DynamicVL: Benchmarking Multimodal Large Language Models for Dynamic City Understanding

Weihao Xuan[1,2*] **Junjue Wang**[1*] **Heli Qi**[2,3] **Zihang Chen**[4]
Zhuo Zheng[5] **Yanfei Zhong**[4] **Junshi Xia**[2] **Naoto Yokoya**[1,2†]
[1] The University of Tokyo [2] RIKEN AIP [3] Waseda University
[4] Wuhan University [5] Stanford University

Abstract

Multimodal large language models (MLLMs) have demonstrated remarkable capabilities in visual understanding, but their application to long-term Earth observation analysis remains limited, primarily focusing on single-temporal or bi-temporal imagery. To address this gap, we introduce **DVL-Suite**, a comprehensive framework for analyzing long-term urban dynamics through remote sensing imagery. Our suite comprises 14,871 high-resolution (1.0m) multi-temporal images spanning 42 major cities in the U.S. from 2005 to 2023, organized into two components: **DVL-Bench** and **DVL-Instruct**. The *DVL-Bench* includes six urban understanding tasks, from fundamental change detection (*pixel-level*) to quantitative analyses (*regional-level*) and comprehensive urban narratives (*scene-level*), capturing diverse urban dynamics including expansion/transformation patterns, disaster assessment, and environmental challenges. We evaluate 18 state-of-the-art MLLMs and reveal their limitations in long-term temporal understanding and quantitative analysis. These challenges motivate the creation of *DVL-Instruct*, a specialized instruction-tuning dataset designed to enhance models' capabilities in multi-temporal Earth observation. Building upon this dataset, we develop **DVLChat**, a baseline model capable of both image-level question-answering and pixel-level segmentation, facilitating a comprehensive understanding of city dynamics through language interactions. Project: https://github.com/weihao1115/dynamicvl.

1 Introduction

Sustainable city, as a key goal in "The 2030 Agenda for Sustainable Development"[3], has proposed new requirements for urban resilience, convenience, and comfort. Remote sensing technology enables us to monitor urban development over time by analyzing satellite imagery, allowing us to track large-scale changes in urban landscapes [8, 45]. However, research in this field has been largely limited to comparing images from only two time points [54, 42], primarily due to the scarcity of well-aligned vision-language datasets spanning longer time series. This limitation has constrained our ability to conduct a comprehensive, large-scale understanding of urban dynamics.

The recent emergence of MLLMs [20, 26, 40] represents a significant advancement in visual-language understanding. These models mark a shift from specialized, single-purpose systems to versatile frameworks capable of handling multiple tasks, including but not limited to visual grounding [27], image captioning [5], and visual question answering [51, 52]. While recent MLLM research has demonstrated promising results in multi-image [29, 15] and video understanding [43, 12], these

[*] Equal contribution.
[†] Corresponding author.
[3] https://sdgs.un.org/goals/goal11

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

https://doi.org/10.52202/085713-1114

Figure 1: Diverse tasks in the **DVL-Bench**. Our framework encompasses multiple levels of temporal understanding: from pixel-precise change detection and quantification to regional evolution analysis and dense temporal captioning. This hierarchical task design enables systematic evaluation of MLLMs' capabilities in multi-temporal Earth observation understanding.

efforts primarily focus on in-the-wild daily imagery. In the context of remote sensing, existing research [13, 37] on multi-temporal analysis is typically limited to bi-temporal or short-sequence comparisons. Moreover, current multi-temporal MLLMs in remote sensing are primarily tested on high-level semantic understanding, lacking pixel-precise analysis capabilities crucial for quantitative change assessment. These limitations pose significant challenges for dynamic city understanding applications, which require both long-term understanding beyond bi-temporal inputs and precise quantitative analysis of environmental changes.

