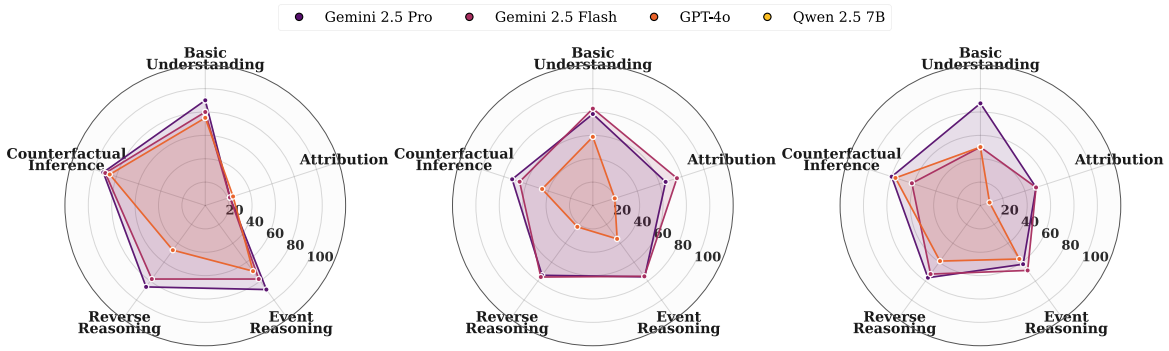


Figure 13. Comparison of four leading multimodal models (*Gemini 2.5 Pro*, *Gemini 2.5 Flash*, *GPT-4o*, and *Qwen 2.5 7B*) across five *VideoQA* reasoning dimensions, highlighting variations in perceptual and causal understanding.

Model Type	Model Name	Pedestrian-centric						Vehicular-centric					
		BU	Atr	ER	RR	CI	Overall	BU	Atr	ER	RR	CI	Overall
Proprietary	**Gemini 2.5 Pro**	93.43	83.03	73.32	76.39	92.81	87.90	66.67	33.84	69.96	77.10	84.72	68.59
	Gemini 2.5 Flash	68.93	87.27	72.12	78.12	94.77	80.21	71.11	38.62	72.46	73.68	78.60	67.93
	GPT-5	68.93	80.61	72.12	76.39	92.16	78.01	71.11	39.49	55.13	59.75	82.10	60.21
	GPT-4o	59.42	25.45	56.49	28.91	83.01	55.29	42.22	7.32	47.25	44.58	71.53	44.56
Open Source	**Qwen 2.5 32B**	87.90	81.82	59.38	41.38	86.93	80.07	53.33	63.48	47.95	54.49	65.07	56.74
	Qwen 2.5 7B	78.10	79.39	65.63	46.55	90.85	77.85	53.33	65.65	53.63	50.46	48.91	53.87
	VideoLLaMA3	86.74	16.97	64.06	37.93	86.27	65.86	46.67	59.13	50.16	47.37	71.62	55.85
	NVILA 8B	83.57	16.97	60.94	29.31	86.19	63.68	57.78	56.81	64.01	52.06	46.72	55.18
	Llava-NeXT-Video 7B	69.45	7.88	37.50	20.69	24.84	38.62	37.78	8.70	43.22	21.67	9.61	22.93
	InternVL 38B	87.83	28.66	62.14	39.66	89.47	69.47	37.78	50.58	47.62	50.47	77.09	55.42

Table 3. Comparison of model performance across *pedestrian-centric* and *vehicular-centric* scenarios. Color intensity indicates relative accuracy within each reasoning subcategory.

2.5 VL-7B and *Qwen 2.5 VL-32B* show meaningful improvements in evening conditions (e.g., *Qwen-7B*: 61.38%), closing the gap to proprietary models on several categories. Some architectures (e.g., *Llava-NeXT-Video 7B*) still struggle at fine-grained perception (very low attribution in some lighting conditions), indicating architecture- or training-data-induced limits.

Failure of Fine-grained Attribute Localization. Across models, we observe systematic attribution errors on fine-grained visual items (e.g., road markings). Example: when asked, *"What large marking is painted in the central lane near the camera?"* ground truth: *"A white 'X' road marking"*, model predictions ranged widely (*"diamond"*, *"bicycle symbol"*, *"white arrow"*), evidencing a decoupling between scene-level reasoning and low-level visual recognition.

6.4. Qualitative Insights and Error analysis

Qualitative study of model outputs reveals consistent patterns that quantitative metrics alone do not capture(see Table 4, Table 5 and Table 6). Two recurring phenomena stand out: ***(1)*** an *inference–attribution scores*, where models succeed at high-level inference but fail at low-level visual grounding, and ***(2)*** *systematic temporal and premise reasoning failures*. Below, we summarize these failure modes, illustrate them with examples from our test splits, and discuss their implications.

Inference–Attribution Scores Models frequently demonstrate strong inferential ability, correctly proposing plausible outcomes or causal links while simultaneously failing to ground those inferences in the visual evidence. This produces seemingly good *counterfactual* or *event-reasoning* scores even as the underlying visual attribution is incorrect.

Example (Counterfactual / Hallucination):

- ***Question:*** *"If the van had not yielded, what might have occurred?"*
- ***Ground Truth:*** *"It might have collided with the oncoming car."*
- ***Prediction:*** *"The impact could have resulted in severe injuries for both the pedestrian and the van's occupants."*

Analysis. The model correctly predicts a collision (matching the ground truth's high-level outcome) but hallucinates

Model	Gemini 2.5 Pro	Gemini 2.5 Flash	GPT-4o	GPT-5	Qwen 2.5 VL - 7B (Fine Tuned)	Qwen 2.5 VL - 7B	InternVL3-38B	Qwen 2.5 VL 32B	VideoLlama3	NVILA 8B	Llava-NeXT-Video
Question	Is there any overtaking observed?										
Ground Truth	No										
Best Answer	No overtaking is observed as there are no moving vehicles	No	No	No	No overtaking is observed in the image.	No, there is no overtaking observed [...]. The road appears to be empty with no vehicles in motion [...]	No, there is no overtaking observed.	No, there is no overtaking observed [...]. The scene shows an empty road with no vehicles or pedestrians present. [...]	The video does not show any overtaking.	No overtaking is observed in the image.	Based on the image provided, it appears that there is no overtaking place. [...]
Question	How many cars exited the foreground?										
Ground Truth	Four										
Worst Answer	Zero cars exited the foreground as the car moved away from the camera.	One car exited the fore-ground	Two	0	One car exited the fore-ground in the image.	[...] there is one car that appears to have just exited the foreground on the right side [...]	One car exited the foreground	[...] there is one car that exits the foreground. It can be seen moving from the right side [...] eventually exiting the frame.	[...] it appears that one car has exited the foreground. The road is mostly empty [...]	There is one car that exited the foreground.	There are three cars in the foreground of the image.

