

UrBench: A Comprehensive Benchmark for Evaluating Large Multimodal Models in Multi-View Urban Scenarios

Anonymous submission

Appendix

In this appendix, we provide additional details that could not be included in the main paper due to limited space, which comprises the details about the benchmark and evaluation. We discuss:

- Additional statistics of UrBench.
- Additional benchmark curation details.
- Additional experiment details and results.
- Comprehensive case studies of UrBench.
- Limitations and future work.

More Benchmark Statistics

Image Statistics

As illustrated in Table 1, UrBench comprises approximately 9.98K images, including three view types: monocular-street-view, panoramic-street-view, and satellite-view. In some cases, a visual prompt in the form of a bounding box is utilized to indicate the object of questioning. Note that some of the images are utilized for both visual prompt and non-visual prompt questions. Consequently, there are three types of images in terms of visual prompt.

	Image	Test	Val
	Total	8783	1199
View Type	monocular-street-view	1582	195
	panoramic-street-view	3100	463
	satellite-view	4101	541
Visual Prompt	with visual prompt	2751	463
	with & without prompt	947	30
	without visual prompt	5085	706
Source	in-house	6082	887
	VIGOR	1007	117
	Cityscapes	628	72
	Mapillary Traffic Sign	447	53
	IM2GPS	619	70

Table 1: Detailed statistics of UrBench in terms of image view type, with or without visual prompt, and image source.

Benchmark Comparison

Table 2 compares remote sensing and urban benchmarks (Lobry et al. 2020; Wang et al. 2024b; Li, Ding, and Elhosseiny 2024; Feng et al. 2024), general benchmarks (Wang et al. 2024a; Yue et al. 2024; Liu et al. 2023) and our UrBench. UrBench outperforms most other benchmarks in terms of images per question, image views per question, and annotation source. Although the remote sensing benchmarks outperform in terms of question quantity, they have limited question quality as shown in Fig. 1 in the main paper. Note that for general benchmarks, if the total number of images of a view type is less than 10, that image view type will be considered unsupported.

Benchmark Curation Details

Data Collection Details

This subsection provides a comprehensive overview of the collection process for all panoramic street-view images, monocular street-view images, and satellite-view images, as well as the annotations from OpenStreetMap (OSM) and PLUTO within UrBench.

The panoramic street-view images of UrBench are obtained from Google Street View¹, and the satellite-view images are obtained from Google Earth² (Level 19). Based on the geographic coordinates of the collected street-view images, we cropped and partitioned the large satellite-view images into patches of size 512×512 , ensuring that each satellite-view image corresponds geographically to its corresponding street-view image. Specifically, the geographical coordinates of the center points of the satellite-view images align with the camera locations of the street-view images. Furthermore, the satellite-view images used in the City Retrieval task are obtained from Google Earth (Level 19), aligned with the latitude and longitude coordinates of the street-view images in the IM2GPS (Hays and Efros 2008) dataset.

The OSM annotation is obtained from OpenStreetMap³ and contains ground segmentation and semantic information for various urban features. Additionally, the number of floor

¹<https://svd360.istreetview.com/>

²<https://www.google.com/earth/about/versions/>

³<https://www.openstreetmap.org>

Dataset	Question Size	Single	Multiple	Sat V	Str V	Cross V	Anno	Repu	OSM	Open	MC	Other
RSVQA	> 1M	✓	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗
EarthVQA	> 100K	✓	✗	✓	✗	✗	✗	✓	✗	✓	✗	✗
VRSBench	> 100K	✓	✗	✓	✗	✗	✓	✓	✗	✓	✓	✓
CityBench	unknown	✓	✗	✓	✓	✗	✗	✗	✓	✓	✓	✓
MUIRBench	2.6K	✗	✓	✓	✓	✗	✓	✓	✗	✗	✓	✗
MMBench	3K	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓	✗
MMMU	11.5K	✓	✓	✗	✓	✗	✓	✗	✗	✓	✓	✗
UrBench	11.6K	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗

Table 2: The comparison of remote sensing and urban benchmarks, general benchmarks, and our UrBench. Sat V, Str V, and Cross V are the abbreviations for satellite-view, street-view, and satellite-street cross-view. Anno, Repu, and OSM are annotated, repurposed, and OpenStreetMap abbreviations. Open and MC mean open-ended and multiple-choice respectively. Other means other answer types are supported in some benchmarks.

annotations used in the cross-view floor counting sub-task of the Object Attribute Recognition task, as well as the land use type annotations used in the land use recognition sub-task, are sourced from PLUTO⁴ annotations. For each satellite-view image, all ground objects with OSM or PLUTO annotations that appear in the range of this image are grouped as the image’s OSM/PLUTO annotations.

Data Pre-processing Details

This subsection details the cross-view detection-matching algorithm and the annotation database. The cross-view detection-matching algorithm is used for providing building matching across the satellite-view and street-view images and for providing building bounding boxes in street-view images. As illustrated in Fig.5 in the main paper, with a given street-view image I_{st} , and a satellite-view image I_{sa} , we first use a trained Grounding DINO model to detect the buildings of I_{st} , thus obtaining a group of bounding boxes B_{st} . Secondly, we calculate the azimuthal range A_{st} of B_{st} using the ray-tracing algorithm. Meanwhile, we calculate the azimuthal range A_{sa} of the bounding boxes B_{sa} of I_{sa} . Finally, we match A_{st} and A_{sa} with an IoU threshold of 0.5, thus successfully match B_{st} and B_{sa} . To facilitate the answer extraction of the questions generated from the rule-based method, we calculate the values such as the bounding boxes, building footprint areas, building count, crosswalk count, etc., and organize them into a small database. Table 3 lists the data sources for all tasks in UrBench.

Annotation and QC Tool

Due to the discrepancy in perspective, the severe distortion at the top of a panoramic image, the capture time difference between the images, and the hallucination of LMMs, Q&As generated by rule-based and LMM-based methods are not always valid. Consequently, human examination is employed on these questions. As illustrated in Fig.1, LabelLLM⁵ is utilized for human annotation and quality control.

⁴<https://data.cityofnewyork.us/City-Government/Primary-Land-Use-Tax-Lot-Output-PLUTO-64uk-42ks>.

