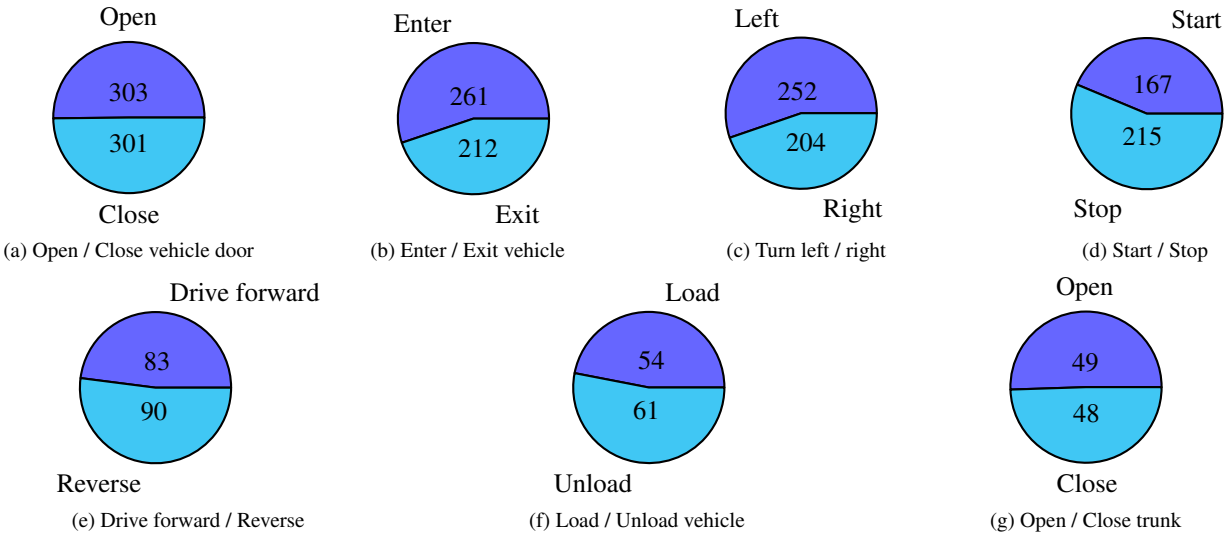


(a) Open / Close vehicle door (b) Enter / Exit vehicle (c) Turn left / right (d) Start / Stop

(e) Drive forward / Reverse (f) Load / Unload vehicle (g) Open / Close trunk

Figure A.1. **Number of video samples per action class within each opposite pair in the SOVABench (Intra-Pair) benchmark.**

The supplementary material contains the following information.

- Section A shows the sample numbers of the SOVABench (Intra-pair) classes.
- Section B shows the instructions used in each dataset for the task-aware prompting strategy.
- Section C shows the effect of not incorporating distracting samples into SOVABench (Inter-pair).
- Section D analyzes the latency differences between models.
- Section E performs an error analysis on some SOVABench (Intra-pair) samples.

## A. Intra-pair Distribution

Figure A.1 shows the number of video clips per action class within the set of opposite action pairs in SOVABench (Intra-pair). All pairs have close to even distributions.

## B. Instruction per Dataset

These are the instructions provided to MLLMs in each dataset:

- **SpatialBench:** List all spatial relationships between objects (e.g., position, size, distance, or orientation) in short sentences.
- **VSR:** List all pairwise spatial relations between objects in the image.
- **What's Up:** List all pairwise spatial relations between objects in the image.
- **CountBench:** Describe the image in a short caption that accurately states the number of main objects (in words) and includes a brief descriptive phrase.
- **Visual7W-Count:** Describe the image in a short caption that accurately states the number of main objects (in words) and includes a brief descriptive phrase.
- **SOVABench:** Briefly classify the actions occurring in this video. (+ System prompt: You are an expert video analysis model specialized in action recognition. Focus on how subjects and objects change and move over time rather than on static appearances or backgrounds. Infer the actions by reasoning about motion, temporal progression, and interactions across the video frames.)