Table 4. Representative examples of best and worst model-generated answers under the *daytime* scenario.

Model	Gemini 2.5 Pro	Gemini 2.5 Flash	GPT-4o	GPT-5	Qwen 2.5 VL - 7B (Fine Tuned)	Qwen 2.5 VL - 7B	InternVL3-38B	Qwen 2.5 VL 32B	VideoLlama3	NVILA 8B	Llava-NeXT-Video
Question	Is the road surface wet from rain?										
Ground Truth	No the road looks dry										
Best Answer	No, the road surface is not wet from rain; it appears completely dry.	No, the road surface appears dry, not wet from rain.	No	No	No, the road surfaceis not wet from rain. The road appears dry and clear of any water.	No, there is no wetness from rain observed [...]. The road appears to be dry with no visible signs of water [...]	No, the road surface appears dry.	No, there is no wetness from rain observed [...]. The road surface appears to be dry with no visible signs of water or reflections [...]	The road surface appears dry, with no visible signs of rain or wetness.	No, the road surface is not wet from rain.	No, there is no wetness from rain observed [...]. The road surface appears to be dry and clear with it not raining at the time [...]
Question	What large marking is painted in the central lane near the camera?										
Ground Truth	A white 'X' road marking										
Worst Answer	A white bicycle symbol.	A large diamond shape (?) is painted in the central lane.	Straight white arrow	A large white marking is painted in the central lane.	The large marking painted in the central lane near the camera is a white arrow pointing to the right	The large marking appears to be a white arrow pointing to the right with it indicating a turn lane [...]	The large marking painted in the central lane near the camera is a 'no parking' sign	The large marking appears to be a "Bus Only" or "HOV" lane marking with this being indicated by the red color [...]	The large marking appears to be a white arrow pointing to the right with it indicating a lane change or turn direction [...]	The large marking painted in the central lane near the camera is a white arrow.	The large marking painted in the central lane near the camera is a "No Parking" sign.

Table 5. Representative examples of best and worst model-generated answers under the *afternoon* scenario.

an absent pedestrian and invents injury details. In many cases, the model relies on a generic *"traffic accident"* script learned from language priors rather than verifying whether a pedestrian is actually present in the clip. This behavior inflates performance on inference-style tasks while undermining factual fidelity.

Temporal and Premise Reasoning Failures. A second common failure mode concerns temporal localization and premise validation. Models often fail to respect explicit temporal windows or to check the truth of the question's premise against the video evidence.

Example (Temporal / Premise Error on Evening Set):

- ***Question:*** *"Which happened first: the first car entering the foreground or the second one?"*
- ***Ground Truth:*** *"No car enters the foreground."*
- ***Prediction (multiple models):*** *"The first car..."*

Analysis. Here, models assert an order of events despite the correct answer indicating that no such event occurred within the queried interval. This indicates failures in ***(a)*** isolating the exact temporal window referenced by the question, and ***(b)*** verifying whether the question's premise holds before attempting a comparative judgment.

Fine-grained Perception Failure. Notably, attribution-style questions that require precise localization or recognition of small, high-frequency visual primitives (e.g., *road markings, signage, small symbols*) are among the weakest for all evaluated models.

Example (Attribution Failure on Morning Set):

- ***Question:*** *"What large marking is painted in the central lane near the camera?"*
- ***Ground Truth:*** *"A white 'X' road marking."*
- ***Prediction:****Qwen-7B: "A white arrow pointing to the right." VideoLLaMA 3: "A yellow diamond shape."*, and *Gemini 2.5 Pro: "A white bicycle symbol."*

Analysis. Despite being a visually local and ostensibly simple task, models produce a variety of incorrect but plausible labels. This demonstrates a persistent gap between ***(i)*** scene-level, multi-agent reasoning and ***(ii)*** reliable low-level visual recognition. In practice, this gap means that models may reach correct conclusions for the wrong perceptual reasons or fail important downstream decisions because they cannot consistently perceive simple visual cues.

Implications and Directions. Together, these qualitative failure modes suggest two complementary research priorities for *VideoQA* and multimodal reasoning:

- ***Tighter Visual–Verbal Grounding:*** Improve models' ability to verify perceptual premises (*"is X present?"*), and to condition any inference on an explicit check of visual evidence rather than on language priors alone.