⁵<https://github.com/opendatalab/LabelLLM>

Task	Open	GE	GSV	OSM	PLUTO
OR	-	✓	✓	✗	✗
IR	VIGOR	✗	✗	✗	✗
CR	IM2GPS	✓	✗	✗	✗
CL	VIGOR	✗	✗	✗	✗
TSR	MTSD	✗	✗	✗	✗
RBR	VIGOR	✗	✗	✗	✗
VPR	Cityscapes	✗	✗	✗	✗
CO	-	✓	✗	✓	✗
SC	-	✓	✓	✓	✗
RU	-	✓	✓	✓	✗
SR	-	✓	✓	✓	✗
OM	-	✓	✓	✓	✗
OAR	-	✓	✓	✓	✓
OG	Cityscapes	✓	✗	✓	✗

Table 3: The image and annotation source of each task in UrBench. Tasks are grouped according to the evaluation dimension. Open, GE, and GSV are the abbreviations for the open dataset, Google Earth, and Google Street View. MTSD means Mappillary Traffic Sign Dataset.

Experiment Details & Results

Baseline Models

Our evaluations include 21 recent LMMs, including both closed-source and open-source models trained to receive multi-image and single-image inputs. We list the baseline models here as follows:

1. GPT (Achiam et al. 2023): We use the gpt-4o-2024-05-13 version for GPT-4o during our evaluations.
2. Claude (Anthropic 2024): We use the latest Claude-3.5-Sonnet version for the Claude model family.
3. Gemini (Reid et al. 2024): We use the latest Gemini-1.5-Flash to represent the Gemini model family. We note that the API of Gemini-1.5-Pro has yet to be released by the time of our evaluations.

Judge whether this is a valid question.
If it is, judge whether the answer is correct.

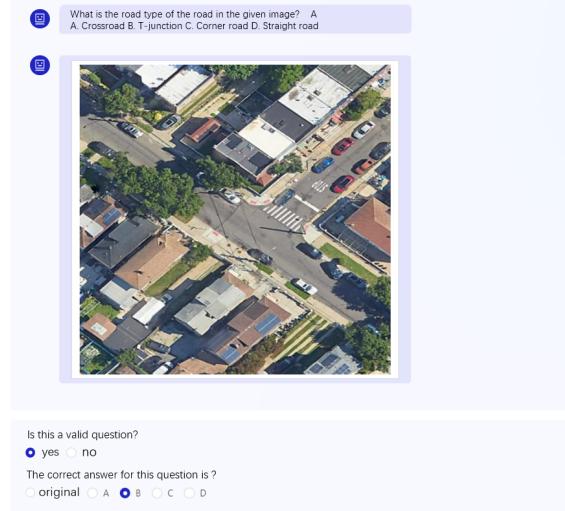


Figure 1: An example of the user interface of LabelLLM. This example illustrates the human verification of a question in the Road Understanding task.

4. TinyLLaVA (Zhou et al. 2024): TinyLLaVA is a small-scale version of the LLaVA-1.5 model family. We use the 1.1B model checkpoint released in huggingface for our evaluation.
5. LLaVA (Liu et al. 2024b): We include the LLaVA-NeXT (7B, 8B, 13B, and 34B) versions as well as the multi-image version: LLaVA-NeXT-Interleave.
6. InternVL2 (Chen et al. 2024): InternVL2 models across different scales are included: 2B, 4B, 8B, and 26B.
7. Mantis (Jiang et al. 2024): We include two strong Mantis checkpoints: Idefics2 and LLaMA3-SigLIP.
8. Idefics2 (Laurençon et al. 2024): We consider the fine-tuned version of Idefics2-8b-base for evaluations.
9. XComposer (Zhang et al. 2023): We use the 4K checkpoint of XComposer2 that can process high-resolution images.
10. InstructBLIP (Li et al. 2023): We utilize the 7B version of InstructBLIP, which features EVA-CLIP as the visual backbone and Vicuna-7B as the language backbone.
11. VILA-1.5 (Lin et al. 2024): We consider the 13B checkpoints and the 40B checkpoints for VILA-1.5 for our evaluations.

Evaluation Protocols

We follow standard setups in MMMU (Yue et al. 2024) and use rule-based methods to extract answers from LMMs' responses. For the multi-choice questions, if no candidate options are extracted from the responses, we randomly assign an option for this response. For the open-ended questions, we use regular expression to match the correct answers with models' responses.

Evaluation Results of Model Scaling

We present models' overall performances across different scales in Fig. 2. Our evaluation results show that the performance of InternVL family (Chen et al. 2024), VILA family (Lin et al. 2024) and LLaVA family (Liu et al. 2024a) approximately increases with model sizes.

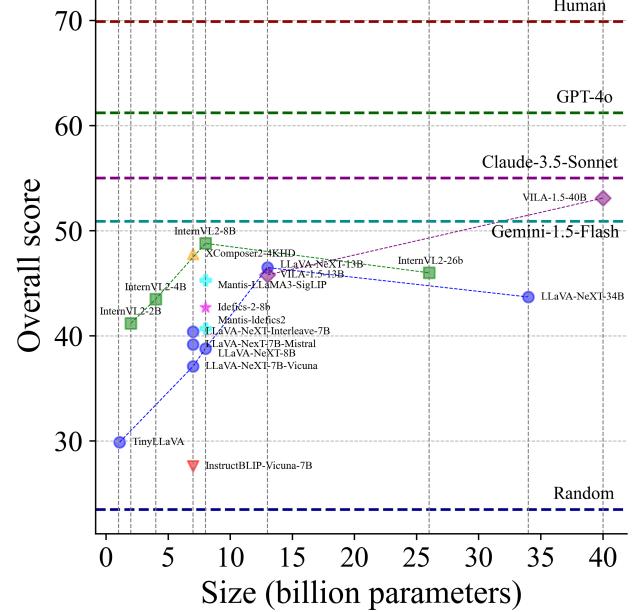


Figure 2: Model performances across different scales.

Evaluation Results of Different Views

Table 4 and Table 5 present our detailed evaluation results across different views. In the main paper, we only include the overall performance of different tasks. However, each task can be further categorized according to their respective views. We report the results of Geo-Localization and Object Understanding in Table 4 and Scene Reasoning and Scene Understanding in Table 5.