To address these challenges, we present **DVL-Suite**, a comprehensive framework for analyzing urban dynamics over time using remote sensing imagery. Our framework introduces **DVL-Bench**, a large-scale benchmark designed for rigorous evaluation of vision-language models in urban contexts. Building on recent advances in video understanding benchmarks [36, 27, 49], we develop a structured taxonomy that identifies six core capabilities essential for sustainable urban understanding: 1) `[BCA]` Basic Change Analysis: Focuses on identifying and comparing multi-temporal changes in land-use patterns. 2) `[CSE]` Change Speed Estimation: Tracks and quantifies temporal trends of key urban elements, such as building expansion rates and vegetation loss. 3) `[EA]` Environmental Assessment: Evaluates urban livability and economic indicators through visual analysis. 4) `[RCD]` Referring Change Detection: Tests models' capabilities in dense reasoning and precise spatial localization of changes. 5) `[RCC]` Regional Change Captioning: Generates detailed change descriptions for user-specified geographical areas. 6) `[DTC]` Dense Temporal Captioning: Generates comprehensive reports documenting long-term temporal changes, highlighting critical events across long time series.

To ensure comprehensive coverage, DVL-Bench encompasses diverse urban scenarios, such as urban expansion, housing crises, natural disasters, urban heat island effects, and green space development. Unlike traditional bi-temporal understanding, it enables systematic evaluation of long-term city dynamics, facilitating deeper insights into sustainable city development through multi-temporal Earth observation.

Through extensive experimentation on DVL-Bench, we discovered that state-of-the-art MLLMs, both commercial and open-source models, face significant challenges in long-term temporal visual understanding, primarily due to insufficient training data spanning extended time periods. To address these limitations, we introduce **DVL-Instruct**, a specialized instruction-tuning dataset designed for dynamic city understanding in remote sensing. Using this dataset, we develop **DVLChat**, a baseline model that enhances multi-temporal urban understanding and caption generation, while introducing referring change detection capabilities.

The key contributions of this paper are as follows:

1. We introduce DVL-Suite, comprising DVL-Bench and DVL-Instruct, with 14,871 high-resolution (1.0m) images spanning 42 U.S. cities, featuring an average of 6.73-6.94 temporal frames per scene from 2005 to 2023, enabling long-term urban dynamics analysis with

https://doi.org/10.52202/085713-1114

a coherent task taxonomy, multi-level analysis capabilities, and thematic focus on urban development patterns.

2. We evaluate 18 vision-language models, revealing critical limitations: the best-performing model, o4-mini, achieves only 34.1% accuracy on DVL-Bench's overall QA average, demonstrating significant deficiencies in complex temporal tasks and quantitative analysis.

3. Based on DVL-Instruct, we develop DVLChat, a baseline that surpasses its base Qwen2.5-VL 7B by significant improvements, enabling multi-temporal urban analysis and referring change detection from a single model.

2 Related Work

2.1 Large Multimodal Language Models

The rapid advancement of MLLMs has sparked significant interest in their applications to complex visual understanding tasks, including understanding temporal dynamics and multiple image inputs, which are central to multi-temporal remote sensing analysis. In generic vision-language models, early pioneering works such as Flamingo [1] demonstrated the capability to process interleaved sequences of visual and textual data, including video frames, through mechanisms like Perceiver Resampler. Subsequent developments have enhanced video understanding, with models like Video-LLaVA [24] unifying image and video representations before LLM projection, and LLaVA-OneVision [20] offering a unified framework for single-image, multi-image, and video tasks via the "Higher AnyRes" strategy. Qwen2-VL series [40, 2] introduced Multimodal Rotary Position Embedding (M-ROPE) aligned with absolute time for long video comprehension, and InternVL3 [55] employed Variable Visual Position Encoding (V2PE) for extended multimodal contexts including lengthy video sequences. Concurrently, the challenge of multi-image understanding has been addressed by models such as LLaVA-NeXT-Interleave [21], which utilizes a data-centric approach with the M4-Instruct dataset to handle diverse multi-image scenarios.

Despite these advances in temporal and multi-image processing, existing models still fall short in dynamic, long-term remote sensing analysis, particularly for precise quantitative assessment of urban changes. To address this gap, we introduce DVL-Suite and DVLChat to advance multi-temporal urban understanding.

Table 1: Comparison with existing multi-temporal remote sensing vision-language datasets.