Model	Gemini 2.5 Pro	Gemini 2.5 Flash	GPT-4o	GPT-5	Qwen 2.5 VL - 7B (Fine Tuned)	Qwen 2.5 VL - 7B	InternVL3-38B	Qwen 2.5 VL 32B	VideoLlama3	NVILA 8B	Llava-NeXT-Video
Question	Is the road surface wet from rain?										
Ground Truth	No the road looks dry										
Best Answer	No, the road surface is not wet from rain; it appears completely dry.	No, the road surface appears dry, not wet from rain.	No	No	No, the road surfaceis not wet from rain. The road appears dry and clear of any water.	No, there is no wetness from rain observed [...]. The road appears to be dry with no visible signs of water [...]	No, the road surface appears dry.	No, there is no wetness from rain observed [...]. The road surface appears to be dry with no visible signs of water or reflections [...]	The road surface appears dry, with no visible signs of rain or wetness.	No, the road surface is not wet from rain.	No, there is no wetness from rain observed [...]. The road surface appears to be dry and clear with it not raining at the time [...]
Question	What large marking is painted in the central lane near the camera?										
Ground Truth	A white 'X' road marking										
Worst Answer	A white bicycle symbol.	A large diamond shape (?) is painted in the central lane.	Straight white arrow	A large white marking is painted in the central lane.	The large marking painted in the central lane near the camera is a white arrow pointing to the right	The large marking appears to be a white arrow pointing to the right with it indicating a turn lane [...]	The large marking painted in the central lane near the camera is a 'no parking' sign	The large marking appears to be a "Bus Only" or "HOV" lane marking with this being indicated by the red color [...]	The large marking appears to be a white arrow pointing to the right with it indicating a lane change or turn direction [...]	The large marking painted in the central lane near the camera is a white arrow.	The large marking painted in the central lane near the camera is a "No Parking" sign.

Table 6. Representative examples of best and worst model-generated answers under the *evening* scenario

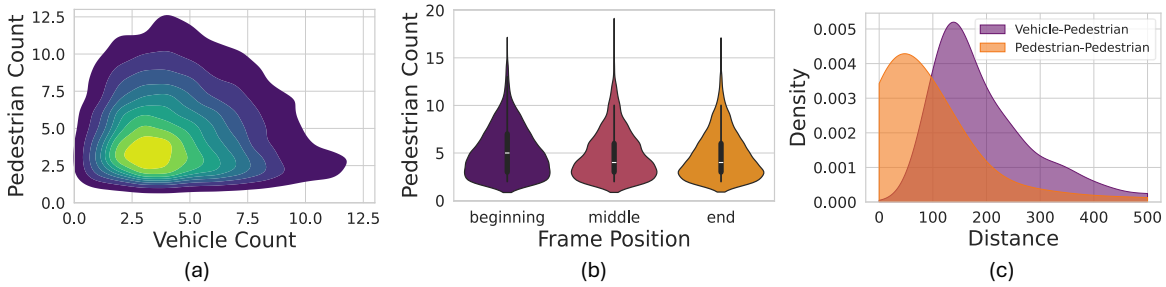


Figure 14. Analysis of multi-agent dynamics: (a) most frequent multi-agent scenario observed in the dataset, (b) temporal stability of pedestrian activity across video segments, and (c) proximity density plots highlighting frequent close interactions between vehicle–pedestrian and pedestrian–pedestrian pairs.

- ***Temporal Precision and Localized Perception:*** Enhance temporal alignment and small-object recognition through either improved video encoders, targeted supervision (e.g., *contrastive/object-centric objectives*), or hybrid symbolic checks that force premise validation before inference.

Addressing these issues will reduce hallucination, improve the fidelity of counterfactual and event-reasoning evaluations, and close the gap between *"reasoning"* and *"seeing"* that we observe across current SOTA models. Examples for *RR* are shown in Figure 18, *ER* in Figure 19, *RR* in Figure 20, *ER* in Figure 21, and *CF* in Figure 22, respectively.

6.5. Multi-Agent Behavioral Analysis

Traffic intersections represent highly dynamic environments where diverse agents-vehicles, pedestrians, and micro-mobility users-share limited space and interact under varying contextual constraints. We hypothesize that such intersections naturally promote frequent, observable multi-agent interactions that make them ideal for studying spatio-temporal reasoning. Our empirical analysis of the *UDVideoQA* dataset supports this hypothesis through three key observations: ***(1)*** strong agent co-occurrence, ***(2)*** temporal consistency of activity, and ***(3)*** measurable proximity-based interaction potential.

Agent Co-occurrence A fundamental prerequisite for studying interactions is the concurrent presence of multiple agents within the same frame. Figure 14(a) shows a density heatmap of vehicle and pedestrian counts across all annotated frames. The highest-density region (brightest yellow) lies not near the origin, representing empty or single-agent scenes, but around 2–5 *vehicles* and 2–4 *pedestrians*. This distribution confirms that the majority of footage captures scenes with concurrent multi-agent activity rather than sparse or unoccupied intersections. Such consistent co-occurrence validates the choice of intersections as natural laboratories for multi-agent reasoning, where spatial dependencies and interaction opportunities emerge continuously.

Temporal Consistency of Activity To verify that these interactions are temporally stable rather than transient, we analyzed pedestrian counts across three temporal segments of each 10s clip *beginning*, *middle*, and *end*. The violin plots in Figure 14(b) reveal near-identical distributions across

Figure 15. Example frame illustrating the model's use of visual evidence, specifically the "AP" license plate prefix-to support its prediction in the *RoadSocial* dataset.

these intervals: the median counts (white dots), interquartile ranges (black bars), and overall distribution shapes remain remarkably consistent. This indicates that the level of pedestrian activity and, by extension, the overall scene dynamics remain stable throughout the clips. The dataset, therefore, provides a balanced temporal sampling of urban traffic behavior rather than isolated or event-biased moments.

Quantifying Proximity and Interaction Potential Beyond co-occurrence, we measure physical proximity to estimate interaction potential. Figure 14 (c) shows pedestrian–pedestrian distances peaking near 50–75 units, while vehicle–pedestrian distances peak around 150 units, with a significant tail of low-distance events. These *high-interaction zones* confirm that intersections force multi-agent co-location via traffic constraints like lights and crosswalks. This proximity generates rich behaviors like yielding and conflict, demonstrating that intersections are *optimal real-world settings* for the complex, causally entangled reasoning *UDVideoQA* aims to benchmark.

