



DynamicVL: Benchmarking Multimodal Large Language Models for Dynamic City Understanding

Weihaio Xuan*, Junjue Wang*, Heli Qi, Zihang Chen, Zhuo Zheng,

Yanfei Zhong, Junshi Xia, Naoto Yokoya†

* Equal Contribution † Corresponding Author

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



東京大学
THE UNIVERSITY OF TOKYO



Motivation & Challenges

Research Gap

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

✗ 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



1) Human
annotate

1) Human
annotate

Q1: 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.

Q1: What has happened in the red box 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 had disappeared in both parking lot and plaza.

2) GPT
generation

2) GPT
generation

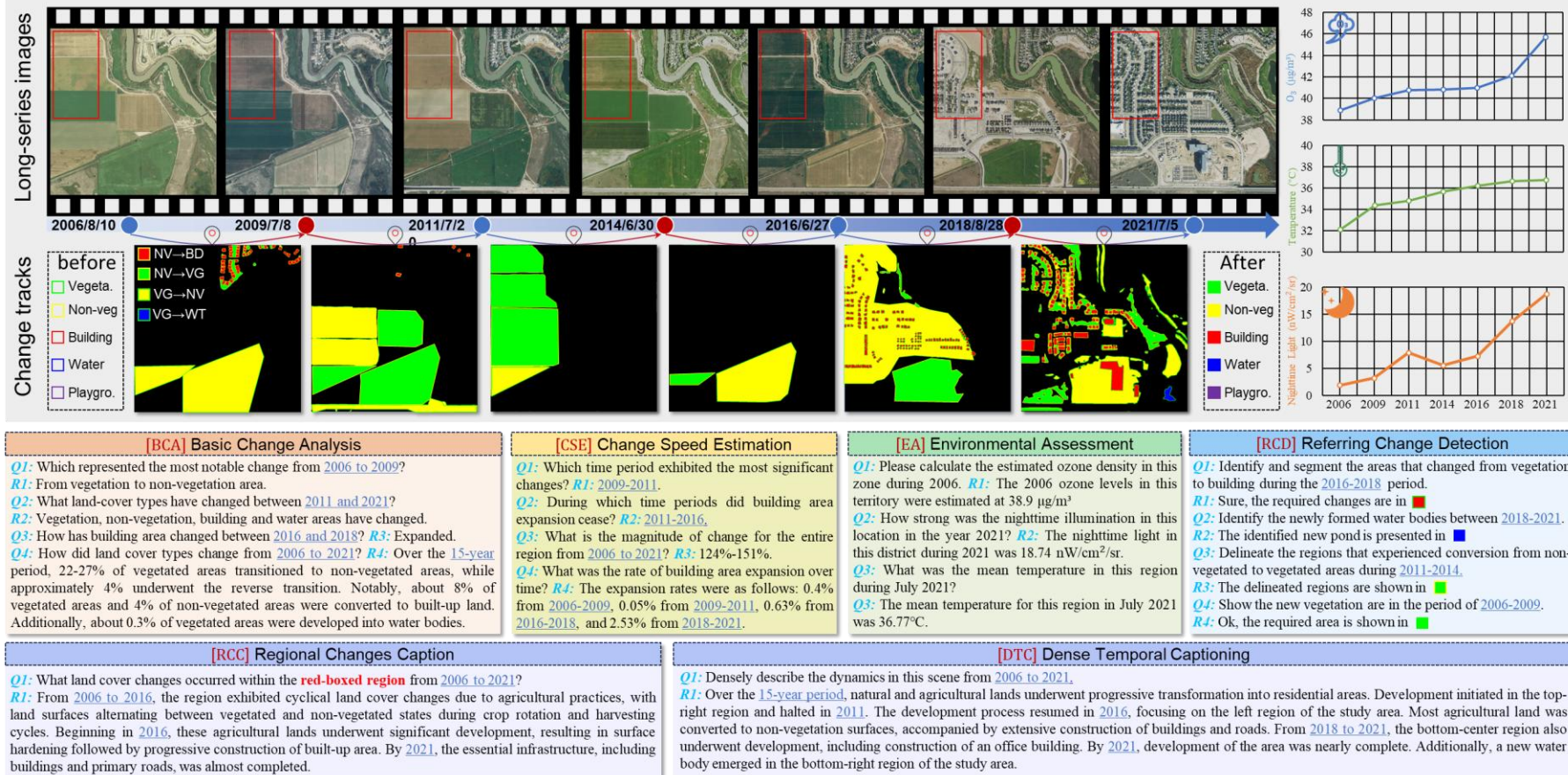
[DTC] Dense Temporal Caption

Q1: 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.

