# Omni-RGPT: Unifying Image and Video Region-level Understanding via Token Marks

## Supplementary Material

In this supplementary material, we provide additional insights and details to support the main paper. In Sec. 8, we outline the implementation details, including dataset setups and training strategies. Sec. 9 presents extended quantitative results, offering further validation of our method's performance. To provide a more comprehensive understanding, Sec. 10 includes a set of qualitative results, showcasing visual examples that demonstrate our model's effectiveness in diverse scenarios. Sec. 11 provides a visualization analysis, including an exploration of the limitations of our model. In Sec. 12, we offer additional details about our newly constructed dataset, RegVID-300k. Finally, in Sec. 13, we discuss important ethical considerations.

## 8. Implementation Details

We further provide detailed setups for the image dataset. Our training consists of two stages: image pre-training, followed by image-video joint fine-tuning, which incorporates our newly proposed region-level video instruction dataset, RegVID-300k. For pretraining, we utilize a image-text paired dataset alongside an image-region pre-training dataset as proposed in RegionGPT [19]. Specifically, following VILA [37], we use interleaved image-text data [95] and conventional image-text pairs [4]. The image-region pre-training dataset includes Visual Genome [25], RefCOCOg [82], and V3Det [65]. For the image region-level dataset in the joint fine-tuning stage, we utilize the RegionGPT's fine-tuning dataset, which includes the ReCapD, a high-quality, GPT-assisted region-aware image dataset.

During pre-training, all learnable parameters, except those in the visual encoder and the language model, are trained for one epoch. In the fine-tuning stage, the model is trained for one epoch, with the vision encoder frozen while allowing all other parameters to be updated. We use a batch size of 16 and a learning rate of $5 \times 10^{-5}$. The entire training process, including both pre-training and fine-tuning stages, is completed within 24 hours using 8 nodes of 8×A100 GPUs.

## 9. More Quantitative Results

### 9.1. VCR

In Tab. 10, we present evaluation results on VCR [85] validation set under various settings, including category name usage, different region prompt representations (*e.g.*, bounding boxes and masks), and dataset-specific fine-tuning. The first row in the table corresponds to the results presented in the main paper, following the same evaluation protocol as existing methods [5, 87].

| Fine-tuning | Category | Region | Q → A | QA → R | Q → AR |
|---|---|---|---|---|---|
| ✓ | ✓ | mask | **88.5** | **90.1** | **79.9** |
| ✓ | | mask | 88.2 | **90.1** | 79.8 |
| ✓ | | bbox | 87.7 | 90.0 | 79.2 |
| | | bbox | 85.8 | 87.5 | 75.3 |

Table 10. **Accuracy (%) on VCR with different settings.**

**Category Name Usage.** Manually specifying category names for each target can be inefficient in real-world, user-interactive applications. For instance, consider the question "Why is `<region1>` moving towards `<region2>`?" paired with region prompts that include category names, such as `<region1>:person` and `<region2>:animal`. Incorporating these categories transforms the question into, "Why is person `<region1>` moving towards animal `<region2>`?"

In the second row, we evaluate our model's performance without utilizing such category names for the provided region prompts. While the results show a slight decrease in performance, they highlight that Omni-RGPT internalizes the semantic knowledge of regions within the question. This is achieved by leveraging visual context through a *direct region-language connection* using Token Mark.

Notably, unlike RoI-based methods [19, 87], which inject explicit visual features into the text prompt to embed semantic knowledge, Omni-RGPT accomplishes this by utilizing randomly sampled Token Mark within the text prompt. This highlights its ability to generalize and infer semantics directly from the visual context itself.

**Region Prompt Representation.** VCR provides two types of localization information for targets: bounding box coordinates and polygon-style segmentations. In the third row, we report the result using bounding box coordinates as the region prompt representation, which offers a less fine-grained target representation compared to masks. The overall performance shows a marginal decrease compared to the results obtained using masks, highlighting the flexibility of our framework in accommodating dual region prompt types.

**Task-specific Fine-tuning.** In the last row, we evaluate the performance of Omni-RGPT without VCR-specific fine-tuning. We use joint image-video trained weights, maintaining consistency with the settings used for other quantitative results presented in the main paper. The results here demonstrate that Omni-RGPT retains competitive perfor-

| RegVID -300k | Causal-VidQA | | | | Extended -Elysium |
|---|---|---|---|---|---|
| | D (A) | E (A) | P (AR) | C (AR) | |
| ✓ | **84.0** | **84.6** | **76.9** | **64.3** | **19.3** |
| | $82.7_{-1.3}$ | $84.4_{-0.2}$ | **76.9** | $63.5_{-0.8}$ | $13.5_{-5.8}$ |

Table 11. **Effectiveness of RegVID-300k.** Accuracy for Causal-VidQA and METOR scores for Extended-Elysium are reported.

mance even without task-specific fine-tuning, showcasing its robustness and broader applicability.

### 9.2. Effectiveness of RegVID-300k

In Tab. 11, we report the effectiveness of RegVID-300k, our new region-level video instruction dataset, on Causal-VidQA [30] and Extended-Elysium. The first row is the results reported in the main paper, obtained using the fine-tuned weight that incorporates the new dataset. In contrast, the second row shows results where the model is trained under the same settings but without including the new dataset, relying solely on annotations from public video datasets. The results underscore the effectiveness of RegVID-300k, as evidenced by the performance degradation, including the Description (D) task within Causal-VidQA ($-1.3$) and Extended-Elysium ($-5.8$). In particular, the large performance drop in Extended-Elysium highlights the contribution of our new dataset's diverse and detailed descriptions, which enhance the model's region-level captioning capability in the video domain.

### 9.3. Hallucination

Following the POPE [32] evaluation protocol, we report object hallucination evaluation results in Tab. 12. Performance is evaluated on the MSCOCO dataset under three settings: Random, Popular, and Adversarial. In the Random setting, our method achieves the best Accuracy and a competitive F1 Score, trailing the best-performing model (InstructBLIP [11]) by only $-0.47$. In the Popular setting, our Accuracy is marginally below the best-performing model (RegionGPT) by $-0.57$ while achieving the highest F1 Score among all methods. Finally, in the Adversarial setting, both Accuracy and F1 scores are lower than RegionGPT, with differences of $-4.0$ and $-1.97$, respectively. Overall, our method demonstrates competitive performance compared to the image-specialized counterparts.

## 10. More Qualitative Results

We showcase visualizations of Omni-RGPT's region-level understanding capabilities across both image and video scenarios. For all examples, we use a jointly fine-tuned model using the 7B language model. The inputs, including mask proposals, are sourced from the VIPSeg [45], which is not included in the training dataset (see Tab. 13).

### 10.1. Video Region-level Understanding

**Brief Region-level Captioning.** We present Omni-RGPT's brief region-level video captioning capability in Tab. 14. The results show that Omni-RGPT effectively generates concise captions for the given prompts, incorporating visual context across the entire sequence. For example, in the second case, the caption for the red backpack (`<region0>`) notes that it is unzipped, a detail observable from the third frame.

**Detailed Region-level Captioning.** Tab. 15 visualizes examples of detailed region-level video captioning. Key information is highlighted in light orange color. The outputs effectively capture transitions in the target's location and motion across the sequence. For instance, in the first case, Omni-RGPT describes the pillow (`<region0>`) as initially lying flat, later being lifted and adjusted by the person.

**Region-level QA.** In Tab. 16, we present various examples of region-level question-answering in videos. Omni-RGPT generates answers to diverse questions, including those about multi-object relationships and the location of a target within a sequence (*e.g.*, see the second example).

### 10.2. Image Region-level Understanding

**Region-level Captioning.** In Tab. 17, we demonstrate Omni-RGPT's captioning capabilities, including both brief and detailed descriptions. The results highlight the model's ability to describe specific details about the target object (*e.g.*, the names of instruments), including its visual features such as color, material, and size.

**Region-level QA.** In Tab. 18, we present a question-and-answering example that reflects a visual commonsense reasoning scenario. In this example, Omni-RGPT successfully generates answers by considering the commonsense context within the scene. Specifically, the output recognizes that wearing a helmet in an office is unusual, explains why, and suggests multiple possible reasons for such a case.

**Referring Expression Comprehension (REC).** In Tab. 19 and Tab. 20, we provide a visualization comparison with existing image-specialized methods [19, 42] on REC task using the RefCOCOg [82]. Among multiple proposals in the image, Omni-RGPT accurately identifies the most suitable one for the given caption. As shown in the examples, the challenge in this task lies in the presence of multiple objects with similar appearances in the scene, requiring a precise understanding of the spatial location of each visual element.