6.6. Generalization Tests on available Datasets

To evaluate cross-domain transferability, we tested the fine-tuned *Qwen 2.5-VL-7B* model on 500 *QA* pairs each from the validation splits of *RoadSocial* [8] and *SUTDTrafficQA* [11]. These datasets were chosen for their structural similarity to *UDVideoQA* in terms of question format and reasoning taxonomy, allowing for a controlled assessment of generalization across different video sources and annotation protocols. The evaluation produced a few noteworthy observations (see Table 7 and Table 8). On the *RoadSocial* benchmark, which includes deliberately adversarial questions targeting perception robustness, the model demonstrated strong visual grounding even when its prediction diverged from the labeled ground truth. For example:

- **Question:** *"In which Indian state did this road event likely occur?"*

Training Configuration	QA-Group				Overall Score
	Factual	Complex	Imaginative	Adversarial	
Fine Tuned Model	68.2	41.8	74.4	55	**60**
Baseline	66.2	3	83.7	58.1	**52.75**

Table 7. Evaluation of the *UDVideoQA* fine-tuned model's generalization performance on the *RoadSocial* dataset across four *QA* categories - *factual*, *complex*, *imaginative*, and *adversarial*.

Training Configuration	QA-Group				Overall Score
	BU	Att	CI	ER	
Fine Tuned Model	54.5	73.3	75	48	**62.7**
Baseline	48.9	69.6	73.4	44	**59.05**

Table 8. Evaluation of the *UDVideoQA* fine-tuned model's generalization performance on the *SUTDTraffic* dataset across four *QA* categories *factual*, *complex*, *imaginative*, and *adversarial*.

- ***Ground Truth:*** *"The incident likely occurred in Telangana, India."*
- ***Model Prediction:*** *"The bus in the video has a license plate that starts with 'AP,' which stands for Andhra Pradesh... therefore, it is likely that this road event occurred in Andhra Pradesh."*

While the prediction (*"Andhra Pradesh"*) does not match the ground-truth label (*"Telangana"*), visual inspection confirmed that the bus's license plate indeed displayed the prefix *"AP"*. This outcome highlights the model's capacity for *visual evidence-based reasoning*, deriving logically coherent answers grounded in observed cues. Figure 15 illustrates the corresponding frame used in this reasoning process. Beyond individual examples, the fine-tuned *Qwen* model achieved consistent quantitative improvements across both benchmarks. Under the weighted semantic-scoring framework introduced in Section 6.2, it recorded a 7.3% gain on *RoadSocial* and a 3.61% gain on *SUTDTrafficQA* over its baseline. These improvements demonstrate that exposure to high-density, multi-agent intersection data in *UDVideoQA* enhances a model's ability to generalize causal and perceptual reasoning across unseen traffic domains. In summary, fine-tuning on *UDVideoQA* not only boosts in-domain accuracy but also strengthens cross-domain robustness, particularly in tasks requiring grounded visual interpretation and resistance to hallucination.

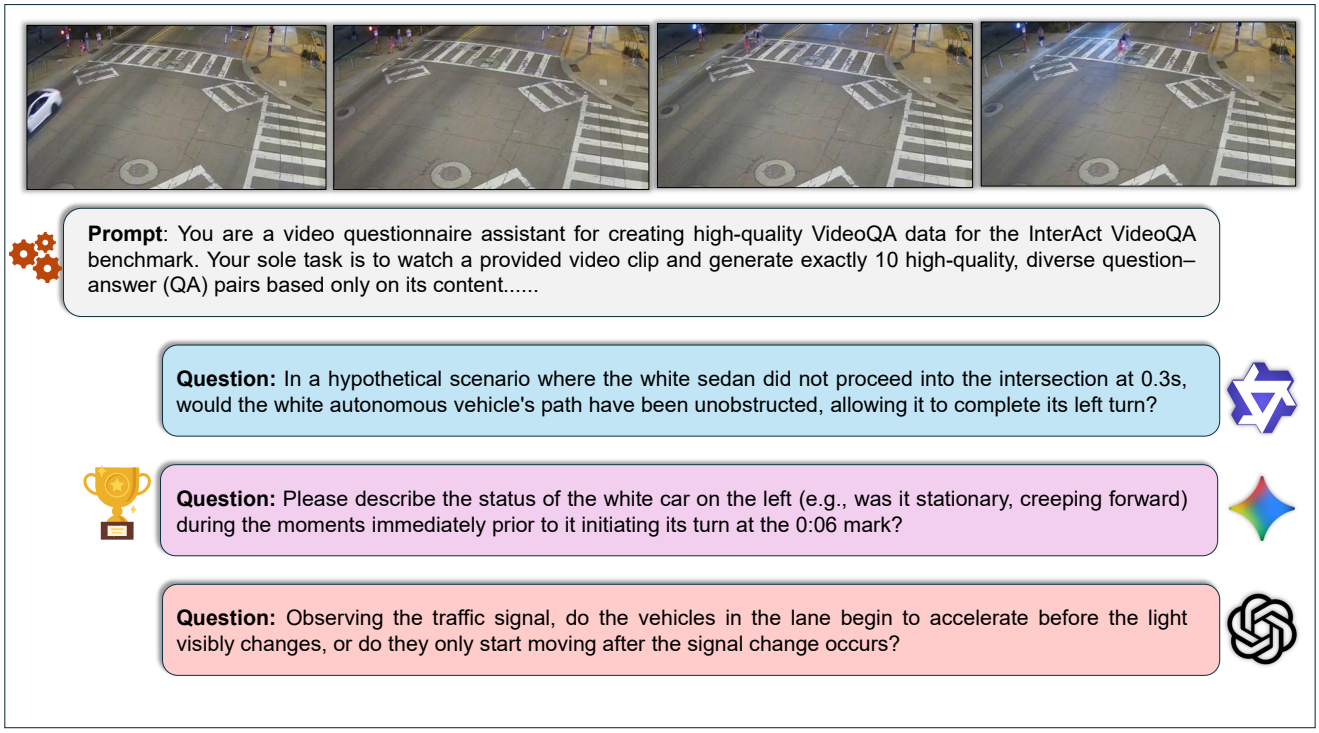


Figure 16. Example *VideoQGen* outputs generated for night-time clips. *Gemini 2.5 Pro* produces the most contextually grounded questions, followed closely by *Qwen 3*, while *GPT* models generate more generic question sets.