Dataset	Self-contained	Average Temp.	Text Pairs	MT Images	Image Size	[BCA]	[CSE]	[EA]	[RCD]	[RCC]	[DTC]
RSICap [28]	✓	1	2585	0	512	×	×	×	×	×	×
LHRS-Bot [30]	✓	1	1.2M	0	768	×	×	×	×	×	×
VRSBench [23]	✓	1	205k	0	512	×	×	×	×	×	×
GeoChatSet [17]	×	1	318k	0	504	×	×	×	×	×	×
CDVQA [50]	✓	2	122k	122k	512	✓	×	×	×	×	×
LEVIR-MCI [25]	✓	2	50.3k	50.3k	256	×	×	×	×	×	×
OVG-360k [22]	×	2	360k	360k	512	✓	×	×	✓	×	×
ChangeChat [6]	×	2	87k	87k	256	✓	×	×	×	×	×
GeoLLaVA [7]	×	2	100k	100k	336	×	×	×	×	×	×
CC-Expert [41]	×	2	135k	135k	384	×	×	×	×	×	×
TEOChatlas [13]	×	2.07	554k	245k	224	✓	×	×	×	✓	×
EarthDial [37]	×	1.01	11M	64.6k	448∼1024	✓	×	×	×	✓	×
DVL-Bench	✓	**6.94**	8,682	3,469	**1024**	✓	✓	✓	✓	✓	✓
DVL-Instruct	✓	**6.73**	63,771	11,402	**1024**	✓	✓	✓	✓	✓	✓

2.2 Multimodal Benchmarks in Remote Sensing

Remote Sensing (RS) domain has witnessed the emergence of numerous specialized multimodal datasets [44]. LHRS-Bot [30], VRSBench [23], and GeoChatSet [17] pioneered single-temporal instruction datasets for classification, detection, and visual question answering (VQA). Subsequently, CDVQA [50] introduced change-aware VQA, while LEVIR-MCI [25] integrated pixel-level masks. GeoLLaVA [7] and CC-Expert [41] enhanced interactive bi-temporal captioning, and OVG-360k [22] provided fine-grained spatial semantic supervision. Although surpassing single-temporal analyses, these efforts remain limited to bi-temporal image pairs. Recently, TEOChatlas [13] curates temporal instruction-following tasks such as those derived from xBD [10] and fMoW [4]. DisasterM3 [39] provides a multi-hazard, multi-sensor, and multi-task remote sensing vision-language benchmark

https://doi.org/10.52202/085713-1114

with 26,988 bi-temporal satellite images and diverse disaster assessment tasks. However, existing datasets predominantly focus on bi-temporal understanding and lack comprehensive evaluations of models' capabilities in processing extended temporal sequences and performing long-term spatiotemporal reasoning. To enable MLLM to excel in understanding long-term remote sensing images, we introduce DVL-Bench, a large-scale vision-language benchmark for remote sensing analysis within long time series that offers three key advantages: **1) Coherent task taxonomy.** Unlike composite datasets assembled from heterogeneous sources, DVL-Bench introduces a systematically designed task taxonomy built upon newly collected data with consistent annotation standards. **2) Diverse temporal tasks.** DVL-Bench includes multiple analysis granularities, progressing from fine-grained pixel-level change detection and region-based dynamic captioning to holistic temporal reasoning and comprehensive environmental assessment, thereby facilitating systematic city dynamics understanding. **3) Practical and thematically focused.** In contrast to existing datasets addressing wide-ranging geospatial tasks, DVL-Bench specifically targets the analysis and representation of long-term urban development dynamics. In addition, the developed DVLChat can be directly integrated with the NAIP platform, serving as an AI assistant for urban understanding applications.

3 DVL-Suite Curation Pipeline

To ensure data diversity and quality, we built DVL-Suite using high-resolution (1.0m GSD) remote sensing imagery from the National Agriculture Imagery Program (NAIP), covering 42 major U.S. cities. The imagery was first geo-referenced and processed into 14,871 patches of 1024×1024 pixels, comprising 2,193 multi-temporal scenes. For compatibility with generic MLLMs, we utilized three optical bands. By collecting environmental datasets from diverse Earth observation platforms (sources are detailed in Appendix § D), all data were spatially resampled to match the resolution of collected remote sensing imagery, ensuring one-to-one correspondence between images and environmental indicators. Based on the multi-source Earth observation data, we hired several experts and a well-trained annotation team to label and examine the DVL-Suite.