Case Study

In this section, we provide some question cases of different UrBench tasks and the responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet (Fig.3 - Fig.28). Incorrect descriptions in responses are marked in red.

Limitations & Future Work

There are several limitations to our work. First, the satellite-view and street-view images are not captured simultaneously, so the non-static ground objects in the image pair, such as moving cars, are not exactly paired up between views. Second, our work focuses primarily on 2D image understanding of the current LMMs. Future work could extend our work by including more modalities, such as video, 3D point clouds, audio, etc.

Model	Geo-Localization							Object Understanding					
	CR			IR	CL	OR	OM	OG			OAR		
	satellite	street	cross	cross	cross	cross	cross	satellite	street	satellite	street	cross	
GPT-4o	66.1	84.8	86.4	85.9	<u>35.3</u>	<u>30.7</u>	28.0	<u>62.3</u>	36.0	43.8	58.9	48.8	
Gemini-1.5-Flash	59.3	71.6	77.8	25.9	<u>25.9</u>	24.0	22.0	<u>55.9</u>	27.9	36.2	51.5	37.9	
Claude-3.5-Sonnet	59.2	<u>74.2</u>	<u>83.3</u>	55.8	30.8	33.3	22.0	81.5	48.0	50.9	43.2	<u>43.7</u>	
TinyLLaVA	48.9	53.8	53.0	23.2	24.7	27.9	22.9	54.6	32.0	26.5	31.9	7.8	
InternVL2-2B	40.3	56.4	53.9	23.8	31.9	29.0	25.5	29.1	31.1	39.6	50.5	35.6	
InternVL2-4B	43.9	62.5	58.3	24.2	27.4	23.1	30.6	34.4	27.3	33.3	50.9	38.2	
XComposer-4KHD	53.6	66.2	65.7	26.0	27.5	25.9	16.2	53.5	35.2	45.2	64.9	38.3	
LLaVA-NeXT-7B-Mistral	45.7	52.5	56.5	25.9	24.2	24.0	43.9	36.6	30.8	26.3	31.9	31.0	
LLaVA-NeXT-7B-Vicuna	46.1	51.1	56.5	24.6	27.2	23.3	31.8	28.9	26.1	23.1	28.9	38.3	
InstructBLIP-Vicuna-7B	35.1	44.1	41.9	25.7	25.4	27.1	22.9	25.3	20.9	19.6	21.3	14.0	
LLaVA-NeXT-8B	48.5	54.4	60.2	27.0	27.8	26.0	<u>39.8</u>	44.7	39.7	23.1	26.1	32.5	
InternVL2-8B	37.2	61.6	53.5	26.6	31.8	25.2	<u>30.3</u>	32.2	31.8	34.2	49.3	40.9	
Idefics-2-8B	57.9	67.5	70.9	23.8	26.0	24.1	24.8	15.4	24.9	23.6	37.2	24.3	
LLaVA-NeXT-13B	42.6	54.1	59.1	24.5	27.7	26.7	31.8	48.0	25.9	24.3	29.6	25.7	
InternVL2-26B	53.1	66.4	64.4	23.0	32.3	24.7	26.1	39.2	36.0	43.4	<u>64.5</u>	42.7	
LLaVA-NeXT-34B	47.9	65.1	62.1	26.0	28.5	27.8	28.7	47.3	36.0	24.7	33.0	19.0	
LLaVA-NeXT-Interleave-7B	54.4	58.3	60.8	41.6	27.6	25.5	37.3	38.8	36.7	<u>45.6</u>	47.9	35.9	
Mantis-LLaMA3-SigLIP	63.3	67.4	70.3	32.4	27.0	27.2	<u>25.2</u>	44.0	27.6	40.0	37.4	38.6	
Mantis-Idefics2	<u>64.4</u>	70.3	72.1	29.9	27.0	25.7	29.6	52.7	27.8	34.6	41.7	32.8	
VILA-1.5-13B	53.9	65.8	68.2	33.7	28.6	24.1	25.8	29.3	34.2	38.2	47.2	33.3	
VILA-1.5-40B	60.7	<u>74.2</u>	75.4	<u>62.5</u>	36.8	27.9	34.1	59.7	<u>40.6</u>	39.3	50.7	33.6	

Table 4: The LMMs experimental results of the different views of tasks under Geo-Localization and Object Understanding evaluation dimensions. The maximum value and the next largest value of each task are indicated by the **bold** and underlined text, respectively. Task names are abbreviated for brevity.

Model	Scene Reasoning						Scene Understanding						
	RBR			TSR		VPR	SR			RU		CO	SC
	street	street	street	satellite	street	cross	satellite	street	satellite	street	satellite	street	
GPT-4o	79.6	68.2	77.9	61.7	61.3	71.9	42.4	86.4	40.1	<u>79.2</u>	78.6		
Gemini-1.5-Flash	77.8	75.8	69.8	59.6	54.7	59.5	40.5	87.4	29.1	68.3	65.5		
Claude-3.5-Sonnet	73.7	37.7	66.7	<u>64.7</u>	52.6	40.0	54.7	67.7	48.0	84.0	70.0		
TinyLLaVA	40.3	32.7	48.6	17.5	7.2	1.1	3.9	47.4	9.5	27.4	28.6		
InternVL2-2B	64.8	45.9	54.5	47.5	45.9	50.3	20.2	52.6	28.6	30.8	27.7		
InternVL2-4B	73.0	56.6	62.6	51.9	51.9	53.0	36.6	52.6	22.1	36.1	51.8		
XComposer-4KHD	75.5	62.6	60.8	54.1	52.5	60.0	<u>49.0</u>	53.5	35.8	26.4	44.6		
LLaVA-NeXT-7B-Mistral	59.2	42.7	45.5	55.2	50.8	60.5	38.8	52.6	34.1	28.6	27.7		
LLaVA-NeXT-7B-Vicuna	49.6	51.5	54.5	56.3	51.4	60.5	28.5	52.6	34.3	25.2	28.6		
InstructBLIP-Vicuna-7B	30.7	25.7	29.3	36.1	35.9	27.0	14.4	0.0	20.4	27.6	29.5		
LLaVA-NeXT-8B	55.2	52.8	58.6	60.1	48.6	58.4	40.6	52.6	37.3	23.0	27.7		
InternVL2-8B	74.9	54.8	62.6	54.6	50.3	54.1	37.6	70.6	27.4	49.6	58.0		
Idefics2-8B	64.7	60.4	42.8	50.8	51.4	54.6	34.6	52.6	22.9	28.8	24.1		
LLaVA-NeXT-13B	54.0	52.1	52.3	54.6	52.5	54.6	41.5	52.6	34.1	23.5	31.3		
InternVL2-26B	<u>77.9</u>	63.8	<u>71.2</u>	68.3	<u>58.0</u>	<u>68.6</u>	32.1	53.1	<u>43.0</u>	49.4	64.3		
LLaVA-NeXT-34B	65.6	59.3	56.8	58.5	54.1	62.2	41.1	52.6	25.9	49.6	69.6		
LLaVA-NeXT-Interleave-7B	65.8	55.9	63.1	55.2	45.9	56.8	38.9	33.8	27.4	46.2	56.3		
Mantis-LLaMA3-SigLIP	67.6	41.6	57.7	62.8	51.9	62.7	32.4	23.9	28.2	49.6	62.5		
Mantis-Idefics2	68.9	50.6	56.3	53.0	48.6	49.2	33.6	59.6	30.1	52.8	67.9		
VILA-1.5-13B	66.3	43.8	46.4	50.3	43.6	49.2	31.0	54.8	21.6	46.7	55.4		
VILA-1.5-40B	76.4	55.5	61.3	59.0	48.1	53.5	35.6	70.2	32.1	64.6	<u>74.1</u>		