## C. Inter-pair Evaluation Constrained to Queries

Table C.1 shows the performance of the models in the inter-pair evaluation protocol constrained to the set of queries (1,423 samples). The results indicate similar trends to those found in the default inter-pair evaluation setting. However, this constrained evaluation reveals a larger advantage for MLLMs with MLLM-to-Embedding framework over contrastive VLMs: more MLLM-based configurations surpass the strongest contrastive baseline, and the performance gaps widen (best MLLM: 44.8 mAP, best contrastive VLM: 41.4 mAP). This setting is also more challenging than the default one despite returning higher absolute values, since the margin from random performance (23.7 mAP) is compressed. This increased difficulty arises from the higher similarity among samples, where all clips depict vehicle-related activities and the distracting human-only clips are removed. In summary, when samples are more similar to each other (without distractors), the advantage of MLLMs becomes more pronounced.

| Model | Efficiency | Inter-pair (Constr.) |
|---|---|---|
| Random | – | 23.7 |
| *Contrastive Image-VLMs* | | |
| CLIP-ViT-L-14 | **22.88** | 37.4 |
| SigLIP2-Giant | 3.94 | 38.7 |
| MERU | 21.25 | 36.9 |
| *Contrastive Video-VLMs* | | |
| VideoCLIP | 0.47 | 41.4 |
| CLIP4Clip | 9.63 | 36.6 |
| ActionCLIP | 7.78 | 36.0 |
| *General MLLMs* | | |
| **InternVL3.5 8B$_{\text{GENERAL}}$** | 0.26 | 39.4 |
| **MiniCPM-V 4.5$_{\text{GENERAL}}$** | 0.10 | 42.2 |
| **InternVL3.5 8B$_{\text{TASK-AWARE}}$** | 0.33 | 44.2 |
| **MiniCPM-V 4.5$_{\text{TASK-AWARE}}$** | 0.26 | **44.8** |
| *Video-MLLMs* | | |
| **VideoLLaVA 7B$_{\text{GENERAL}}$** | 0.16 | 33.1 |
| **VideoLlama3 7B$_{\text{GENERAL}}$** | 0.42 | 40.7 |
| **VideoChat-R1 7B$_{\text{GENERAL}}$** | 0.06 | 36.3 |
| **VideoLLaVA 7B$_{\text{TASK-AWARE}}$** | 0.22 | 35.8 |
| **VideoLlama3 7B$_{\text{TASK-AWARE}}$** | 0.44 | 40.5 |
| **VideoChat-R1 7B$_{\text{TASK-AWARE}}$** | 0.13 | 42.7 |
| *API MLLMs* | | |
| **Gemini 2.5 Flash$_{\text{GENERAL}}$** | – | 38.1 |
| **Qwen3-VL 235B A22B$_{\text{GENERAL}}$** | – | 31.6 |
| **Gemini 2.5 Flash$_{\text{TASK-AWARE}}$** | – | 43.0 |
| **Qwen3-VL 235B A22B$_{\text{TASK-AWARE}}$** | – | 42.8 |

Table C.1. **Performance comparison of models in SOVABench (Inter-pair) restricted to the set of query samples, and efficiency comparison.** Efficiency is calculated as instances processed per second in the evaluation of the default SOVABench (Inter-pair) benchmark including distracting samples. SOVABench (Inter-pair) uses mAP.

## D. Efficiency Analysis

Table C.1 shows the number of instances processed per second for each model[1]. The table confirms that MLLMs are naturally heavier and, therefore, slower than usual contrastive VLMs. However, the key observation is that task-aware configurations consistently deliver faster inference than their general counterparts. This means that task-aware prompting offers not only performance gains, but also an efficiency advantage, allowing models to produce shorter and more task-relevant outputs.

[1]All values are obtained using GPUs of type NVIDIA GeForce RTX 3090.

## E. Error Analysis of SOVABench (Intra-pair)

We analyze the answers generated by MiniCPM-V 4.5$_{\text{TASK-AWARE}}$ in the opposite action pair <Open trunk, Close trunk> in the intra-pair evaluation protocol. Based on the error analysis, we group the errors detected into a set of 4 error modes. The subsequent list provides the description of each error mode and its counts from the analysis (shown in brackets). The list is prioritized by severity, meaning that if an error belongs to more than one mode, the most serious one is assigned. Examples of each error mode are illustrated in Figure E.1.

