



# DynamicVL: Benchmarking Multimodal Large Language Models for Dynamic City Understanding

Weihao Xuan\*, Junjue Wang\*, Heli Qi, Zihang Chen, Zhuo Zheng,

Yanfei Zhong, Junshi Xia, Naoto Yokoya<sup>†</sup>

\* Equal Contribution † Corresponding Author

The University of Tokyo • RIKEN AIP • Waseda University • Wuhan University • Stanford University











### Motivation & Challenges

### **Research Gap**

Existing multi-temporal analysis is limited to bi-temporal or short sequences, lacking comprehensive long-term urban dynamics understanding.

### X Current Limitations

- Bi-temporal focus only
- Limited pixel-level analysis
- Insufficient training data

### **✓ Our Solution**

- Long-term sequences (5-10 frames)
- Pixel-precise detection
- Specialized instruction tuning



01: Densely describe all events in this scene from 2005 to 2023

Human: 2005-2019, there is no significant change. The farmland on the left is just normal crop rotation, 2019-2020, the top left farmland is being developed and the ground is being hardened. 2020-2023, the top left residential area was already built. The R bottom left farmland is being developed.

01: What has happened in the red box

Human: 2005-2014, the left parking lot was

occupied by homeless people with tents of

during 2005 to 2023.

### [DTC] Dense Temporal Caption

O1: Densely describe all events in this scene from 2005 to 2023.

RI: From 2005-2019, no significant changes were observed. The farmland on the left exhibited typical crop rotation patterns. Between 2019-2020, development began on the top left farmland. resulting in ground compaction. By 2020-2023, construction of the top left residential area was completed, featuring an artificial lake and primary roads, while development commenced on the bottom left farmland following a similar development pattern.

### [RCC] Regional Changes Caption

Q1: What has happened in the red box during 2005 to 2023.



RI: From 2005-2013, the left parking lot was occupied by individuals experiencing homelessness, with tents of various colors visible throughout the area. Between 2014-2018, this homeless encampment relocated from the parking lot on the left to the plaza on the right. By 2018-2023, due to the eviction measures, homeless settlements had disappeared from both the parking lot and plaza

various colors. 2014-2018, the homeless people move from the parking lot on the left to the plaza on the right, 2018-2023, Homeless settlements have disappeared in both parking lot and plaza.

### Six Core Urban Understanding Tasks



### [BCA] Basic Change Analysis

- O1: Which represented the most notable change from 2006 to 2009?
- R1: From vegetation to non-vegetation area.
- 02: What land-cover types have changed between 2011 and 2021?
- R2: Vegetation, non-vegetation, building and water areas have changed.
- 03: How has building area changed between 2016 and 2018? R3: Expanded.
- 04: How did land cover types change from 2006 to 2021? R4: Over the 15-year

period, 22-27% of vegetated areas transitioned to non-vegetated areas, while approximately 4% underwent the reverse transition. Notably, about 8% of vegetated areas and 4% of non-vegetated areas were converted to built-up land. Additionally, about 0.3% of vegetated areas were developed into water bodies.

### [CSE] Change Speed Estimation

- 01: Which time period exhibited the most significant changes? RI: 2009-2011.
- 02: During which time periods did building area expansion cease? R2: 2011-2016.
- 03: What is the magnitude of change for the entire region from 2006 to 2021? R3: 124%-151%.
- 04: What was the rate of building area expansion over time? R4: The expansion rates were as follows: 0.4% from 2006-2009, 0.05% from 2009-2011, 0.63% from 2016-2018, and 2.53% from 2018-2021.

### [EA] Environmental Assessment

- 01: Please calculate the estimated ozone density in this zone during 2006. R1: The 2006 ozone levels in this territory were estimated at 38.9 μg/m<sup>3</sup>
- 02: How strong was the nighttime illumination in this location in the year 2021? R2: The nighttime light in this district during 2021 was 18.74 nW/cm<sup>2</sup>/sr.
- 03: What was the mean temperature in this region during July 2021?
- 03: The mean temperature for this region in July 2021 was 36 77°C

### [RCD] Referring Change Detection

- 01: Identify and segment the areas that changed from vegetation to building during the 2016-2018 period.
- R1: Sure, the required changes are in
- 02: Identify the newly formed water bodies between 2018-2021.
- R2: The identified new pond is presented in
- 03: Delineate the regions that experienced conversion from nonvegetated to vegetated areas during 2011-2014.
- R3: The delineated regions are shown in
- 04: Show the new vegetation are in the period of 2006-2009.
- R4: Ok, the required area is shown in