Table 5: The LMMs experimental results of the different views of tasks under Scene Reasoning and Scene Understanding evaluation dimensions. The maximum value and the next largest value of each task are indicated by the **bold** and underlined text, respectively. Task names are abbreviated for brevity.

City Retrieval

Question:

Where is this picture possibly taken?

Option:

- (A) Palo Alto
- (B) Mountain View
- (C) Santa Clara
- (D) San Jose

**Answer: D**

GPT-4o: C. The image shows the Adobe Lodge at Santa Clara University. The structure and surrounding area are characteristic of historical buildings found at the university. The Adobe Lodge is well-known for its distinctive architecture, stucco walls, and wooden beams, which this photo displays.



VILA1.5-40B: D. San Jose.



Claude-3.5-Sonnet: D. San Jose

This image likely depicts a historic adobe building in San Jose, possibly at History Park or a similar historical site. The architectural style, with its adobe walls, wooden porch, and rustic features, is characteristic of early California structures often preserved in San Jose's historical areas.

Figure 3: A question case of the City Retrieval task in street-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

City Retrieval

Question:

Where is this picture possibly taken?

Option:

- (A) Chicago
- (B) St. Louis
- (C) Cincinnati
- (D) Madison

**Answer: A**

GPT-4o: A. The image shows an aerial view of a stadium and its surrounding area. The distinctive architecture and the presence of the "Gallagher" branding are clues. Gallagher is associated with the naming rights for the Gallagher Way, an entertainment district adjacent to Wrigley Field, the home stadium of the Chicago Cubs in Chicago. The neighborhood layout and the patterns of the buildings are also consistent with the Wrigleyville area in Chicago.



VILA1.5-40B: A. The answer is A.



Claude-3.5-Sonnet: B. St. Louis

This image is likely taken in St. Louis due to the visible "Gallagher" sign, which is a reference to the Enterprise Center (formerly Scottrade Center), home of the St. Louis Blues NHL team. The aerial view shows a mix of urban buildings and structures consistent with a major city like St. Louis.

Figure 4: A question case of the City Retrieval task in satellite-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

City Retrieval

Question:

Where is this picture possibly taken?

Option:

- (A) Berlin
- (B) Hamburg
- (C) Oslo
- (D) Aarhus

**Answer: D**

GPT-4o: D. The second image shows a public square with a red-brick building featuring "De Augustin Nygaards Minde" inscribed on it. This building is a well-known landmark in Aarhus, Denmark. The architectural style and the visible signage also match the typical appearance of some parts of Aarhus. The first image of the aerial view complements this by showing a similar urban layout and architectural style distinct to this Danish city. Therefore, the most likely location for these images is Aarhus.



VILA1.5-40B: C. Oslo.



Claude-3.5-Sonnet: D. Aarhus

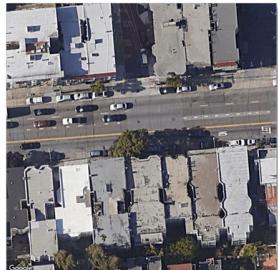
The architecture and urban layout in the second image are characteristic of Danish cities, particularly Aarhus. The cobblestone square, the mix of traditional brick buildings and colorful facades, and the presence of an H&M store are typical of Aarhus city center.

Figure 5: A question case of the City Retrieval task in cross-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

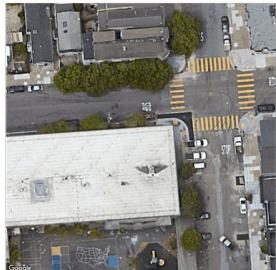
Image Retrieval

Question:

Which of the following four images correctly represents the satellite view of the image?



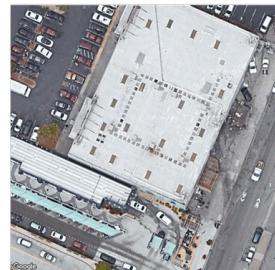
(A)



(B)



(C)



(D)

Answer: D



GPT-4o: D. In the street view image, you can see a distinctive building labelled "United Rentals" with a parking area and equipment in front of it. The surroundings are industrial with large buildings and wide streets typical of commercial areas. Image D accurately depicts a large building with industrial characteristics, along with a significant number of parked vehicles and equipment, aligning with the elements visible in the street view image.



VILA-1.5-40B: B. The answer is B.



Claude-3.5-Sonnet: D.

The street view shows a wide intersection with a large white building and parking areas, which matches the satellite view in image D. This aerial view shows the same large white building, parking lots, and street layout visible in the street-level image.

Figure 6: A question case of the Image Retrieval task with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Camera Localization

Question:

Compare the key landmarks or distinctive features in both images to determine which quadrant (1, 2, 3, 4) of the satellite image shows the shooting location of the street view image.