1. **Generation Errors / Hallucination (34):** The model produces semantically incorrect descriptions or fabricated entities, objects or actions.
2. **Under-Description (12):** The model provides a description that is factually correct but insufficiently expressive, omitting dynamic event details (*e.g.*, omitting any reference to the trunk of a vehicle when there is a trunk action).
3. **Temporal Misunderstanding (16):** The model identifies the correct action type but assigns an incorrect temporal direction, such as reversing "opening" and "closing" a trunk.
4. **Action Assumption (9):** The model correctly detects the high-level action but omits required intermediate steps, resulting in an incomplete or overly compressed event sequence (*e.g.*, describing "walking to a vehicle and loading an object" without mentioning the intermediate action of opening the trunk). Although not an error, it complicates the retrieval of actions belonging to the same class.

Following the numbers for each error mode, there are only 26 totally correct samples out of 97. In addition, the examples in Figure E.1 show that most retrieval failures originate from textual-generation issues rather than embedding quality alone.

**Model Answer:** A person is standing on the left side of the car, and another person is standing on the right side of the car. The person on the right side of the car opens the trunk of the car, and the person on the left side of the car also opens the trunk of the car. The person on the right side of the car is taking something out of the trunk of the car, and the person on the left side of the car is also taking something out of the trunk of the car. The person on the right side of the car is taking something out of the trunk of the car and putting it into a bag, and the person on the left side of the car is also taking something out of the trunk of the car and putting it into a bag. ...
**Observation:** Degenerated generation.

(a) Generation Errors / Hallucination

**Model Answer:** A person is opening the trunk of a car, and then the trunk is closed.
**Observation:** Incoherent answer.

(b) Generation Errors / Hallucination

**Model Answer:** A person is standing on the roof of a car, the car door is open, the car trunk is open, and the person is moving.
**Observation:** Hallucinated illogical person standing on the roof of a car.

(c) Generation Errors / Hallucination

**Model Answer:** Two people are standing in the back of a pickup truck, and they are looking down at something.
**Observation:** Statement with no key actions.

(d) Under-Description

**Model Answer:** A car is parked, and a person is getting in.
**Observation:** Trunk-related actions are not mentioned.

(e) Under-Description

**Model Answer:** A man is pushing a cooler on a dolly towards a blue truck. Another man is standing next to the truck, opening the tailgate. The man with the cooler moves the dolly towards the truck, and the other man continues to hold the tailgate open.
**Observation:** Wrong temporal understanding since the tailgate is being closed.

(f) Temporal Misunderstanding

**Model Answer:** A man walks to the back of a car, opens the trunk, and places a box inside.
**Observation:** Although the temporal direction within the close / open trunk pair is correctly determined, the next action should be unloading a box, not loading.

(g) Temporal Misunderstanding

**Model Answer:** A man is walking towards a car, and another man is standing next to the car. The man in the black shirt walks up to the car and the man in the red shirt puts a box in the trunk.
**Observation:** No mention to opening the trunk as intermediate action.

(h) Action Assumption

**Model Answer:** A man is walking to the back of a car, a man is standing in the back of a car, a man is loading a suitcase into the back of a car.
**Observation:** No mention to opening the tailgate as intermediate action.

(i) Action Assumption

**Model Answer:** A man is standing next to a car, then he opens the trunk of the car.

(j) Correct

Figure E.1. **Examples of generation errors obtained from MiniCPM-V 4.5$_{\text{TASK-AWARE}}$ in the <Open trunk, Close trunk> pair.** Errors are prioritized from more to less sever, including Generation Errors / Hallucination (a)-(c), Under-Description (d)-(e), Temporal Misunderstanding (f)-(g), and Action Assumption (h)-(i). In addition, (j) shows a successful case. Each example is composed by a filmstrip of the video, the model answer and an observation indicating why is wrong.