Our understanding of Omni-RGPT's robust REC capacity is that Token Mark establishes a strong ability to refer to visual tokens at the pixel level. This enables better differentiation between visually similar objects compared to RoI-based methods, which primarily rely on similarity using averaged appearances to locate the target region prompt.

| POPE | Metrics | Omni-RGPT | RegionGPT [19] | Shikra [7] | InstructBLIP [11] | MiniGPT4 [94] | LLaVA [38] | MM-GPT [18] | mPLUG-Owl [78] |
|---|---|---|---|---|---|---|---|---|---|
| Random | Accuracy (↑) | 88.76 | 87.80 | 86.90 | 88.73 | 77.83 | 86.00 | 50.03 | 53.30 |
| | Precision (↑) | 91.15 | 97.75 | 94.40 | 85.08 | 75.38 | 87.50 | 50.02 | 51.71 |
| | Recall (↑) | 86.60 | 78.13 | 79.26 | 93.33 | 82.67 | 84.00 | 100.00 | 99.53 |
| | F1 Score (↑) | 88.82 | 86.85 | 86.19 | 89.29 | 78.86 | 85.71 | 66.68 | 68.06 |
| | Yes | 48.96 | 41.20 | 43.26 | 55.20 | 54.83 | 48.00 | 99.97 | 96.23 |
| Popular | Accuracy (↑) | 86.63 | 87.20 | 83.97 | 81.37 | 68.30 | 76.67 | 50.00 | 50.63 |
| | Precision (↑) | 86.20 | 95.44 | 87.55 | 75.07 | 64.27 | 72.22 | 50.00 | 50.32 |
| | Recall (↑) | 86.60 | 78.13 | 79.20 | 93.33 | 82.40 | 86.67 | 100.00 | 99.27 |
| | F1 Score (↑) | 86.40 | 85.92 | 83.16 | 84.35 | 72.21 | 78.79 | 66.67 | 66.79 |
| | Yes | 50.23 | 40.93 | 45.23 | 62.57 | 64.10 | 60.00 | 100.00 | 98.63 |
| Adversarial | Accuracy (↑) | 81.67 | 85.67 | 83.10 | 74.37 | 66.60 | 73.33 | 50.00 | 50.67 |
| | Precision (↑) | 78.82 | 91.99 | 85.60 | 67.67 | 62.45 | 69.02 | 50.00 | 50.34 |
| | Recall (↑) | 86.60 | 78.13 | 79.60 | 93.33 | 83.27 | 84.67 | 100.00 | 99.33 |
| | F1 Score (↑) | 82.53 | 84.50 | 82.49 | 78.45 | 71.37 | 66.32 | 66.67 | 66.82 |
| | Yes | 54.93 | 42.47 | 46.50 | 68.97 | 66.67 | 61.33 | 100.00 | 98.67 |

Table 12. **Results on the object hallucination benchmark using the POPE evaluation pipeline on MSCOCO.**

# 11. Visualization Analysis

**Elysium dataset.** As discussed in the main paper, we observe that the video region-level captioning dataset from Elysium [64] is limited in capturing the rich information within video sequences. In Tab. 21, we illustrate Elysium's original inputs with ground-truth captions alongside responses generated by our model. From the examples, we note that: 1) the videos typically exhibit monotonous motion with limited dynamics, 2) the dataset primarily contains single dominant objects, which restricts its ability to evaluate multi-object scenarios, 3) the original ground truth captions are closer to noun-based descriptions rather than full captions. To facilitate better evaluation, we refined its annotations (referred to as Extended-Elysium in our main paper) using the same pipeline as our dataset curation.

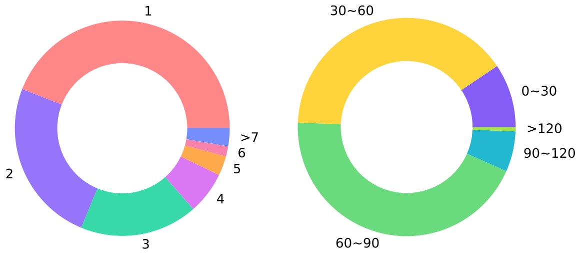
**Failure case.** In Tab. 22, the first example shows that the current design occasionally exhibits unstable performance to understand small objects. Additionally, since we employ frame-independent visual encoding without explicit temporal embeddings, our model has limited capability in interpreting the direction of objects.

**Occluded objects.** In Tab. 23, we analyze a heatmap of the Temporal Region Guide Head in diverse scenarios of occluded objects from [51].

# 12. RegVID-300k

## 12.1. Data Preparation

We select public datasets containing videos and manually annotated regions (*e.g.*, masklets, tracklets, single-frame bounding boxes) along with their associated region nouns. The full list of data sources, number of videos, and tasks used in the original datasets is summarized in Tab. 13. Since each dataset inherently contains biases due to the dataset construction process——such as video selection, annotation rules, or targeted scenes——we diversify the source datasets based on the targeted tasks. We use only a subset of the ElysiumTrack-1M [64] and AVA-Kinetics [27] datasets to diversify the video sources instead of increasing the videos within a single dataset. Because the original MeViS dataset [12] does not provide objects' nouns, we generate them by using GPT-4 [47] to extract target nouns from annotated referring sentences for each object. Also, since the ActivityNet-Entities dataset [92] consists of relatively long frames and includes scene changes, we divide each video into shorter clips based on the ground-truth event segments. Detailed statistics are visualized in Fig. 6. Our dataset includes multiple regions within each video, with detailed descriptions for each region.



(a) Number of regions (b) Caption length

Figure 6. **RegVID-300k statistics.** (a) The dataset contains multiple regions in videos. (b) The captions primarily range from 30 to 60 words, resulting in diverse instruction-following samples.

## 12.2. Data Visualization

The QA instruction-following samples and their corresponding videos are presented in Tabs. 24 and 25. These visualizations demonstrate the automatic generation of diverse and high-quality question-and-answer pairs related to the regions within the videos. The automated, multi-stage approach used to generate these region-level instruction-following samples is detailed in the following sections.

| Dataset | Original Task | Region Annotation | Number of Used Videos |
|---|---|---|---|
| MeViS [12]<br>Ref-YTVOS [55] | Referring video object segmentation | masklet | 1.6k<br>3.4k |
| Vid-STG [90]<br>HC-STVG [58] | Spatio-Temporal Grounding | tracklet | 5.3k<br>10.1k |
| BDD100k [80]<br>ElysiumTrack-1M [64] | Object Tracking | tracklet | 4.1k<br>10k |
| LV-VIS [63] | Open-vocabulary video instance segmentation | masklet | 3k |
| ActivityNet-Entities [92] | Dense Video Grounding | bounding box | 38.8k |
| AVA-Kinetics [27] | Action Localization | bounding box | 13.4k |
| Action Genome [23] | Action Graph | bounding box | 8.4k |

Table 13. **Source video dataset summary table.** We collect the videos and corresponding region annotations from diverse datasets.

### 12.3. GPT4o-Assisted Region-level Captioning

Since not all datasets have annotated object masks per frame (*i.e.*, masklets), we apply SAM [24] with ViT-H backbone [13] to the datasets that only contain the tracklets, such as Vid-STG [90], HC-STVG [58], BDD100k [80], and ElysiumTrack-1M [64]. Moreover, for the datasets that only contain the single frame bounding box, such as ActivityNet-Entities [92], AVA-Kinetics [27], and Action Genome [23], we apply SAM2 [54] to propagate the annotated bounding box to before and after the annotated frames to obtain the masklets. We followed the official code of SoM [77] to overlay the masklets onto the image frame-by-frame. Due to the context length and the budget limitation, we randomly subsample 16 frames from the video and resize to $334 \times 334$ to input to GPT4o with the text prompt Tab. 26. We input SoM-processed videos along with the object nouns to GPT4o to obtain the region-level captions. We employ **`gpt-4o-2024-08-06`** model in the paper.

### 12.4. Visual Hallucination Mitigation

We employ LLaMA-3.1-8B [14] for the LLM in the first and third stages (decomposing the caption into questions and re-generating the caption) and Qwen2-VL-7B [66] for MLLMs for the second step (multi-modal fact-checking). In the first step, the multiple questions, which ask about the contents in the captions, are generated from the region-level caption using the text prompt Tab. 27. As mentioned in the text prompt, we query the LLM only to include the questions that ask the contents in the input captions. In the second step, we input the generated questions into the Qwen2-VL with the original videos to verify the contents. We ask Qwen2-VL to answer the multi-choice closed-ended questions (Tab. 28). We input 8 video frames due to the context length. We collect the questions whose answers are "(D) Definitely No" and input these questions with the original captions to remove the non-validated contents from the original captions and re-generate them (Tab. 29).

### 12.5. Region-level Instruction Data Generation

From the refined region-level captions, we generate diverse instruction-following data. Three types of instruction-following data are generated: detailed descriptions, brief descriptions, and general conversations.