Answer: 4



GPT-4o: 3. By comparing these significant features in both images, it is evident that the shooting location of the street view image corresponds to Quadrant 3 on the satellite image.



VILA1.5-40B: 2. The street view image shows a road with a blue car parked on the side and a few trees. The satellite image is divided into four quadrants, and the exact shooting point of the street view image corresponds to a specific location on the satellite view image. By comparing the key landmarks or distinctive features in both images, it can be determined that quadrant 2 of the satellite image shows the shooting location of the street view image.



Claude-3.5-Sonnet:

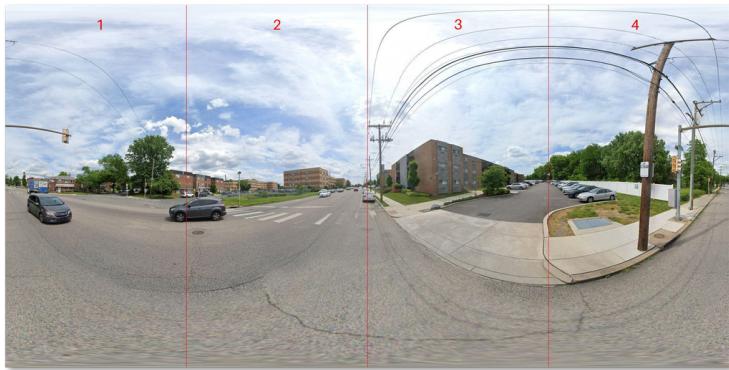
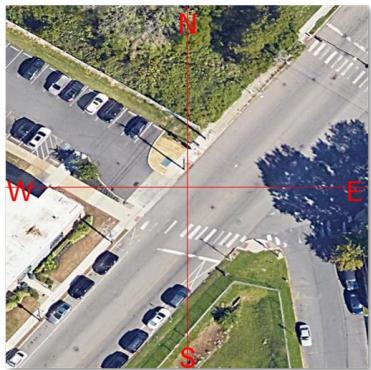
The street view image appears to be taken from quadrant 3 of the satellite image. This can be determined by the curved road visible in both images and the pattern of trees and open spaces that match between the street view and the lower left quadrant of the satellite image.

Figure 7: A question case of the Camera Localization task with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Orientation

Question:

Which section of the panoramic street view image corresponds to the true north direction?



Answer: 4



GPT-4o: 4. Since the satellite image marks the parking lot on the northwest, section 4 must correspond to looking directly north.



VILA1.5-40B: 2. In the panoramic street view image, section 2 is at the top. So, the section that corresponds to the true north direction is section 2.



Claude-3.5-Sonnet: 3

Explanation: The satellite image shows that North is toward the top-right corner, which aligns with the direction of the large apartment buildings visible in section 3 of the panoramic street view image.

Figure 8: A question case of the Orientation task with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Object Grounding

Question:

Identify the extent of the object in the description below in the format of [xmin, ymin, xmax, ymax].

Description: A person wearing a white shirt and dark pants is standing on the sidewalk near the center of the image, holding a bag in their left hand.

Option:

- (A)[219, 324, 274, 507]
- (B)[677, 752, 730, 929]
- (C)[876, 369, 928, 543]
- (D)[915, 542, 972, 731]

**Answer: C**

GPT-4o: C. [876, 369, 928, 543].



VILA-1.5-40B: A. The answer is A.



Claude-3.5-Sonnet: C. [876, 369, 928, 543]

Explanation: This bounding box most accurately encompasses the person described in the image. The person is standing near the center of the image on the sidewalk, and this box captures their position and approximate size relative to the scene.

Figure 9: A question case of the Object Grounding task in street-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Object Grounding

Question:

Identify the extent of the object in the description below in the format of [xmin, ymin, xmax, ymax].

Description: The building with a light brown roof is located at the bottom-left corner of the image, adjacent to a large grassy field and near a black-roofed building.

Option:

- (A)[340, 250, 483, 333]
- (B)[0, 38, 136, 116]
- (C)[281, 425, 394, 491]
- (D)[59, 433, 196, 512]

**Answer:** D

GPT-4o: B. The coordinates [0, 38, 136, 116] correctly identify the extent of the building with the light brown roof located at the bottom-left corner of the image.



VILA-1.5-40B: B.



Claude-3.5-Sonnet: D. [59, 433, 196, 512]

This coordinate range accurately captures the location of the light brown-roofed building in the bottom-left corner of the image. The building is situated adjacent to the parking lot and near the edge of the frame, matching the description provided.

Figure 10: A question case of the Object Grounding task in satellite-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Object Attribute Recognition

Question:

What is the land use type of the building framed in red in the given image?

Option:

- (A) Residential
- (B) Mixture of residential and commercial
- (C) Industrial, manufacturing, public facilities or institutions
- (D) Mixture of commercial and office
- (E) Others

**Answer: B**

GPT-4o: C. The building framed in red has retail stores on the ground floor, including a business named "FELISSA" and "LCT #1," which suggests commercial use. The upper floors appear to be used for offices, as indicated by the windows, which are typical of office spaces rather than residential units.



VILA-1.5-40B: B. The land use type of the building framed in red in the given image is a mixture of commercial and office.



Claude-3.5-Sonnet: B. mixture of residential and commercial.

The building framed in red appears to be a multi-story structure with commercial storefronts on the ground level and likely residential units on the upper floors. This type of mixed-use building is common in urban areas, combining retail or service businesses at street level with apartments or offices above.

Figure 11: A question case of the Object Attribute Recognition task in street-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Object Attribute Recognition

Question:

What is the land use type of the building framed in red in the given image?

Option:

- (A) Residential
- (B) Mixture of residential and commercial
- (C) Industrial, manufacturing, public facilities or institutions
- (D) Mixture of commercial and office
- (E) Others

**Answer: D**

GPT-4o: D. The building framed in red in the satellite image appears to have features typical of industrial or institutional structures, such as large rooftop installations, a layout that suggests large interior spaces, and the presence of service areas, consistent with manufacturing, public facilities, or institutional uses rather than residential or commercial functions.



VILA-1.5-40B: D. The land use type of the building framed in red in the given image is industrial, manufacturing, public facilities, or institutions.