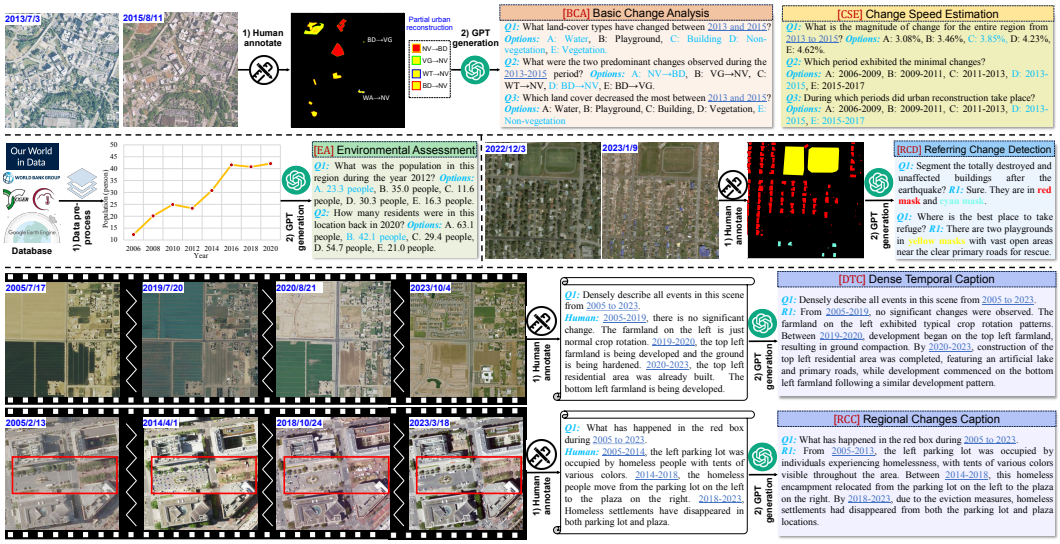

Figure 2: The annotation pipeline of the proposed DVL-Suite. Four common urban dynamics are depicted from top to bottom: partial urban reconstruction, natural disasters, farmland conversion, and homeless encampments. In our semi-auto pipeline, urban experts perform the basic annotations, while GPT4.1 integrates this information to generate the paired instructions.

Figure 2 illustrates our multi-stage annotation pipeline for the DVL-Bench dataset. The annotation process includes several specialized tasks: For bi-temporal analysis tasks (`[BCA]` and `[CSE]`), annotators first segmented semantic change areas between adjacent temporal images. These changes were categorized across five primary land-cover types: vegetation, non-vegetated surfaces, water, buildings, and playgrounds. This categorization, adapted from SECOND [46], yielded 20 distinct change event categories. GPT4.1 then generated diverse task-specific instructions using these segmentation masks and categories. For `[BCA]` questions (e.g., "What land-cover type changed most between 2015 and 2017?"), the system calculated correct answers from the masks and generated

https://doi.org/10.52202/085713-1114

four incorrect options from other land-cover types. For [CSE] tasks, change speeds were computed from the masks, with alternative options varying by ±20% and ±40%. The [EA] task followed a similar pipeline, but utilized the multi-source environmental indicators as reference data. For [RCD] tasks, domain experts designed event-specific prompts, followed by manual mask annotation and GPT4.1-based language enhancement of prompts and answers. For temporal narrative tasks ([DTC] and [RCC]), annotators first identified keyframes containing significant changes to segment the temporal sequence, then crafted period-specific captions. GPT4.1 refined these draft captions to enhance detail and coherence. While both tasks follow identical procedures, [RCC] focuses on user-specified local areas rather than global changes.