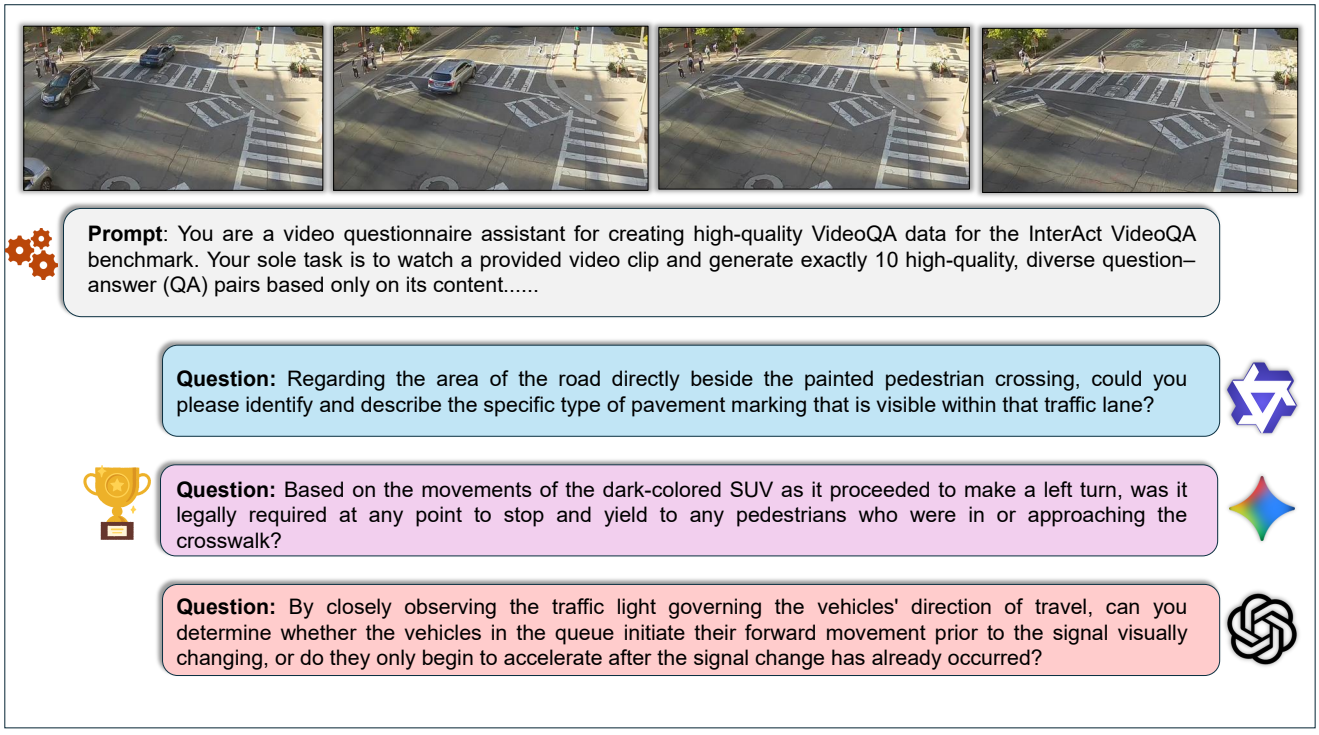


Figure 17. Example *VideoQGen* outputs generated for early hours of the day clips. *Gemini 2.5 Pro* produces the most accurate and contextually grounded questions, followed closely by *Qwen 3*, while *GPT* models generate more generic question sets.