Claude-3.5-Sonnet: D. industrial, manufacturing, public facilities, or institutions.
The building framed in red appears to be a flat-roofed structure with industrial or institutional characteristics, such as HVAC units visible on the roof. Its size and layout are not typical of residential or commercial buildings, suggesting it's more likely an industrial or institutional facility.

Figure 12: A question case of the Object Attribute Recognition task in satellite-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Object Attribute Recognition

Question:

What is the land use type of the building framed in red in the given image?

Option:

- (A) Residential
- (B) Mixture of residential and commercial
- (C) Industrial, manufacturing, public facilities or institutions
- (D) Mixture of commercial and office
- (E) Others

**Answer: A**

GPT-4o: A. The building framed in red is primarily used for residential purposes, as indicated by its design and structure, which includes features typical of houses such as a porch, windows, and a driveway. The surrounding environment in both the satellite and street view images also depicts a neighborhood with similar residential buildings.



VILA-1.5-40B: A. The first image is a satellite image, and the second image is a street view image. Both images show a building framed in red. The building is surrounded by other buildings, and there are cars parked on the street. The street view image shows that the building has a front yard and a sidewalk in front of it. These details suggest that the building is likely a residential building.



Claude-3.5-Sonnet: A. residential

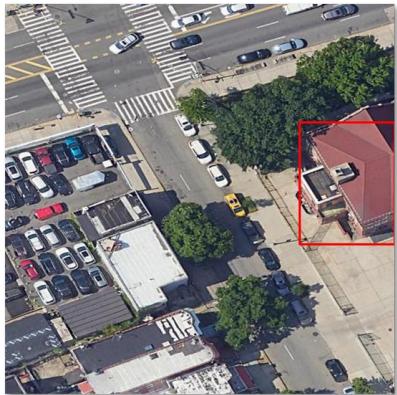
The framed building appears to be a typical residential house in a neighborhood of similar structures. Its architecture and setting among other houses strongly indicate it is a residential dwelling.

Figure 13: A question case of the Object Attribute Recognition task in cross-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Object Attribute Recognition

Question:

Provide the number of floors of the building.



Answer: 3



GPT-4o: 2. To locate the highlighted building, I compared the shape and relative size of the rooftops in the satellite image with the corresponding structures visible in the panoramic street view image. The building highlighted in red in the satellite image is situated at a corner with its rooftop showing a pointed or triangular outline. This matches the brick building on the left side of the panoramic image. **Observing its front view in the street view, this building has 2 floors.**



VILA-1.5-40B: 3. The building highlighted with a red box in the satellite image is a three-story building.



Claude-3.5-Sonnet:

The building highlighted in the satellite image appears to be the brick structure on the left side of the panoramic street view. Based on its visible facade, it has 3 floors, including what looks like a ground floor and two upper stories.

Figure 14: A question case of the Object Attribute Recognition task in cross-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Role-Based Reasoning

Question:

As a tourist visiting this area, I want to try the classic American fast food. Which store should I go to?

Options:

- (A) Five Guys
- (B) Burger King
- (C) McDonald's
- (D) Wendy's



Answer: A



GPT-4o: A. Five Guys

Explanation: In the image, "Five Guys" is clearly visible on one of the storefront signs, indicating that this fast-food restaurant is present in the location shown. No signs for Burger King, McDonald's, or Wendy's are seen.



VILA-1.5-40B: C. The image shows a busy city street with tall buildings and a clear blue sky. There are cars parked on the side of the road, and people can be seen walking on the sidewalk. The street is lined with trees, and there are traffic lights at the intersection. **The McDonald's restaurant is visible on the right side of the image.**



Claude-3.5-Sonnet: A. Five Guys

In the image, you can see a Five Guys restaurant sign on the left side of the street. This is the only fast-food restaurant clearly visible among the options provided in the urban street scene.

Figure 15: A question case of the Role-Based Reasoning task with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Traffic Sign Reasoning

Question:

What should I be cautious of according to the traffic sign in the red bounding box?

Option:

- (A) Sharp curve
- (B) Steep ascent
- (C) Roadwork
- (D) School crossing



Answer: B



GPT-4o: B. The traffic sign in box 0 depicts a truck on an incline, which indicates a steep ascent or descent. The accompanying text "USE LOW GEAR" reinforces this by advising drivers to switch to a lower gear, a common practice when driving on steep slopes to maintain control of the vehicle.



VILA-1.5-40B: A. Sharp curve



Claude-3.5-Sonnet: A. Sharp curve

The traffic sign in box 0 shows a downward-pointing arrow on an inclined surface, indicating a steep descent ahead. The sign below it reads \"USE LOW GEAR\" in English, which is typically advised for navigating sharp curves or steep descents safely.

Figure 16: A question case of the Traffic Sign Reasoning task with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Visual Prompt Reasoning

Question:

What is the primary function of the vehicle inside the bounding box?

Option:

- (A) Emergency medical services
- (B) Police patrol
- (C) Firefighting
- (D) Taxi service

**Answer: B**

GPT-4o: B. The vehicle inside the bounding box has the word "Polizei" written on it, which is German for "police." This indicates that the primary function of the vehicle is police patrol.



VILA-1.5-40B: A. The primary function of the vehicle inside the bounding box is emergency medical services.



Claude-3.5-Sonnet: D. Taxi service

The vehicle inside the red bounding box appears to be a white van, which is a common type of vehicle used for taxi or shuttle services in many cities. Its position on the road and general appearance suggest it's likely operating as a taxi or transport service rather than an emergency or specialized vehicle.

Figure 17: A question case of the Visual Prompt Reasoning task with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Scene Recognition

Question:

Which class of object appears in this image?

Choose from classes: bus station, fountain, graveyard, memorial or monument, motorway, parking, pitch, rail, residential, stadium, swimming pool, tower.

**Answer:** Parking

GPT-4o: Parking. The image predominantly shows areas used for parking vehicles. To the right side of the image, there is a fenced area with numerous parked cars, which is characteristic of a parking lot. Additionally, there are several cars parked along the street on both sides, further supporting the classification.



VILA-1.5-40B: Residential. The answer is residential.



Claude-3.5-Sonnet: Parking

Explanation: The image shows a street with multiple parked cars and a visible parking lot on the right side of the image. This clearly indicates the presence of parking facilities in the urban scene.