### [RCC] Regional Changes Caption

- 01: What land cover changes occurred within the red-boxed region from 2006 to 2021?
- R1: From 2006 to 2016, the region exhibited cyclical land cover changes due to agricultural practices, with land surfaces alternating between vegetated and non-vegetated states during crop rotation and harvesting cycles. Beginning in 2016, these agricultural lands underwent significant development, resulting in surface hardening followed by progressive construction of built-up area. By 2021, the essential infrastructure, including buildings and primary roads, was almost completed.

### [DTC] Dense Temporal Captioning

- OI: Densely describe the dynamics in this scene from 2006 to 2021.
- R1: Over the 15-year period, natural and agricultural lands underwent progressive transformation into residential areas. Development initiated in the topright region and halted in 2011. The development process resumed in 2016, focusing on the left region of the study area. Most agricultural land was converted to non-vegetation surfaces, accompanied by extensive construction of buildings and roads. From 2018 to 2021, the bottom-center region also underwent development, including construction of an office building. By 2021, development of the area was nearly complete. Additionally, a new water body emerged in the bottom-right region of the study area.

### Semi-automatic annotation pipeline

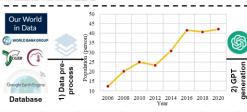


#### [BCA] Basic Change Analysis

- 01: What land-cover types have changed between 2013 and 2015? Options: A: Water, B: Playground, C: Building D: Nonvegetation, E: Vegetation.
- 02: What were the two predominant changes observed during the 2013-2015 period? Options: A: NV→BD, B: VG→NV, C: WT→NV, D: BD→NV, E: BD→VG.
- 03: Which land cover decreased the most between 2013 and 2015? Options: A: Water, B: Playground, C: Building, D: Vegetation, E. Non-vegetation

#### [CSE] Change Speed Estimation

- 01: What is the magnitude of change for the entire region from 2013 to 2015? Options: A: 3.08%, B: 3.46%, C: 3.85%, D: 4.23%, E: 4.62%,
- 02: Which period exhibited the minimal changes?
- Options: A: 2006-2009, B: 2009-2011, C: 2011-2013, D: 2013-2015, E: 2015-2017
- 03: During which periods did urban reconstruction take place?
- Options: A: 2006-2009, B: 2009-2011, C: 2011-2013, D: 2013-



### [EA] Environmental Assessment | 2022/12/3

OI: What was the population in this region during the year 2012? Op A. 23.3 people, B. 35.0 people, C. 11.6 people, D. 30.3 people, E. 16.3 people. 02: How many residents were in this location back in 2020? Options: A. 63.1 people, B. 42.1 people, C. 29.4 people, D. 54.7 people, E. 21.0 people.









#### [RCD] Referring Change Detection

- 01: Segment the totally destroyed and unaffected buildings after the earthquake? R1: Sure. They are in red mask and evan mask.
- Q1: Where is the best place to take refuge? RI: There are two playgrounds with vast open areas near the clear primary roads for rescue.



01: Densely describe all events in this scene from 2005 to 2023

Human: 2005-2019, there is no significant change. The farmland on the left is just normal crop rotation, 2019-2020, the top left farmland is being developed and the ground is being hardened. 2020-2023, the top left residential area was already built. The bottom left farmland is being developed.



### 01: Densely describe all events in this scene from 2005 to 2023.

RI: From 2005-2019, no significant changes were observed. The farmland on the left exhibited typical crop rotation patterns. Between 2019-2020, development began on the top left farmland, resulting in ground compaction. By 2020-2023, construction of the top left residential area was completed, featuring an artificial lake and primary roads, while development commenced on the bottom left farmland following a similar development pattern.

[DTC] Dense Temporal Caption





during 2005 to 2023.

Human: 2005-2014, the left parking lot was occupied by homeless people with tents of various colors. 2014-2018, the homeless people move from the parking lot on the left to the plaza on the right, 2018-2023, Homeless settlements have disappeared in 8 both parking lot and plaza.

01: What has happened in the red box

### [RCC] Regional Changes Caption

OI: What has happened in the red box during 2005 to 2023.