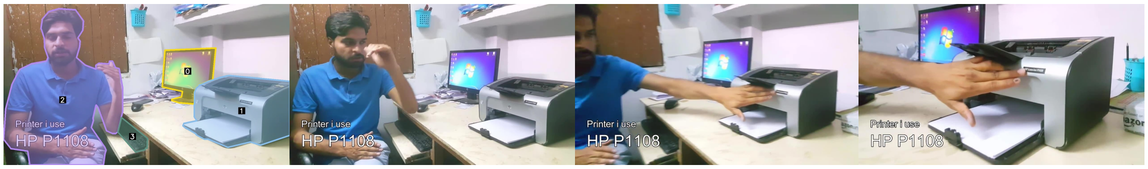
**Region-level Detailed Description.** We directly use the refined captions for each region as an answer. For the instructions, we augmented the input instruction prompt using GPT4 [47] from the original text prompt (`Describe the region in detail`), resulting in diverse input text prompts to enhance the LLM's adaptability to users' various text prompts (Tab. 33). For each region-level caption, we randomly select one of the instruction samples from the list to construct the question-and-answer pairs.

**Region-level Brief Description.** We employ GPT4 [47] to summarize the region-level detailed caption using the text prompt Tab. 30 and generate the region-level short descriptions. Similar to the region-level detailed description, we also diversify the input instruction from (`Briefly describe the region`), allowing users' diverse input questions (Tab. 32), and the question is randomly selected from the list to construct the question-and-answer pairs for the region-level brief description task.

**Region-level Conversation.** Inspired by LLaVA [38], we generate diverse region-specific question-and-answer pairs from the region-level captions using GPT4. The prompt to GPT4 is shown in Tab. 31. Since the captions contain diverse aspects of the regions in the video, the generated questions also cover the diverse aspects of them.

## 13. Ethics Concerns

We utilize the pre-trained weights of the LLM (Llama-2 [61]) and fine-tune them using publicly available image and video datasets, along with our re-annotated datasets. Consequently, societal biases present in the original LLM model weights, which are derived from large-scale web text corpora, may be inherited in the output response.



**User**: Give a brief interpretation of what stands out in `<region0>`.
**Omni-RGPT**: The monitor is on a desk with a printer, indicating a workspace setup.

---

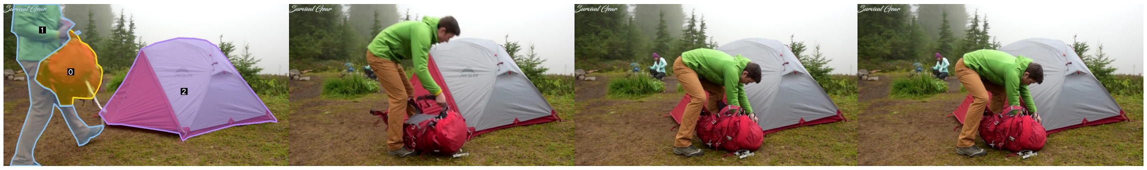
**User**: Provide a short summary of what's visible in `<region1>`.
**Omni-RGPT**: The printer is a sleek, modern device on a desk in a home office setting.

---

**User**: Describe the notable objects or subjects you can see in `<region2>`.
**Omni-RGPT**: A person in a blue shirt operates a printer in a home office setting.

---

**User**: How would you quickly explain what is featured in `<region3>`?
**Omni-RGPT**: The keyboard is on a desk near a printer, used for computer tasks.



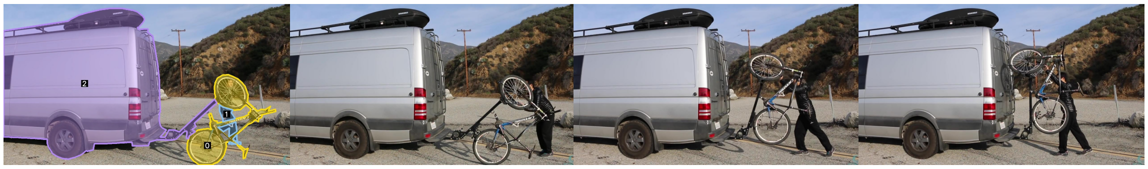
**User**: Provide a concise rundown of what is visible in `<region0>`.
**Omni-RGPT**: A red backpack is being unzipped and organized near a tent in a misty outdoor setting.

---

**User**: Describe the notable objects or subjects you can see in `<region1>`.
**Omni-RGPT**: A person in a green jacket and brown pants is organizing a red backpack near a tent in a foggy outdoor setting.

---

**User**: What are the key points of focus in this section `<region2>`?
**Omni-RGPT**: A red and gray tent is set up on a grassy area in a misty forest.



**User**: Create a short label for the visible content in `<region0>`.
**Omni-RGPT**: A bicycle is being lifted onto a rack on a van in a parking lot.

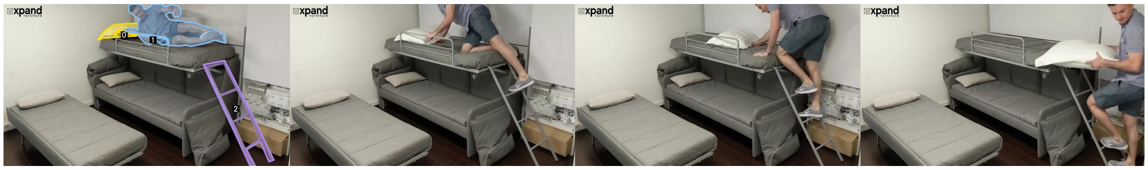
---

**User**: Draft a short title for the video content enclosed by `<region1>`.
**Omni-RGPT**: A person is securing a bicycle to a trailer attached to a van in a rural setting.

---

**User**: Provide a concise label for the highlighted `<region2>`.
**Omni-RGPT**: A silver van is parked on a gravel surface with a bicycle attached to its rear.

Table 14. **Qualitative visualization of brief region-level video captioning capability.**



**User**: Give an extensive description of the area labeled `<region0>`, paying attention to textures, colors, movements, and notable details.

**Omni-RGPT**: The pillow is positioned on the top bunk of a bunk bed. It is initially lying flat, but as the sequence progresses, it is lifted and adjusted by the person. The pillow is white and appears soft, contrasting with the gray bedding. It is situated in a minimalist room with a wooden floor and white walls, providing a simple and clean backdrop.

---

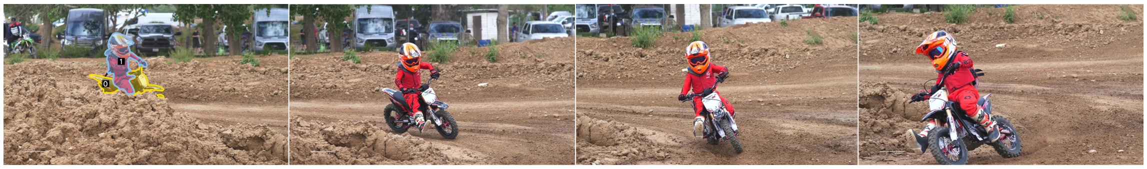
**User**: Explain the visible details of the subject in `<region1>`, including its appearance, movements, and any distinguishing traits.

**Omni-RGPT**: The person is positioned on a bunk bed, initially lying on the top mattress. They are wearing a light-colored shirt and shorts, and appear to be adjusting their position. The room is minimalistic, with a plain wall and a few pieces of furniture, including a ladder leading to the top bunk. The person is seen moving from a reclined position to sitting up, then standing, indicating a transition from rest to activity. The setting suggests a casual, relaxed environment.

---

**User**: Explain the key attributes and features of the subject found in `<region2>` with in-depth detail.

**Omni-RGPT**: The ladder is positioned on the right side of the room, attached to the bunk bed. It is metallic and extends upwards, providing access to the upper bunk. The ladder is stationary, serving as a functional piece of furniture in the room.



**User**: Create a thorough, descriptive caption for the portion of the video located in `<region0>`, touching on its materials, origin, and symbolic elements.

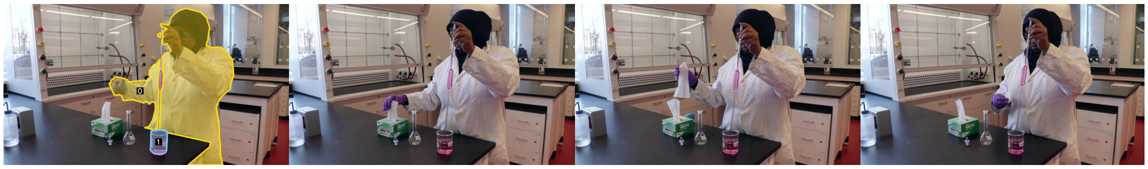
**Omni-RGPT**: The motorbike is a small dirt bike, designed for off-road racing. It is positioned on a dirt track, moving steadily across the uneven terrain. The bike is equipped with knobby tires for better traction on the loose surface. The rider is wearing protective gear, including a helmet and a red racing suit, indicating participation in a competitive event. The background shows a dirt field with several parked vehicles and spectators, suggesting a motocross event. The bike maintains a consistent speed, navigating the bumps and turns of the course.