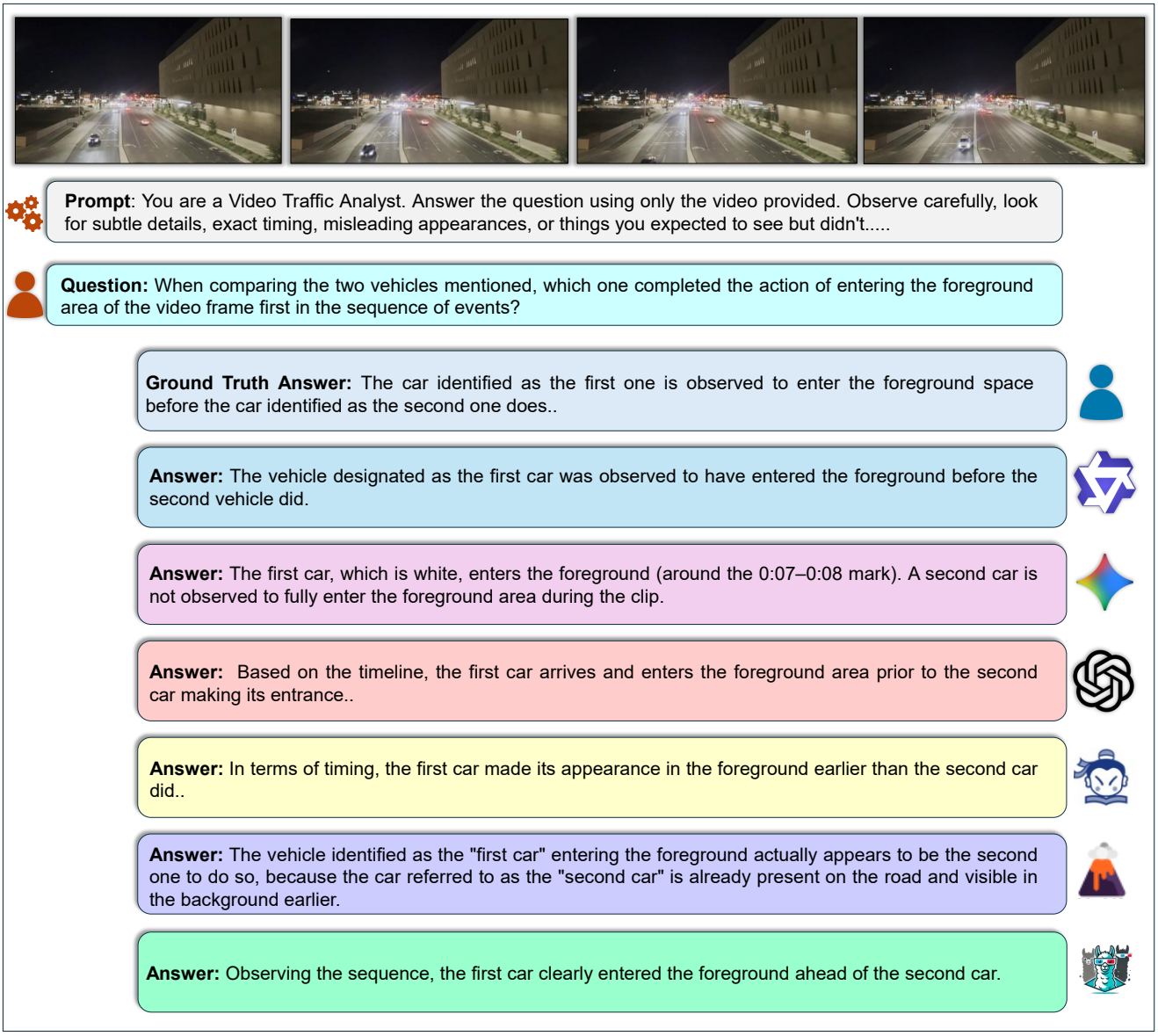


Figure 18. Example of a *reverse reasoning* question evaluating fine-grained temporal understanding. Models must track two distinct objects and determine their correct sequential order. *Gemini 2.5 Pro* provides the most accurate response by focusing on the temporal relationships within the sequence.

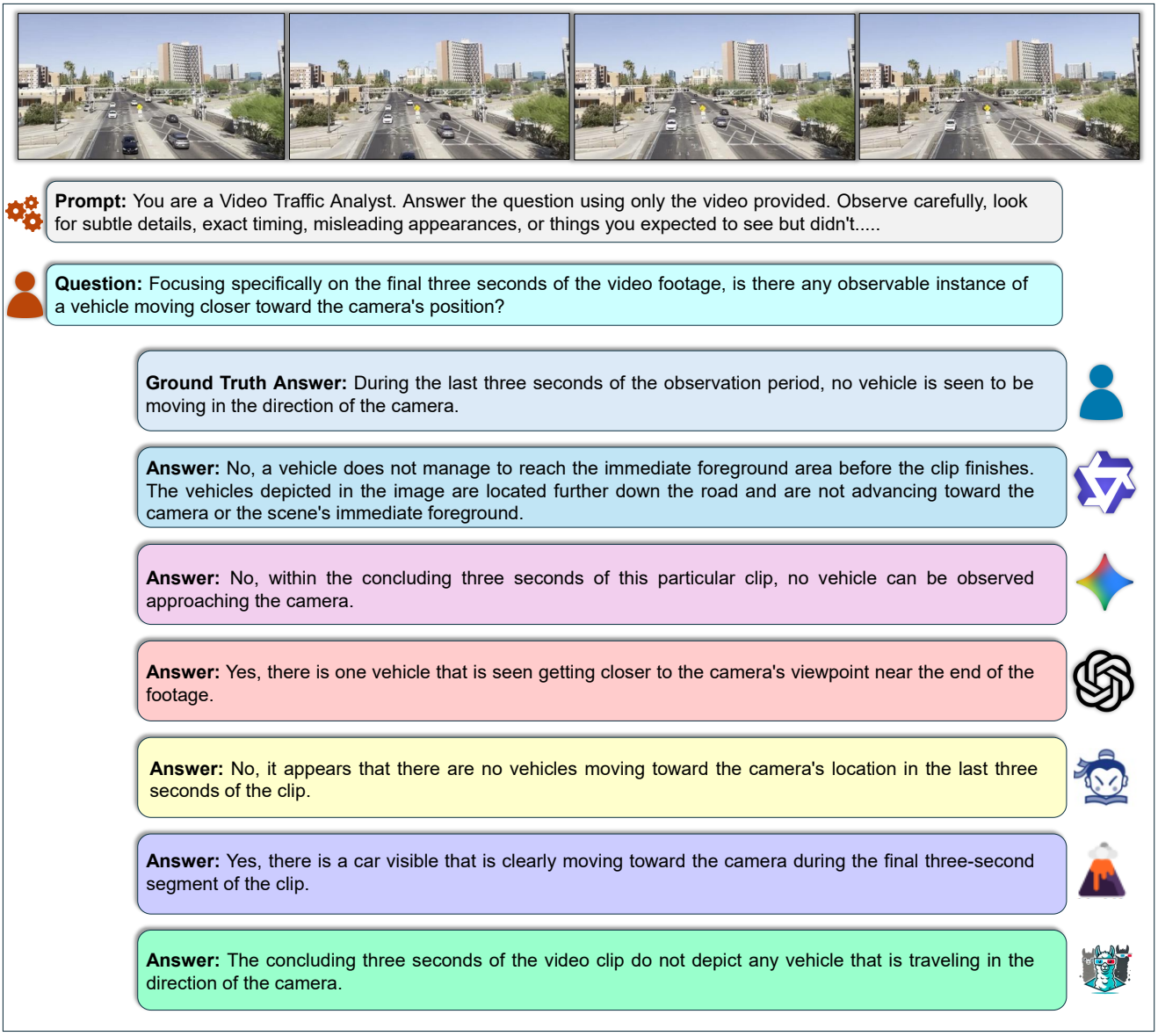


Figure 19. . Example of an *event reasoning* challenge illustrating model failures in temporal grounding. Although the ground truth indicates that no vehicle approaches the camera during the final three seconds, several models (e.g., *GPT* and *LLaVA-NeXT-Video*) incorrectly respond "Yes", revealing hallucinated motion and poor temporal localization.