R1: From 2005-2013, the left parking lot was occupied by individuals experiencing homelessness, with tents of various colors → visible throughout the area. Between 2014-2018, this homeless encampment relocated from the parking lot on the left to the plaza on the right. By 2018-2023, due to the eviction measures, homeless settlements had disappeared from both the parking lot and plaza b locations.

## Key Experimental Findings

Method	AVG	BCA-QA		CSE-QA		EA	BCA-Report				CSE-Report			
		Single	Multi	Single	Multi		AVG	LCT	TPA	CQA	AVG	CRP	TPA	CPA
Commercial models														
o4-mini [34]	34.1	62.8	36.1	33.8	12.4	25.3	3.16	2.85	4.70	1.93	2.34	0.97	3.71	2.33
GPT4.1 [33]	32.5	66.1	39.7	31.3	5.4	20.2	3.02	2.69	4.67	1.72	2.23	0.78	3.84	2.05
GPT4o [31]	29.7	63.3	19.3	32.3	7.3	26.2	2.96	2.55	4.66	1.66	2.21	0.73	3.46	2.43
Gemini 2.5 Flash [9]	24.4	46.3	15.8	21.0	12.1	26.8	2.90	2.40	4.69	1.62	2.19	0.70	3.78	2.09
Open-source models														
TEOChat [13]	17.2	35.1	8.7	17.0	10.8	14.6	0.64	1.61	0.22	0.09	1.22	0.85	1.46	1.33
EarthDial [37]	30.3	62.2	20.3	30.9	12.2	25.9	1.10	2.57	0.01	0.72	1.03	0.85	0.74	1.50
Video-LLaVA [24]	17.7	34.8	10.4	17.7	5.4	20.2	2.01	1.58	3.14	1.33	1.63	0.86	2.48	1.54
LLaVA-OneVision 7B [20]	19.3	41.7	2.8	21.5	4.8	25.9	2.30	2.29	3.20	1.42	1.72	0.95	2.44	1.78
LLaVA-OneVision 72B [20]	25.0	59.9	6.5	25.9	6.2	26.5	3.01	2.70	4.52	1.83	2.05	0.93	3.39	1.83
InternVL3 8B [55]	23.9	55.2	11.5	22.0	7.6	23.1	2.99	2.49	4.68	1.78	2.15	0.95	3.31	2.20
InternVL3 14B [55]	27.2	63.2	15.3	28.8	4.0	24.9	3.02	2.61	4.72	1.74	2.36	0.97	3.65	2.48
InternVL3 78B [55]	27.1	60.5	14.5	28.3	8.6	23.6	3.04	2.74	4.59	1.80	2.25	0.82	3.87	2.06
Qwen2.5-VL 3B [2]	24.7	56.9	6.0	26.1	9.2	25.1	2.99	2.72	4.58	1.65	1.72	0.57	3.42	1.18
Qwen2.5-VL 7B [2]	23.3	54.6	4.8	28.5	13.6	15.0	2.94	2.49	4.70	1.62	1.73	0.25	3.90	1.05
Qwen2.5-VL 32B [2]	31.4	62.0	33.3	36.9	3.2	21.6	3.04	2.65	4.65	1.81	2.60	1.21	3.89	2.71
Qwen2.5-VL 72B [2]	29.7	65.4	24.3	34.6	4.0	20.2	2.99	2.61	4.64	1.71	2.27	0.72	3.76	2.33
Ours														
DVLChat 7B	33.3	64.9	21.3	31.3	18.6	30.6	3.47	3.41	4.72	2.28	2.51	1.48	3.41	2.65

### Challenge

Best model achieves only 34.1% accuracy

### **DVLChat**

Outperforms base model with significant gains

### Limitation

Scaling insufficient - domain data crucial

### Gap

18 models - all show deficiencies

### Contributions & Impact

1 Dataset

First large-scale long-term multi-temporal benchmark with 14,871 images spanning 18 years

3 Model

DVLChat demonstrating **unified capabilities** in QA and pixel-level segmentation

**2** Evaluation

Assessment of **18 MLLMs** revealing critical limitations in temporal understanding

4 Impact

Advancing sustainable city development through **Al-powered**monitoring

Resources

github.com/weihao1115/dynamicvl

huggingface.co/datasets/weihao1115/dvl\_suite