---

**User**: Delve deeply into the description of the contents at `<region1>`, focusing on the interplay of textures, colors, movements, and standout aspects.

**Omni-RGPT**: The person is riding a small dirt bike, navigating through a rugged, dirt-covered terrain. They are wearing a red outfit and a helmet, indicating they are likely participating in a motocross event. The person is positioned centrally on the bike, maintaining balance as they move forward. The background shows a dirt track with several vehicles parked, suggesting a motorsport event or competition. The person appears focused, maneuvering the bike over uneven ground, showcasing their riding skills in this challenging environment.

Table 15. **Qualitative visualization of detailed region-level video captioning capability.** Specifically, the sentences highlighted in orange indicate that our model effectively captures the temporal motion (or dynamics) of the targeted objects.



**User**: What is the setting of the scene?
**Omni-RGPT**: The scene is set in a laboratory environment with a clean and organized workspace.

---

**User**: What objects are near `<region0>`?
**Omni-RGPT**: The person is near a laboratory bench with various laboratory equipment, including a box of tissues, a beaker, and a bottle.

---

**User**: Could you explain what `<region0>` is doing in the scene?
**Omni-RGPT**: The person is engaged in a laboratory experiment, handling a test tube filled with a pink liquid.

---

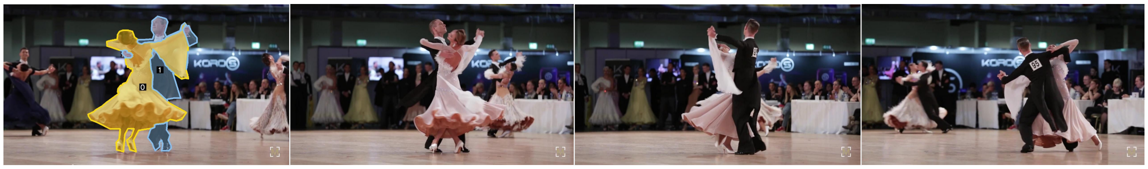
**User**: What is `<region0>` wearing, and what is it used for?
**Omni-RGPT**: The person is wearing a white lab coat and purple gloves, which are used for protection in a laboratory setting.

---

**User**: What is the object `<region0>` is holding in their left hand and what is used for?
**Omni-RGPT**: The person is holding a dropper in their left hand, which is used for precise liquid measurements.

---

**User**: What is the purpose of the object in region `<region1>`?
**Omni-RGPT**: The beaker is used for mixing or measuring substances.



**User**: Describe the setting in the video.
**Omni-RGPT**: The setting is a ballroom with a polished wooden floor, surrounded by an audience seated at tables, creating a formal and elegant atmosphere.

---

**User**: Is there any interaction visible between the regions `<region0>` and `<region1>`?
**Omni-RGPT**: Yes, the couple is engaged in a synchronized dance routine, with the woman in a flowing dress and the man in a formal suit.

---

**User**: What can you say about the movement of the region `<region0>`?
**Omni-RGPT**: The person is engaged in a graceful dance, moving fluidly across the dance floor.

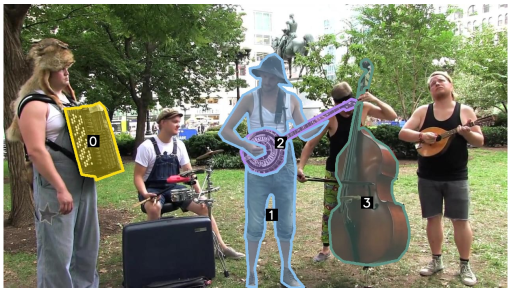
---

**User**: Where is the region `<region1>` located in the frame at the beginning of the video?
**Omni-RGPT**: The person is positioned towards the center of the dance floor.

---

**User**: Where is the region `<region1>` positioned in the frame at the end of the video?
**Omni-RGPT**: The person is positioned towards the center-right of the frame.

Table 16. **Qualitative result of region-level video QA.**



- **Brief captioning**

**User**: Can you briefly explain what you see in the region `<region0>`?
**Omni-RGPT**: A black and white accordion.

---

**User**: What are the key features or subjects captured in this region `<region1>`?
**Omni-RGPT**: A man in a white tank top and blue jean shorts playing a banjo.

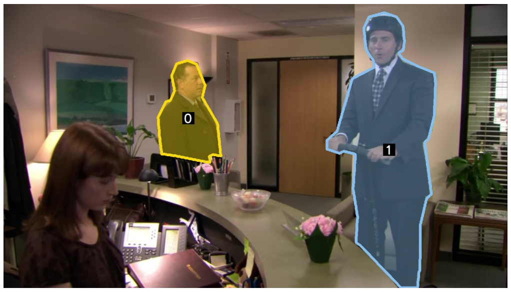
- **Detailed captioning**

**User**: Zoom in on the area indicated by `<region2>` and describe every discernible detail, from texture and color to form and function.
**Omni-RGPT**: The banjo in the close-up is a large, round, and wooden instrument. It has a distinctive shape, with a curved body and a long neck. The banjo is held by a person in the group, who is likely playing it as part of the musical performance. The banjo's color is brown, which is a common color for wooden instruments. The close-up view of the banjo allows for a detailed examination of its design and craftsmanship, showcasing its unique features and the skill of the musician playing it.

---

**User**: Craft a thorough narrative about the piece of the image highlighted by `<region3>`, from its aesthetic qualities to its possible historical context.
**Omni-RGPT**: The double bass in the close-up is a large, wooden instrument with a brown color. It has a distinctive shape, with a curved body and a long neck. The double bass is held by a person in the group, who is likely playing it as part of the musical performance. The instrument's size and style suggest that it is a traditional, acoustic double bass, which is commonly used in jazz and classical music.

Table 17. **Qualitative result of region-level image captioning.**



**User**: Why `<region0>` is looking at `<region1>`?
**Omni-RGPT**: The man in the suit is looking at the man in the helmet, possibly because he is wearing a helmet in an office setting, which is an unusual sight. This could be a conversation starter or a point of interest for the man in the suit.

---

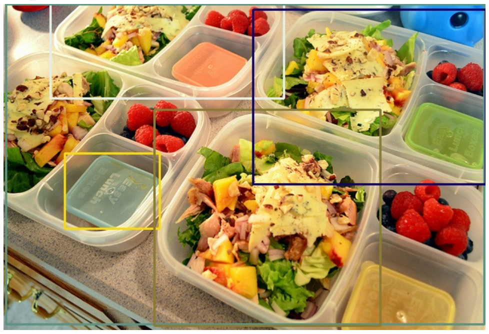
**User**: Please explain why `<region1>`'s outfit is unusual in office.
**Omni-RGPT**: The man's outfit is unusual in the office because he is wearing a suit and a helmet. Typically, suits are associated with formal or professional attire, while helmets are more commonly worn for safety purposes, such as in construction or sports. Wearing a helmet in an office setting is not a common sight and can be considered unconventional or even humorous. This unusual combination of attire might draw attention and spark conversations among the people in the office.

---

**User**: What might be the purpose of `<region1>` wearing a helmet to the office?
**Omni-RGPT**: The man wearing a helmet to the office might be doing so for a variety of reasons. One possibility is that he is a cyclist or a motorcyclist who has just arrived at work and is wearing the helmet as a safety measure. Another possibility is that he is participating in a themed event or a costume party at the office, where wearing a helmet is part of the dress code. Alternatively, he could be wearing the helmet as a prop for a presentation or a demonstration, where the helmet serves a specific purpose related to his job or the topic of his presentation. It is essential to consider the context and the man's role in the office to better understand the purpose behind his helmet.