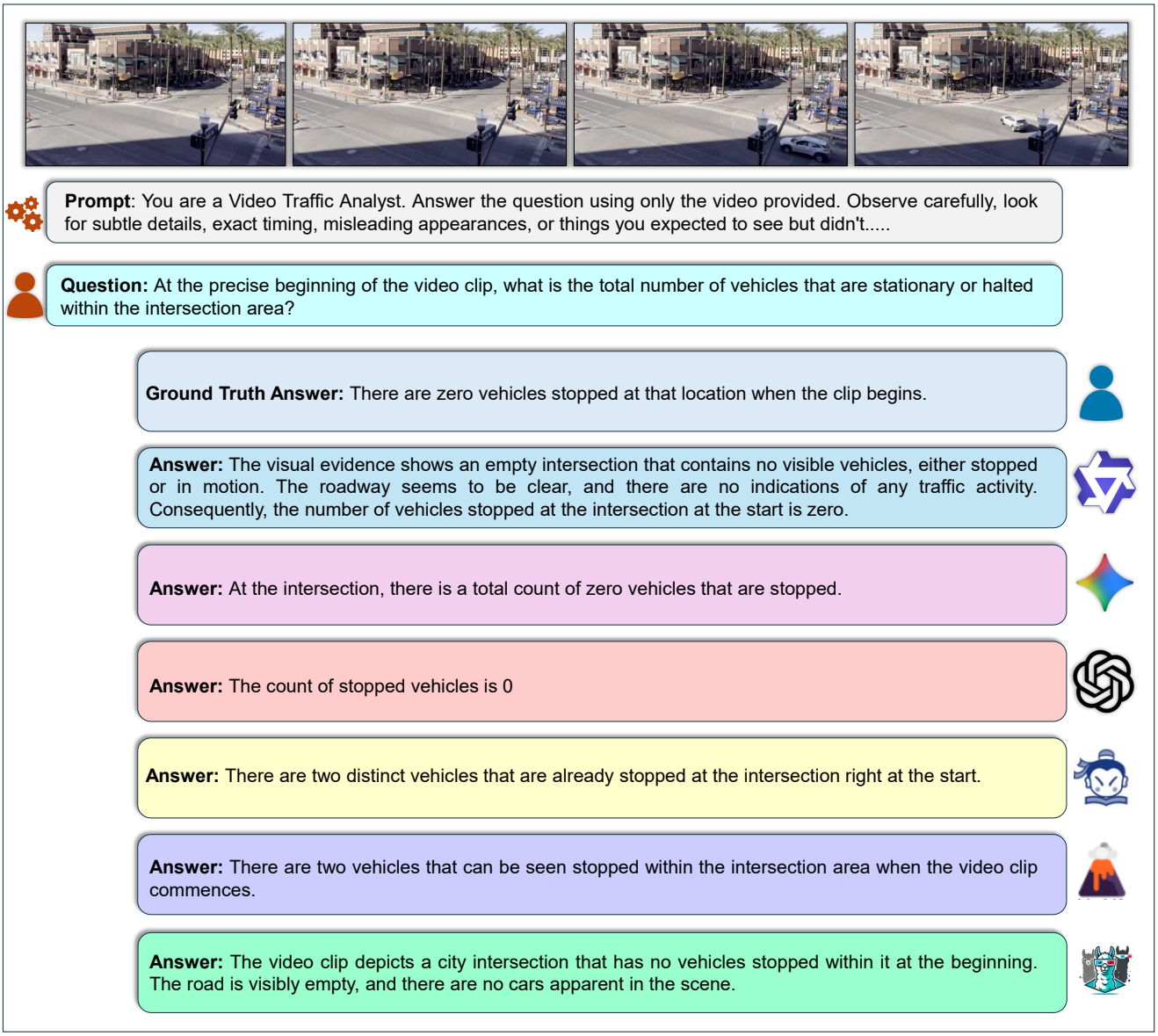


Figure 20. Example of an *attribution* question assessing a specific visual attribute, the number of stopped vehicles at the start of the clip. Most models correctly identify zero vehicles, consistent with the ground truth.