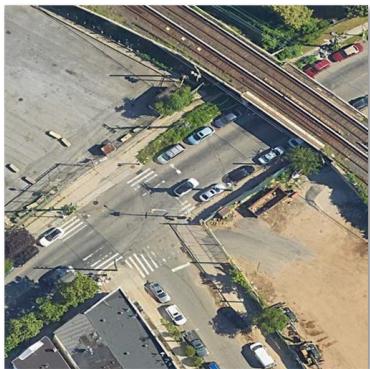
Figure 18: A question case of the Scene Recognition task in street-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Scene Recognition

Question:

Which class of object appears in this image?

Choose from classes: bus station, fountain, graveyard, memorial or monument, motorway, parking, pitch, rail, residential, stadium, swimming pool, tower.



Answer: Rail



GPT-4o: Residential. The satellite image shows a neighborhood with multiple houses, trees, cars parked on driveways and streets, and a building with solar panels on the roof. These features are typically associated with a residential area, which matches the "residential" class from the given options. There are no visible elements indicative of a bus station, fountain, graveyard, memorial or monument, motorway, parking lot, rail, stadium, swimming pool, or tower.



VILA-1.5-40B: Residential. The image is a bird's eye view of a neighborhood with houses, trees, and a parking lot. There are no other objects in the image that match the given classes..



Claude-3.5-Sonnet: Parking

This image shows a satellite view of a neighborhood area with visible parking spaces near buildings. The presence of parked cars and open paved areas for vehicle parking is evident, making "parking" the most appropriate class from the given options.

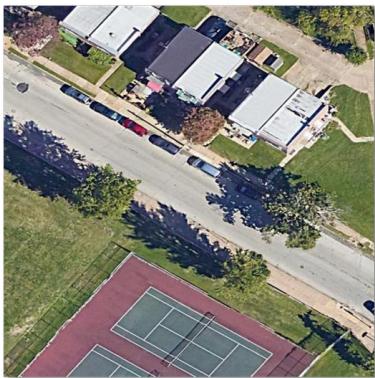
Figure 19: A question case of the Scene Recognition task in satellite-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Scene Recognition

Question:

Which class of object appears in this image?

Choose from classes: bus station, fountain, graveyard, memorial or monument, motorway, parking, pitch, rail, residential, stadium, swimming pool, tower.



Answer: Stadium



GPT-4o: Pitch. In both images, **there is a tennis court visible which falls under the category of a pitch.** Therefore, the correct class according to the provided options is "pitch." While there are residential buildings visible, the presence of the tennis court takes priority over the residential classification.



VILA-1.5-40B: Residential. The class of object that appears in this image is residential.



Claude-3.5-Sonnet: Residential

Explanation: The image shows a typical residential neighborhood with brick houses, parked cars, and a street view. **While there is a fenced area that could be a pitch or field, the dominant features are clearly residential in nature.**

Figure 20: A question case of Scene Recognition task in cross-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Road Understanding

Question:

What is the road type of the road in the given image?

Options:

- (A) Crossroad
- (B) Corner road
- (C) T-junction
- (D) Straight road

**Answer: C**

GPT-4o: Straight road. The image shows an aerial view of a road that continues in a straight direction without any significant turns, corners, or junctions. The road appears to have marked lanes, pedestrian crossings, and visible buildings on either side, indicating that it is a straight road.



VILA-1.5-40B: Crossroad. The road type of the road in the given image is a crossroad.



Claude-3.5-Sonnet: A. crossroad

The image clearly shows a four-way intersection where two roads meet perpendicular to each other, forming a crossroad. This is evident from the visible traffic lanes, crosswalks, and the arrangement of buildings and parking lots around the intersection.

Figure 21: A question case of the Road Understanding task in satellite-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Road Understanding

Question:

Are the following four images captured successively on the same road?



(A)



(B)



(C)



(D)

Answer: No.



GPT-4o: No.



VILA-1.5-40B: Yes. The images are captured successively on the same road.



Claude-3.5-Sonnet: Yes.

The second, third, and fourth images appear to show successive views of the same area around Radio City Music Hall in New York City, with the same distinctive buildings visible from different angles. The first image shows a different location.

Figure 22: A question case of the Road Understanding task in street-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Counting

Question:

How many printed white or yellow crosswalks are shown in this image?



Answer: 5



GPT-4o: 3. Upon examining the satellite image, **there are three areas where white lines forming crosswalks can be seen**: One crosswalk at the top left corner where two roads meet at a crossing. Another crosswalk is slightly below and to the right of the first one, forming a parallel set of white lines across the road. The third crosswalk is located towards the top center-right part of the image where another two roads intersect, showing white lines creating a crossing route for pedestrians.



VILA-1.5-40B: 3. **There are three printed white or yellow crosswalks shown in this image.**



Claude-3.5-Sonnet: Answer: There are 4 printed crosswalks visible in this image.

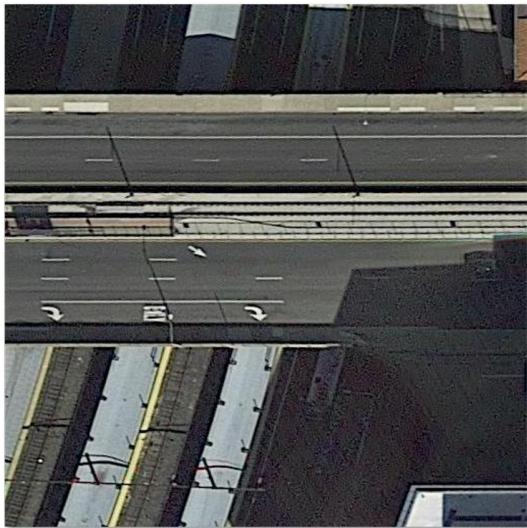
Explanation: The satellite image shows an urban intersection with clearly visible white crosswalk markings on the streets. **These crosswalks appear at the corners of the intersection, creating four distinct pedestrian crossing areas.**

Figure 23: A question case of the Counting task in satellite-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Counting

Question:

How many printed white or yellow crosswalks are shown in this image?

**Answer:** 0

GPT-4o: 1. The crosswalk can be seen on the street near the center of the image, consisting of horizontal white stripes across the road.