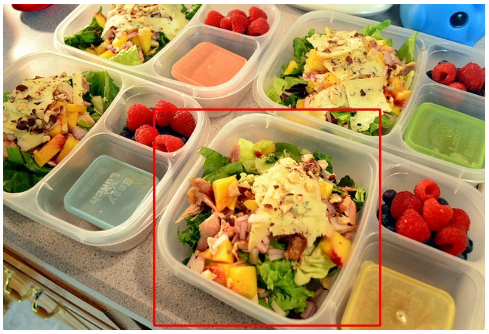
Table 18. **Qualitative result of region-level image QA.**

**Input**:

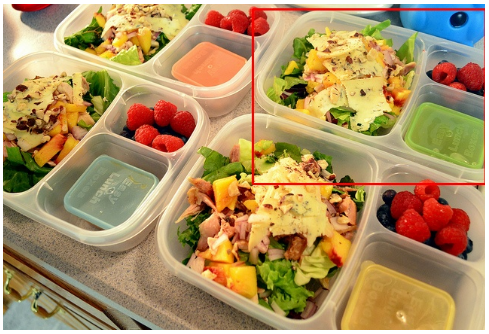


**Caption**: the bottom right tray with a salad in it

**Omni-RGPT**:



**RegionGPT**:



**Groma**:

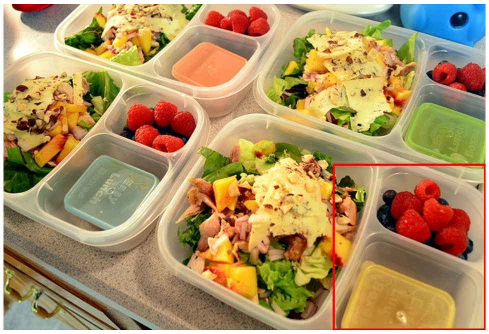
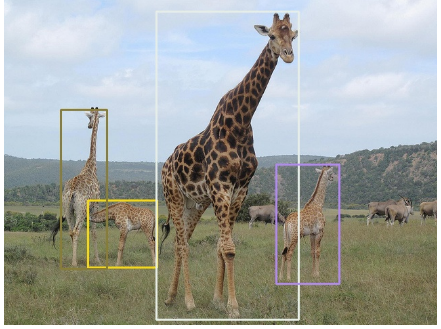


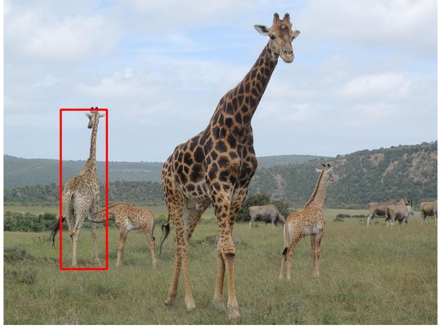
Table 19. **Qualitative comparison of image REC.**

**Input**:

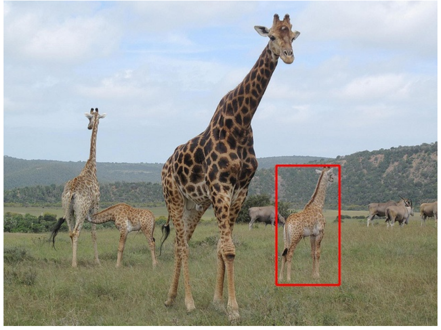


**Caption**: a mother giraffe nursing her young

**Omni-RGPT**:



**RegionGPT**:



**Groma**:

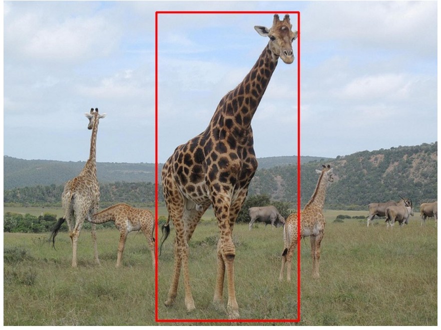
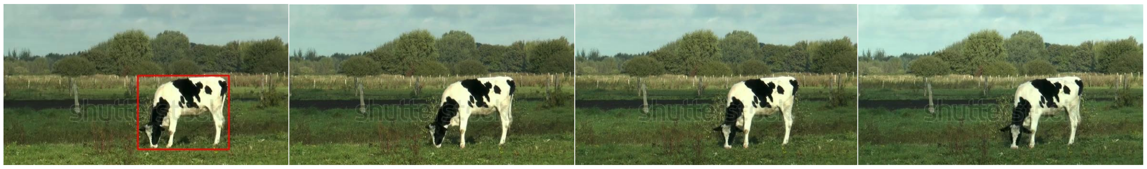


Table 20. **Qualitative comparison of image REC.**



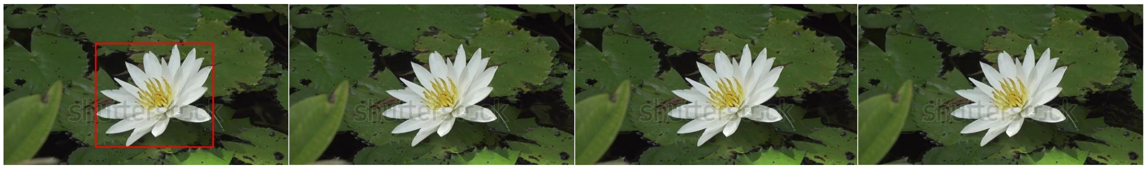
- **Original Elysium**

**User**: Can you briefly explain what you see in the `<region>`?
**Ground Truth**: cattle
**Omni-RGPT**: A black and white cow grazes peacefully in a serene, rural field.

---

- **Extended-Elysium**

**User**: Explain the key attributes and features of the subject found in `<region>` with in-depth detail.
**Ground truth**: The cattle depicted in the series of frames is grazing steadily in a lush, green pasture. Positioned calmly in the foreground, it remains focused on eating the grass beneath. The vast field extends into the distance, bordered by a line of trees providing a natural barrier and creating a serene rural landscape. The bright, cloudy sky casts a gentle light across the meadow, highlighting the cattle's black and white patches. Throughout the frames, the cattle maintains its position, engrossed in the act of feeding.
**Omni-RGPT**: The cow is positioned in the middle of a lush green field, surrounded by a backdrop of dense trees under a clear sky. The cow is grazing, with its head lowered towards the ground, indicating it is feeding on the grass. The cow's body is mostly white with distinct black patches, and it appears to be calmly moving its head as it feeds. The field is expansive, with a few scattered trees and a fence line visible in the distance, suggesting a rural setting. The cow's motion is gentle and unhurried, typical of a peaceful pastoral scene.



- **Original Elysium**

**User**: What short caption would best capture the details of `<region>`.
**Ground truth**: flower
**Omni-RGPT**: A white water lily with a yellow center blooms serenely on a pond surrounded by lily pads.

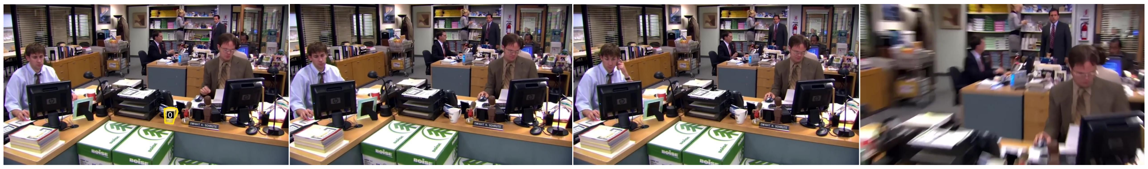
---

- **Extended-Elysium**

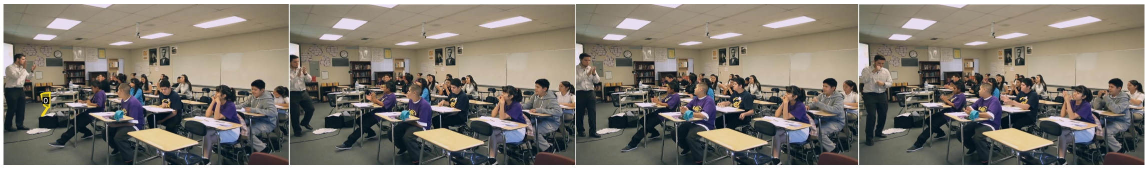
**User**: Offer a full description of the object found in `<region>`.
**Ground truth**: The flower, likely a water lily, is prominently positioned atop broad, flat green leaves on the water's surface. Its pristine white petals radiate outward from a vibrant yellow center, creating a striking contrast against the leafy backdrop. In the serene water setting, the flower remains still, exhibiting a delicate and serene presence. The surrounding green leaves occasionally display small imperfections, adding a natural, untouched feel to the scene. The overall environment is tranquil, exemplifying a typical pond ecosystem where such flowers thrive.
**Omni-RGPT**: The water lily is prominently positioned in the center of the frame, displaying its vibrant white petals that are slightly ruffled, suggesting a gentle breeze. The lily is surrounded by large, broad green leaves that float serenely on the water's surface, providing a lush, natural backdrop. The water lily remains stationary, its petals radiating outward, creating a symmetrical and elegant appearance. The scene is set in a tranquil pond environment, with the water reflecting the surrounding greenery, enhancing the serene and picturesque atmosphere.

Table 21. **Visualized analysis of Elysium dataset. Top:** The original Elysium annotations are short and predominantly noun-centric. **Bottom:** We present Extended-Elysium, refined annotations created using our dataset curation pipeline.