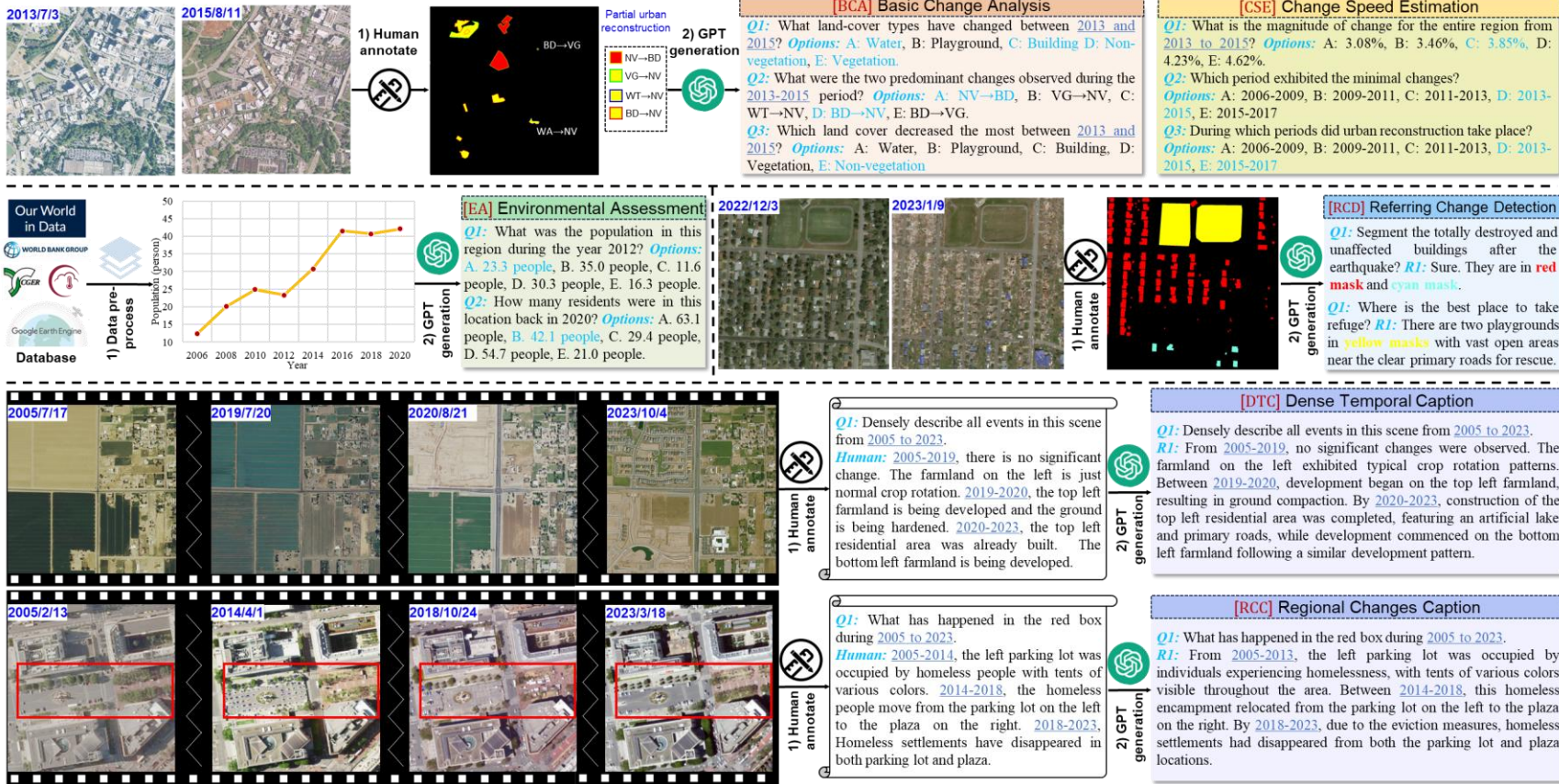
[RCC] Regional Changes Caption

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 locations.

Six Core Urban Understanding Tasks



Semi-automatic annotation pipeline



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**

2 Evaluation

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

3 Model

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

4 Impact

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



Resources

github.com/weihao1115/dynamiccvl

huggingface.co/datasets/weihao1115/dvl_suite