VILA-1.5-40B: 3. There are 3 printed white or yellow crosswalks shown in this image.



Claude-3.5-Sonnet: 0 (zero) crosswalks.

The image shows a bird's-eye view of a multi-lane highway or freeway with railway tracks running parallel. There are no visible crosswalks in this image, which is consistent with high-speed road infrastructure where pedestrian crossings would be unsafe and impractical.

Figure 24: A question case of the Counting task in satellite-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Scene Comparison

Question:

Which image contains the most buildings?



(A)



(B)



(C)



(D)

Answer: D

GPT-4o: B. By counting the distinct buildings in each image, **Image B has at least six visible buildings, which is more than the other images.**



VILA-1.5-40B: B. This image shows a residential area with multiple houses and a few cars parked on the streets. The other images also depict buildings, **but they are either commercial or industrial structures, and there are fewer buildings in each of them.**



Claude-3.5-Sonnet: B. Image B contains the most buildings.

Explanation: Image B shows an aerial view of a neighborhood with multiple buildings, including several houses and larger structures. **The other images show fewer distinct buildings or focus more on individual structures and natural areas.**

Figure 25: A question case of the Scene Comparison task in satellite-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Scene Comparison

Question:

Which image contains the building with the largest plant cover?



(A)



(B)



(C)



(D)

Answer: D



GPT-4o: D. Image D has the most plant cover. It contains numerous large trees covering large portions of the area, whereas the other images (A, B, and C) have fewer and smaller trees that cover significantly smaller areas.



VILA-1.5-40B: D. The image D shows a residential area with a large number of trees surrounding the houses and streets. The trees are lush and green, indicating that they are well-maintained and healthy. The other images, A, B, and C, also show residential areas with some trees, but the number of trees is significantly less than in image D. The trees in these images are also smaller and less dense, suggesting that they may be younger or less well-maintained.



Claude-3.5-Sonnet: D.

Image D has the most plant cover, with numerous large trees dominating the scene and creating a dense canopy over much of the area. The other images have significantly less vegetation visible, mostly limited to scattered trees or small patches of greenery.

Figure 26: A question case of the Scene Comparison task in satellite-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

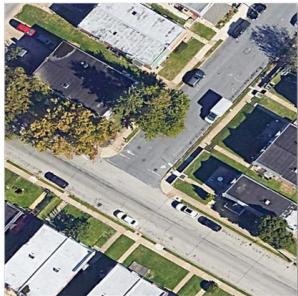
Scene Comparison

Question:

Which image contains the building with the largest area?



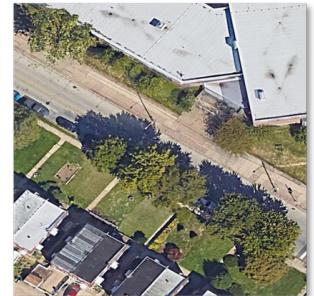
(A)



(B)



(C)



(D)

Answer: D



GPT-4o: D. Image D shows a building with a significantly larger roof area compared to the other images. The building in D occupies a large portion of the visible area in the image, and its roof extends over a much larger surface than the buildings shown in images A, B, and C.



VILA-1.5-40B: D. This image shows a large building with a gray roof, surrounded by smaller buildings and a few trees. The building's size is noticeable compared to the other structures in the image.



Claude-3.5-Sonnet: D.

The building in image D appears to be a large commercial or industrial structure with an expansive flat roof, covering significantly more area than the residential buildings seen in the other images. Its size and shape suggest it's likely the largest building among the options presented.

Figure 27: A question case of the Scene Comparison task in satellite-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

Scene Comparison

Question:

Which image contains the most vehicles?



(A)



(B)



(C)



(D)

Answer: B



GPT-4o: B. Image A shows one visible car parked by the side of the road. Image B has several vehicles, including at least a truck and multiple cars parked along the road and on the far side. Image C shows two visible vehicles. Image D contains one or two vehicles visible.



VILA-1.5-40B: The image that contains the most vehicles is **The one with the street view of a city.**



Claude-3.5-Sonnet: B.

Image B contains the most vehicles. It shows several cars parked along the street and in the distance, as well as a truck on the left side of the image, totaling more visible vehicles than any of the other images.

Figure 28: A question case of the Scene Comparison task in street-view with responses from GPT-4o, VILA-1.5-40B, and Claude-3.5-Sonnet.

References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic.
- Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.
- Feng, J.; Zhang, J.; Yan, J.; Zhang, X.; Ouyang, T.; Liu, T.; Du, Y.; Guo, S.; and Li, Y. 2024. CityBench: Evaluating the Capabilities of Large Language Model as World Model. *arXiv preprint arXiv:2406.13945*.
- Hays, J.; and Efros, A. A. 2008. IM2GPS: estimating geographic information from a single image. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.
- Jiang, D.; He, X.; Zeng, H.; Wei, C.; Ku, M.; Liu, Q.; and Chen, W. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.
- Laurençon, H.; Tronchon, L.; Cord, M.; and Sanh, V. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, X.; Ding, J.; and Elhoseiny, M. 2024. VRSBench: A Versatile Vision-Language Benchmark Dataset for Remote Sensing Image Understanding. *arXiv preprint arXiv:2406.12384*.
- Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26689–26699.
- Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.
- Lobry, S.; Marcos, D.; Murray, J.; and Tuia, D. 2020. RSVQA: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12): 8555–8566.
- Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lilliacrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittweis, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Wang, F.; Fu, X.; Huang, J. Y.; Li, Z.; Liu, Q.; Liu, X.; Ma, M. D.; Xu, N.; Zhou, W.; Zhang, K.; et al. 2024a. Muir-Bench: A Comprehensive Benchmark for Robust Multi-image Understanding. *arXiv preprint arXiv:2406.09411*.
- Wang, J.; Zheng, Z.; Chen, Z.; Ma, A.; and Zhong, Y. 2024b. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5481–5489.
- Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.
- Zhang, P.; Wang, X. D. B.; Cao, Y.; Xu, C.; Ouyang, L.; Zhao, Z.; Ding, S.; Zhang, S.; Duan, H.; Yan, H.; et al. 2023. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.
- Zhou, B.; Hu, Y.; Weng, X.; Jia, J.; Luo, J.; Liu, X.; Wu, J.; and Huang, L. 2024. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.