- **Small object understanding**

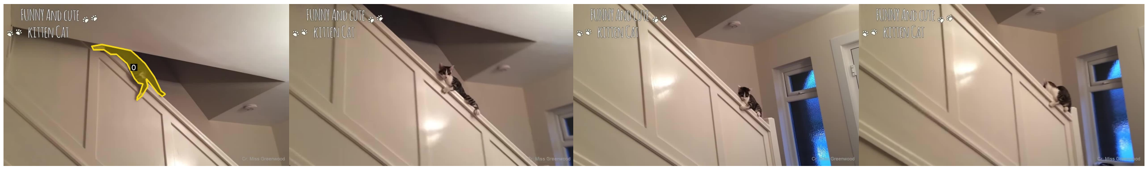


**User**: Please provide a short depiction of the content shown in `<region0>` in the video.
(✓) **Omni-RGPT**: There is a white cup above a table in the office.



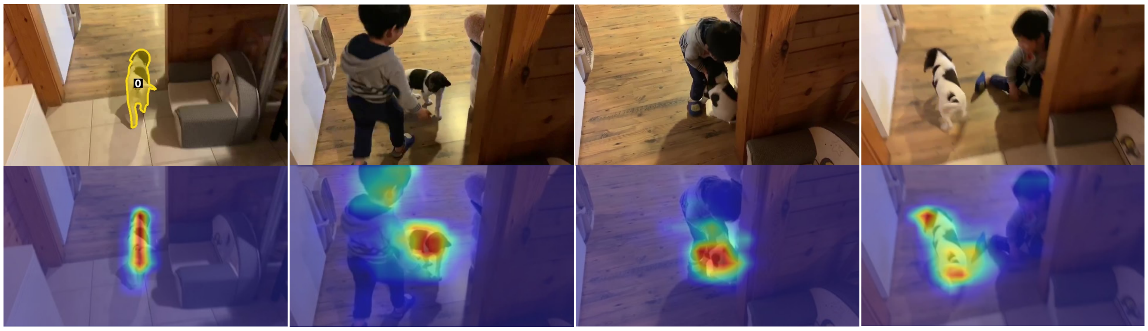
**User**: Provide a description of what is presented in the `<region0>`.
(✗) **Omni-RGPT**: A stationary backpack is on the floor near a desk in a classroom.

- **Direction understanding in video**

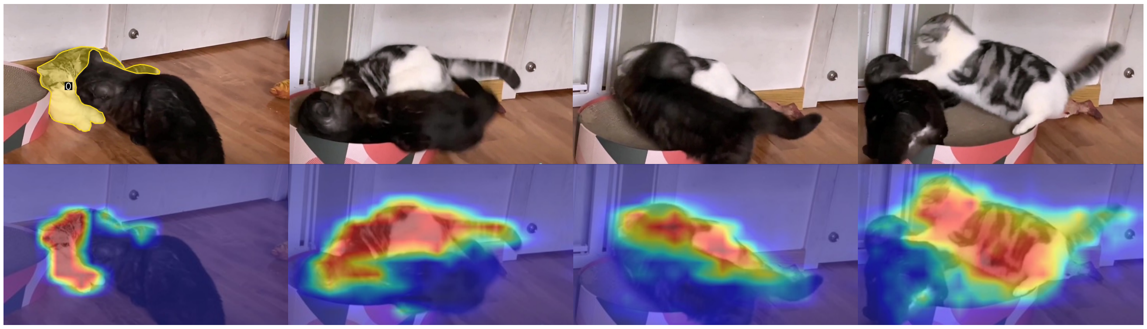


**User**: Give an extensive description of the area labeled `<region0>`, paying attention to textures, colors, movements, and notable details.
(✗) **Omni-RGPT**: The cat is positioned on a staircase, appearing to be in motion as it climbs upwards. Its body is oriented towards the top of the stairs, suggesting it is either ascending or exploring the area. The setting is indoors, with a light-colored wall and ceiling visible in the background. The cat's movement is fluid, indicating a sense of curiosity or playfulness. The scene is well-lit, with a window nearby, adding to the cozy and domestic atmosphere.

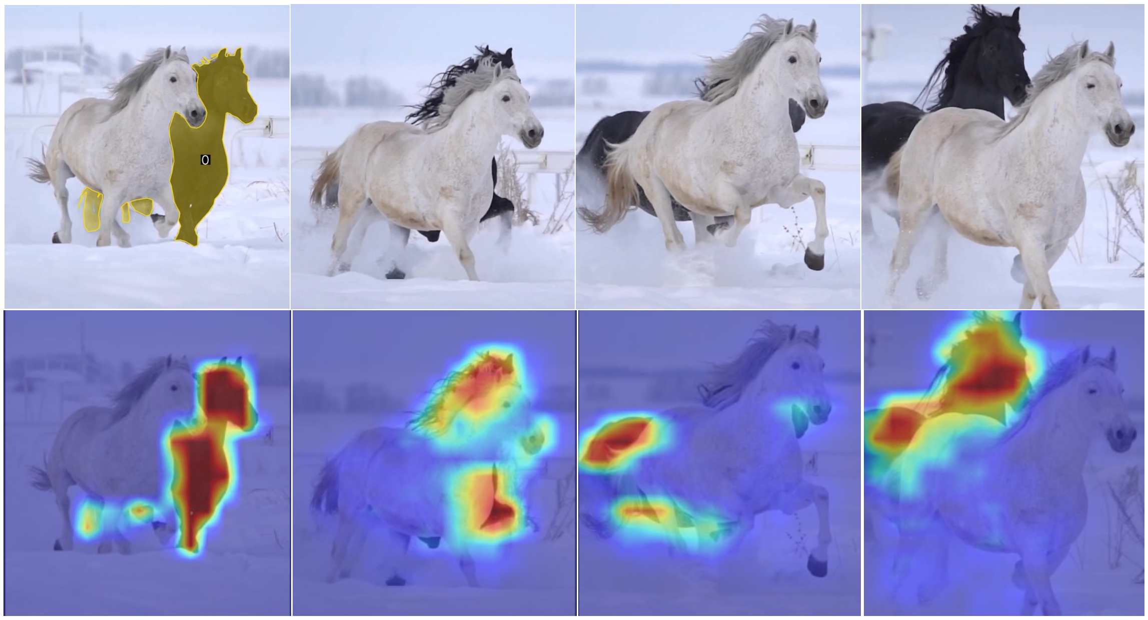
Table 22. **Failure Cases**. **Top:** Omni-RGPT occasionally exhibits unstable performance on small objects. The model mistakenly identifes the black chair as a backpack in the second example. **Bottom:** Omni-RGPT shows limited capacity for understanding the direction of objects. The model answers that the cat is climbing upward, even though it is actually moving downward.



**User**: A short description of `<region0>` from the video:
(**Omni-RGPT**: There is a dog in white and black towards a child in gray.

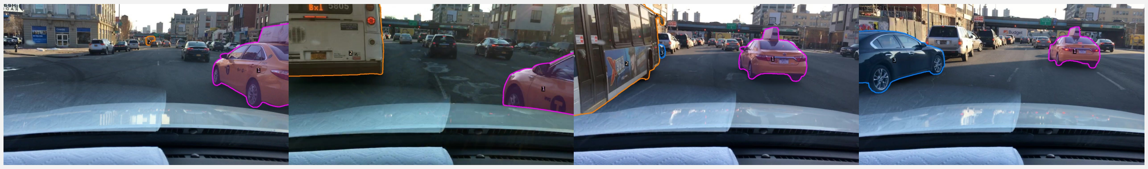


**User**: What stands out to you the most in this `<region0>`?
**Omni-RGPT**: There is a cat hitting another cat on the floor.



**User**: What stands out to you the most in this `<region0>`?
**Omni-RGPT**: A black horse gallops energetically through a snowy landscape.

Table 23. **Heatmap visualization of Temporal Region Guide Head on occluded objects**.



**User**: Generate a long, detailed caption for the segment of the video at `<region1>`, covering aspects such as its origin, material, and any symbolic meaning.
**Assistant**: The car is a yellow taxi positioned on the right side of the road. It stays largely consistent in position, seemingly moving at a steady pace parallel to the dashcam car. The taxi's presence remains constant, indicating it's traveling in the same general direction as the surrounding traffic. The background suggests an urban environment, with multiple lanes and buildings visible, typical of a city street.

---

**User**: Expand upon the visual description of what is contained within the marked area `<region2>`, covering its appearance, size, and stylistic features.
**Assistant**: The bus is located ahead of the taxi and further up the road. Initially, it appears small due to its distance, but gradually becomes larger and more detailed as the dashcam car advances down the road. The bus is traveling in the same direction as the rest of the traffic and navigating through the urban setting. Close proximity to other vehicles and buildings in the scene are evident as they approach a more congested area.

---

**User**: Explain the key attributes and features of the subject found in `<region3>` with in-depth detail.
**Assistant**: The car is a black sedan situated to the left of the bus, becoming visible later in the sequence of frames. Initially, it is further behind but gradually moves up to parallel with the dashcam car. The black sedan is traveling in the same direction as the taxi and bus within the lanes. It doesn't exhibit any erratic movements, maintaining a steady course. The scene's background remains consistently urban, with the overhead signage and multiple vehicles indicating a busy city road.