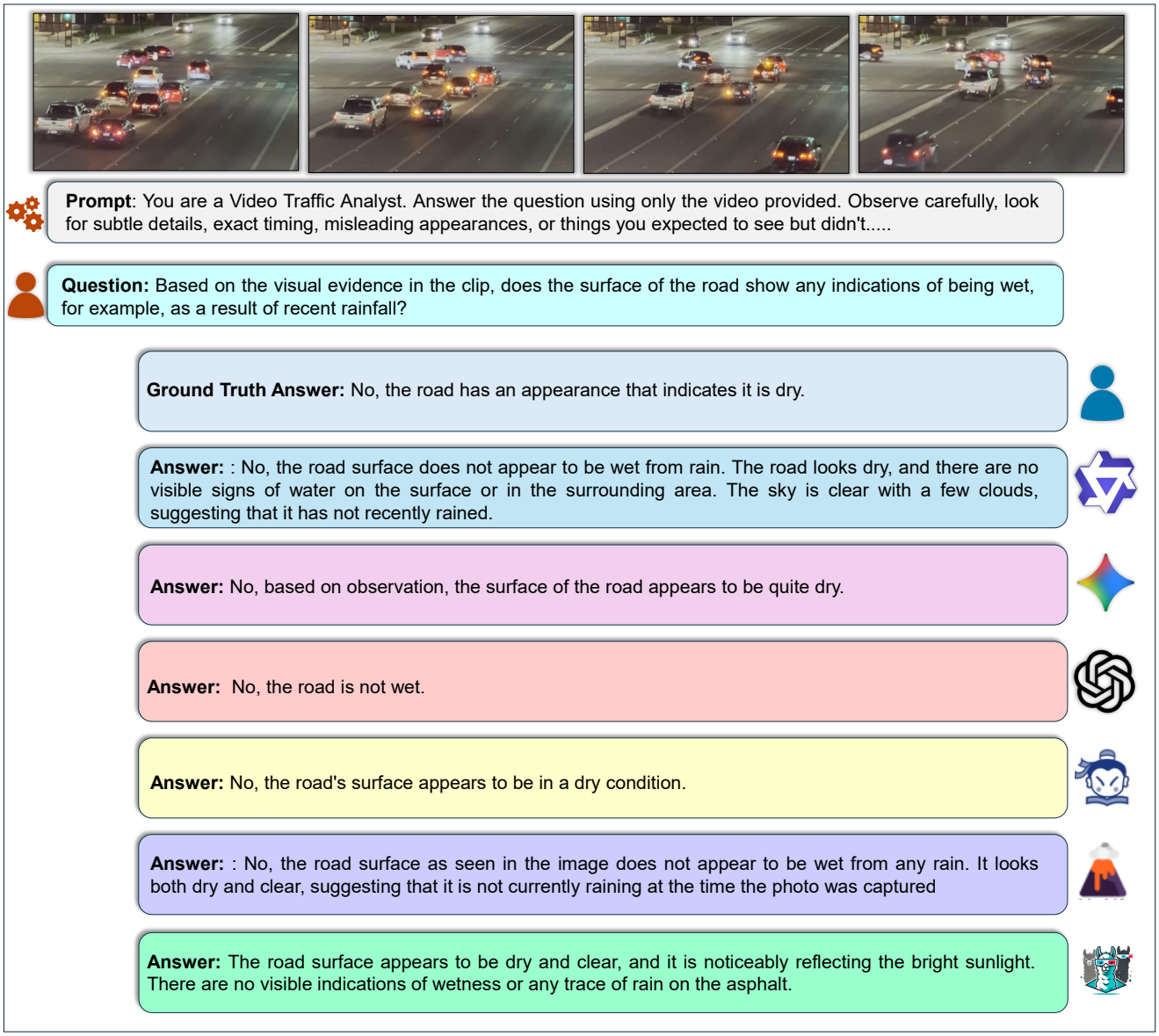
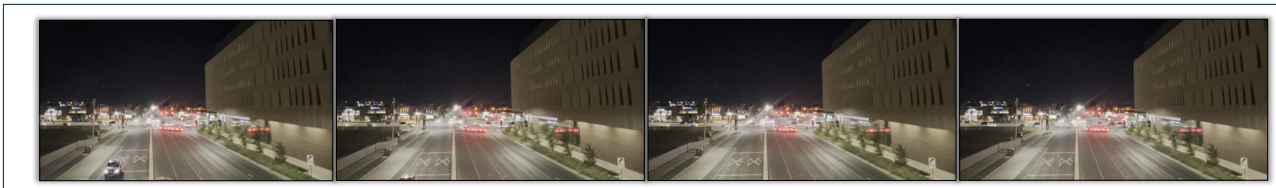


Figure 21. Example of a *basic understanding* question evaluating simple environmental perception, specifically, road surface wetness. All models correctly identify the road as dry, demonstrating high accuracy on this foundational reasoning task.



Prompt: You are a Video Traffic Analyst. Answer the question using only the video provided. Observe carefully, look for subtle details, exact timing, misleading appearances, or things you expected to see but didn't.....

Question: Is the statement accurate that, at any single moment, precisely five cars are simultaneously present in the traffic lane directly beneath the camera?

Ground Truth Answer: That statement is false; at no point during the entire clip does that specific situation (exactly five cars) ever happen.

Answer: This nighttime image uses a long exposure, which turns cars into light streaks and makes an accurate count impossible. While the streaks indicate motion from multiple vehicles, the exact number is uncertain. Therefore, it is incorrect to state that exactly five cars are in the lane under the camera.

Answer: No, that statement is incorrect.

Answer: No, it is not the case that there are precisely five cars situated under the camera at any single point in time.

Answer: No, that assertion is incorrect.

Answer: Yes, judging from the visual information in the image provided, it does appear to be the case that precisely five cars are occupying the lane directly under the camera at the same time.

Answer: No, the statement claiming that exactly five cars are occupying the lane under the camera simultaneously is not accurate.

Figure 22. Example of a *counterfactual inference* question evaluating a model's ability to reject a false premise. While most models correctly classify the statement as false, *LLaVA-NeXT* incorrectly confirms the non-existent event, revealing a failure in precise factual verification.

References

[1] Bharatesh Chakravarthi, M Manoj Kumar, and BN Pavan Kumar. Event-based sensing for improved traffic detection and tracking in intelligent transport systems toward sustainable mobility. In *International Conference on Interdisciplinary Approaches in Civil Engineering for Sustainable Development*, pages 83–95. Springer, 2023. 2

[2] Bharatesh Chakravarthi, Aayush Atul Verma, Kostas Daniilidis, Cornelia Fermuller, and Yezhou Yang. Recent event camera innovations: A survey. In *European Conference on Computer Vision*, pages 342–376. Springer, 2024. 2

[3] Yuhuang Hu, Elia Tsur, Andrea Vianello, Daniel San, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1801–1809, 2021. 3

[4] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 2, 3

[5] Alexander Kirillov and et al. Segment anything model 2. *arXiv preprint arXiv:2407.03991*, 2024. 2

[6] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, Elahe Arani, and Oleg Sinavski. Lingoqa: Visual question answering for autonomous driving. In *ECCV*, 2024. 12

[7] neuromorphicsystems. IEBCS: ICNS Event Based Camera Simulator. https://github.com/neuromorphicsystems/IEBCS, 2025. Accessed: 2025-11-04. 3

[8] Chirag Parikh, Deepti Rawat, Rakshitha R. T., Tathagata Ghosh, and Ravi Kiran Sarvadevabhatla. Roadsocial: A diverse videoqa dataset and benchmark for road event understanding from social video narratives, 2025. 16

[9] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: An open event camera simulator. In *Conference on Robot Learning (CoRL)*, 2018. 3

[10] Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *arXiv preprint arXiv:2408.08870*, 2024. 3

[11] Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *CVPR*, pages 9878–9888, 2021. 16

[12] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box, 2022. 2