---

**User**: How would you briefly summarize the key points in `<region1>`?
**Assistant**: A yellow taxi is traveling steadily alongside the dashcam car on a city street.

---

**User**: Describe what's highlighted within `<region2>`.
**Assistant**: A bus is ahead of the taxi and grows larger as the dashcam car approaches in urban traffic.

---

**User**: Write a short description for the region `<region3>`.
**Assistant**: A black sedan becomes parallel with the dashcam car, moving steadily in city traffic.

---

**User**: What color is the taxi in region `<region1>`?
**Assistant**: The taxi is yellow.

---

**User**: Where is the taxi positioned relative to the road?
**Assistant**: The taxi is positioned on the right side of the road.

---

**User**: Is the yellow taxi moving erratically?
**Assistant**: No, the taxi is moving at a steady pace parallel to the dashcam car.

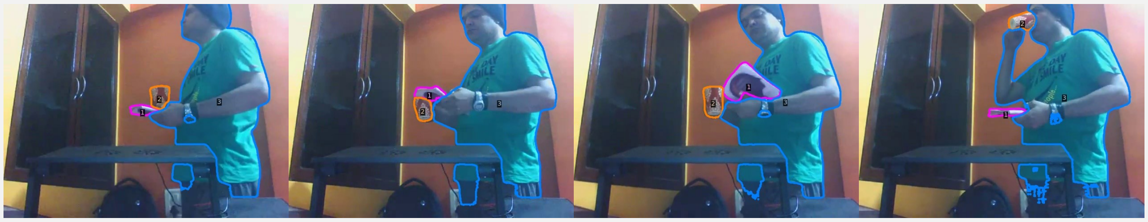
---

**User**: Where is the bus located in relation to the taxi?
**Assistant**: The bus is located ahead of the taxi and further up the road.

---

**User**: When does the black sedan in region `<region3>` become visible?
**Assistant**: The black sedan becomes visible later in the sequence of frames.

Table 24. **Visualized instruction-following sample and the video in RegVID-300k.** The object masks are overlayed onto the video frames. The video is from the BDD100k dataset [80].



**User**: Elaborate on the artwork shown in the region indicated by `<region0>`, mentioning its color, appearance, size, style, and any standout features.
**Assistant**: The book is held in the person's left hand. It appears to be a soft-cover book, with the pages slightly visible as they move and adjust their grip on it. The person keeps the book close to their torso while standing near a table in a room with an orange wall and a wooden cabinet in the background. As they maneuver, the book remains prominently in hand, indicating it could be light and easy to handle.

**User**: Provide a detailed account of the object located at `<region1>` in the video.
**Assistant**: The object, looking like a cup or bottle, is initially near the center of the person's torso and gradually moves toward their mouth in the latter frames. This action suggests it might contain a beverage as they appear to be preparing to drink from it. The cup has a distinct rim, which contrasts against the orange wall, flanked by the cabinet on the left side of the frame.

**User**: Describe comprehensively the object at `<region2>`, including its physical traits, style, actions, and visible characteristics.
**Assistant**: The person is standing, wearing a green T-shirt and a beanie. They are constantly adjusting their position while holding a book in one hand and a cup or bottle in the other, showing a casual and relaxed demeanor. The scene is set in a room with a bold orange paint and a wooden cabinet with reflective glass, indicating a cozy indoor environment.

**User**: Write a short description for the region `<region1>`.
**Assistant**: A cup or bottle moves towards the person's mouth, suggesting a drink.

**User**: How would you briefly summarize the key points in `<region2>`?
**Assistant**: A person in a green T-shirt holds a book and cup in a cozy room.

**User**: Describe the action being performed with the object in region `<region1>`.
**Assistant**: The object, which looks like a cup or bottle, is gradually moving toward the person's mouth, suggesting they are preparing to drink from it.

**User**: What is the clothing color of the person in region `<region2>`?
**Assistant**: The person is wearing a green T-shirt.

**User**: Is there any motion involving the book in region `<region0>`?
**Assistant**: Yes, the book slightly moves as the person adjusts their grip on it.

**User**: In what hand is the object in region `<region1>`?
**Assistant**: The cup or bottle is in one hand, likely not the same where the book is held.

**User**: What is the backdrop environment in the entire video?
**Assistant**: The environment is a room with an orange wall and a wooden cabinet with reflective glass, which gives a cozy indoor feel.

Table 25. **Visualized instruction-following sample and the video in RegVID-300k.** The object masks are overlayed onto the video frames. The video is from the Action Genome dataset [23].

You are an intelligent video captioner designed to generate the description of objects. Your task is to generate detailed descriptions of the objects labeled with a bright numeric ID at the center of video frames.

##INSTRUCTIONS:
- The description should start with the object noun. DO NOT add any irrelevant content to the answer. Each object should be described in around 90 words, but DO NOT include information that you are unaware of.
- INCLUDE the following information in the object descriptions: (1) temporal (motion, action), (2) contextual (spatial position) information of objects, and (3) global (background, scene) information. If you cannot see the object clearly, output the following <object id>: None.
- MUST NOT include the outline color information in the description. Use bracket [] to refer to the object IDs in the descriptions if needed.

These are video frames. Generate descriptions of each labeled object, `[<id1>]: <noun1>`, `[<id2>]: <noun2>`..., .

Table 26. **Input text prompts to GPT4o for the detailed region-level captioning.** We input the above prompts and SoM-processed videos to obtain region-level captions. The placeholders `<idX>` and `<nounX>` are replaced with $X$-th object indices marked in the video and the corresponding object's ground-truth noun, respectively.

Your task is to generate multiple questions comprehensively from given object descriptions in JSON format. The question should ask about the object from the given descriptions. The questions comprehensively ask all the contents in the object descriptions. DO NOT generate a question whose answer will be no. Respond only with valid JSON. Do not write an introduction or summary.

Description: `<ic sample1 caption>`
Question: {“questions”: `[ <ic sample1 q1>`, `<ic sample1 q2>`, ..., `]` }

Description: `<ic sample2 caption>`
Question: {“questions”: `[ <ic sample2 q1>`, `<ic sample2 q2>`, ..., `]` }

Description: `<caption>`
Question:

Table 27. **The text prompts to decompose the object-level to multiple fact-checking questions.** To enhance the output quality, we manually create in-context pairs (input object description and output questions). The placeholder `<ic sampleX caption>` is replaced with the object descriptions for the $X$-th in-context samples, and `<ic sampleX qY>` is replaced with the manually annotated $Y$th question of $X$th in-context sample. The target object caption is injected to the placeholder `<caption>`.

`<Q>` Answer from the following options: (A): Definitely Yes, (B): Possibly Yes, (C): Possibly No, (D): Definitely No, (E): Not sure. Do not write an introduction or summary. Only respond the letter.

Table 28. **The text prompt to input to the MLLM to verify the contents in the captions.** The placeholder `<Q>` is replaced with the input questions generated during the previous stage.

```
Your task is to generate a refined description of the objects based on the given questions about the objects in a list format. The answer is always no, so remove that information from the object description and re-generate the refined description. Do not change the irrelevant content as much as possible. Do not write an introduction or summary, and only write the description in JSON format.
=======
Description: <ic sample original caption>
Questions: [ <ic sample q1>, <ic sample q2>, ..., ]
Refined Description: <ic sample output caption>

======
Description: <original caption>
Questions: [ <q1>, <q2>, ..., ]
Refined Description:
```

Table 29. **The text prompt to refine the region-level captions.** We collect the questions whose answers are "D" in Tab. 28 and prompt the LLM to remove the contents in the questions from the original caption. One in-context sample is provided to improve the output quality. The placeholders `<ic sample original caption>`, `<ic sample qX>`, `<ic sample output caption>` are replaced with the object-level caption, the $X$-th question which includes the non-verified content, and the manually annotated refined caption, respectively. The target caption and non-verified questions are injected to `<original caption>` and `<qX>` placeholders, respectively.

```
You are a helpful assistant in summarizing the long description of the object into a single sentence or a short description. Output the JSON format, for example, [<object ID>]: <description>, ...
##Instructions
- DO NOT include any contents that have not appeared in the long descriptions.
- The short description should briefly cover the abstracted contents of the long description.
- The short description should be a single sentence, and the number of words had better be less than 20.

The long descriptions of objects (<id1>, <id2>, ...) are as follows: [<id1>]: <description1>, [<id2>]: <description2>, ...
```

Table 30. **The text prompt to create a summary of the objects' descriptions.** The placeholders `<idX>` and `<descriptionX>` are replaced with $X$-th object region indices and the corresponding detailed caption.

You are a helpful assistant in creating diverse multi-round conversation question-and-answer pairs from the given object descriptions. Your task is to design a conversation between you and a person asking about objects. The answers should be in a tone that a visual AI assistant is watching the video and answering the question. Output the JSON format with key ’Q’ and value ’A’, for example, `{`’QA_pairs’: `[{`’Q’: ’What is the object next to region `[0]`...’, ’A’: ’It is...’`},{`’Q’: ’Why this object is...’`}` ...`]}`.

Include diverse questions asking about the visual content of the video, including the object types, temporal information, object motions, actions, textures, spatial positions, etc. Only include questions that have definite answers:
(1) one can see the content in the video that the question asks about and can answer confidently;
(2) One can confidently determine from the video that it is not in the video.
Do not ask any questions that cannot be answered confidently.
Please follow the below instructions.
## Instructions
- Do not add additional information to the output answer from the input object descriptions. DO NOT ask about uncertain details.
- If you refer to a certain object from the list, USE region indices (`<id1>`, `<id2>`, ...) in the questions. DO NOT include the region indices in the answer. MUST generate several different QAs.

The object descriptions for `<id1>`, `<id2>`, ... are as below: `[<id1>]: <description1>`, `[<id2>]: <description2>`, ...

Table 31. **The text prompt to generate region-level conversation samples from object-level detailed captions.** The placeholders `<idX>` and `<descriptionX>` are replaced with $X$-th object region indices and the corresponding detailed caption.

- Provide a brief caption for the area indicated by `<region>`.
- Describe in a short phrase the content within the bounds of `<region>`.
- How would you succinctly caption the region highlighted by `<region>`?
- Summarize the scene or object present in the section marked by `<region>`.
- Can you give a concise description of what's depicted in `<region>`?
- Draft a short title for the video content enclosed by `<region>`.
- What brief caption would best describe the visual within `<region>`?
- Offer a succinct interpretation of the area pointed out by `<region>`.
- If you were to provide a short tagline for the content at `<region>`, what would it be?
- Give a one-liner description of the region demarcated by `<region>`.
- How would you encapsulate the essence of the segment labeled `<region>` in a few words?
- Characterize the content of the video portion specified by `<region>` briefly.
- Craft a mini headline for the visual element spotlighted by `<region>`.
- In a few words, how would you narrate the content found within `<region>`?
- Pen down a concise caption for the video section delineated by `<region>`.
- A short caption of region `<region>`:
- A short description of region `<region>`:
- A photo containing the region `<region>`:
- A region `<region>` that shows
- Write a short description for the region `<region>`
- Write a description for the region `<region>`
- Provide a description of what is presented in the region `<region>`.
- Briefly describe the content of the region `<region>`.
- Can you briefly explain what you see in the region `<region>`?
- Could you use a few words to describe what you perceive in the region `<region>`?
- Please provide a short depiction of the region `<region>`.
- Using language, provide a short account of the region `<region>`.
- Use a few words to illustrate what is happening in the region `<region>`.
- Provide an overview of what you see in the region `<region>`.
- Can you break down the main elements present in this region `<region>`?
- What are the key features or subjects captured in this region `<region>`?
- Summarize the primary components of this region `<region>`.
- Walk me through the different aspects of this region `<region>`.
- Highlight the main points of interest in this region `<region>`.
- What stands out to you the most in this region `<region>`?
- If you were to give a brief overview of this region `<region>`, what would you mention?
- List the primary objects or subjects you identify in this region `<region>`.
- Describe the first few things that catch your attention in this region `<region>`.
- How would you introduce this region `<region>` to someone who hasn't seen it?
- What are the defining characteristics of this region `<region>`?
- Give a concise description of the main content in this region `<region>`.
- If you were to caption this region `<region>`, what might you say?
- Describe the scene or setting depicted in this region `<region>`.
- Provide a concise label for the highlighted `<region>`.
- Offer a short depiction of what's enclosed within `<region>`.
- How would you briefly label the segment outlined by `<region>`?
- Summarize the contents of the space marked by `<region>`.
- Can you give a brief explanation of what's featured in `<region>`?
- Suggest a quick headline for the area depicted within `<region>`.
- What short caption would best capture the details of `<region>`?
- Provide a concise overview of the highlighted portion of `<region>`.
- If you were to summarize the content in `<region>`, what would it be?
- Offer a one-line explanation of what is shown in `<region>`.
- How would you briefly summarize the key points in `<region>`?
- Characterize the details of the video enclosed within `<region>` in a few words.
- Draft a small title for the area spotlighted in `<region>`.
- How would you narrate the content observed in `<region>` concisely?
- Write a brief caption for the section of the video enclosed by `<region>`.
- Create a short label for the visible content in `<region>`.
- Offer a quick description of the content shown in `<region>`.
- A snapshot of what `<region>` includes.
- Describe what's highlighted within `<region>`.
- Provide a quick summary of the contents present in `<region>`.
- What's the best way to describe the content within `<region>` briefly?
- What's shown in `<region>`?Please describe.
- How would you quickly explain what is featured in `<region>`?
- Provide a short summary of what's visible in `<region>`.
- Could you break down what you see in `<region>`?
- Can you briefly outline the main elements in `<region>`?
- Give a brief interpretation of what stands out in `<region>`.
- What are the key points of focus in this section `<region>`?
- Provide a concise rundown of what is visible in `<region>`.
- Describe the notable objects or subjects you can see in `<region>`.
- How would you introduce the details shown in `<region>` to someone new?
- List the main features or subjects captured in `<region>`.
- Summarize the primary components present in `<region>`.
- What catches your attention first in `<region>`?

Table 32. **The list of instructions for brief region description.**

- Describe in detail the object located at `<region>` in the video, including its appearance, style, actions, and any visible details.
- Provide a comprehensive description of the area marked by `<region>`, focusing on textures, colors, motions, and any notable features.
- Elaborate on the artwork shown in the region indicated by `<region>`, mentioning its color, appearance, size, style, and any standout features.
- Give a detailed analysis of the scene within the boundary of `<region>`, touching upon its components, ambiance, and any thematic expressions.
- Craft a thorough narrative about the piece of the video highlighted by `<region>`, from its aesthetic qualities to its possible historical context.
- Explain in depth the characteristics and attributes of the subject found in the segment tagged with `<region>`.
- Generate a long, detailed caption for the segment of the video at `<region>`, covering aspects such as its origin, material, and any symbolic meaning.
- Paint a vivid picture with words about the region at `<region>`, diving into the intricacies and nuances present in the area.
- Zoom in on the area indicated by `<region>` and describe every discernible detail, from texture and color to form and function.
- Offer an expanded description of the contents within the area marked by `<region>`, encompassing its color, appearance, size, style, and any remarkable features.
- Provide a detailed account of the object located at `<region>` in the video.
- Offer a full description of the object found in `<region>`.
- Describe comprehensively the object at `<region>`, including its physical traits, style, actions, and visible characteristics.
- Give an extensive description of the area labeled `<region>`, paying attention to textures, colors, movements, and notable details.
- Elaborate thoroughly on the artwork situated in the area indicated by `<region>`, noting its color, size, style, and key features.
- Analyze the scene within the confines of `<region>` in detail, considering its components, atmosphere, and any thematic elements.
- Develop a rich narrative for the part of the video highlighted in `<region>`, exploring its aesthetic aspects and potential historical significance.
- Explain the key attributes and features of the subject found in `<region>` with in-depth detail.
- Compose an extended caption for the video segment located at `<region>`, focusing on its origin, materials, and possible symbolic meaning.
- Describe vividly the region marked `<region>`, exploring the intricacies and nuances present in the area.
- Focus on the area highlighted by `<region>` and provide a meticulous description, covering its texture, color, shape, and function.
- Expand on the description of what's present in the area of `<region>`, addressing its colors, appearance, dimensions, style, and distinctive qualities.
- Give a precise and thorough breakdown of the object located at `<region>`, considering its visual and functional features.
- Delve deeply into the description of the contents at `<region>`, focusing on the interplay of textures, colors, movements, and standout aspects.
- Explore the artwork or object featured in `<region>` with a focus on its color, style, and size.
- Offer a detailed breakdown of the scene in `<region>`, discussing its components, mood, and any visible themes.
- Provide a rich account of the aesthetic qualities and potential historical or cultural relevance of the area highlighted by `<region>`.
- Explain the visible details of the subject in `<region>`, including its appearance, movements, and any distinguishing traits.
- Create a thorough, descriptive caption for the portion of the video located in `<region>`, touching on its materials, origin, and symbolic elements.
- Use vivid language to describe the region marked by `<region>`, diving into the subtle details and unique aspects of the area.
- Take a close look at `<region>` and describe in detail the features you observe, including texture, color, form, and function.
- Expand upon the visual description of what is contained within the marked area `<region>`, covering its appearance, size, and stylistic features.

Table 33. **The list of instructions for detailed region description.**