After the first round of labeling, we conducted a rigorous quality control process: self-examination, cross-examination to correct errors (including false labels, missing details, and inaccurate captions), and supervisor review of 1,000 randomly sampled annotations. Any unqualified annotations were returned for refinement. DVL-Instruct follows the same data curation pipeline, but differs in that it exclusively pairs instructions with ground truths rather than providing multiple-choice options.

4 DVL-Bench Designs

DVL-Bench consists of 3,469 multi-temporal Earth observation images, each accompanied by human-verified annotations. The annotation suite includes 1,391 referring segmentation instructions for precise change localization, 5,854 question-answer pairs for detailed temporal analysis, and 1,437 comprehensive captions documenting urban dynamics. This section outlines the task taxonomy and examines the fundamental challenges in interpreting long-term remote sensing sequences.

Instruction distributions with different tasks. Figure 3 presents the distribution of instructions across task categories. The [BCA], [CSE], [EA], and [RCD] tasks exhibit relatively balanced sample distributions, collectively accounting for 89.9% of the benchmark. In contrast to these tasks, which primarily focus on part of frames within each scene, the [DTC] and [RCC] tasks require comprehensive analysis of complete temporal sequences, thus representing smaller proportions of the overall distribution. With substantial sample sizes across all categories, DVL-Bench enables systematic evaluation of MLLMs in both dynamic understanding and open-ended generation performances.

Geospatial and temporal distributions. Figure 4 shows the spatial distribution across 42 U.S. cities and the temporal length per scene. The samples are evenly distributed in 42 rapidly growing cities, ensuring comprehensive geographical coverage and minimizing regional bias. Unlike the existing multi-temporal dataset focusing on understanding bi-temporal changes, the DVL-Bench features long-term analysis with sequences ranging from five to ten frames per urban scene. This extended temporal scope poses new challenges for modeling discrete temporal transitions in urban environments.

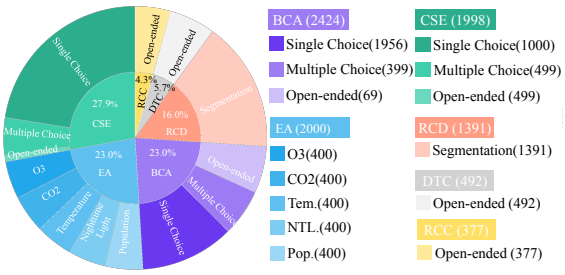

Figure 3: Task taxonomy and sample distribution in DVL-Bench. The multi-level task evaluates MLLM comprehensively.

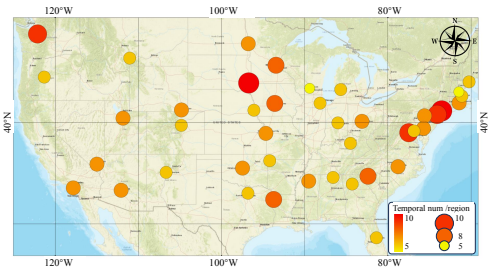

Figure 4: Data distributions across the 42 rapidly growing cities and the temporal number floats from five to ten.

Basic change analysis (BCA). The Sankey diagram in Figure 5 quantifies land cover transition patterns between initial and final states. The visualization reveals a dominant trend of vegetation conversion to developed areas, reflecting rapid urbanization processes. In contrast, only a modest area of building, approximately $2.34km^2$ underwent demolition, primarily transitioning to vegetation and non-vegetation surfaces. These complex dynamics, involving multi-directional transitions among five distinct land cover types and requiring precise quantification across various spatial scales, present significant challenges for MLLMs in terms of both semantic understanding and numerical reasoning.

https://doi.org/10.52202/085713-1114

Change speed estimation (CSE). The temporal analysis in Figure 6 tracks building expansion rates across successive periods, providing insights into U.S. urban development trajectories over the past two decades. Development velocity exhibits a distinct non-linear pattern: accelerating from 2010, reaching peak urbanization rates around 2017, and showing significant deceleration after 2018. This characteristic growth curve, with its pronounced acceleration and subsequent slowdown, represents a typical urban development cycle. Such complex temporal dynamics require MLLMs to maintain precise numerical sensitivity while modeling long-term spatiotemporal variations.

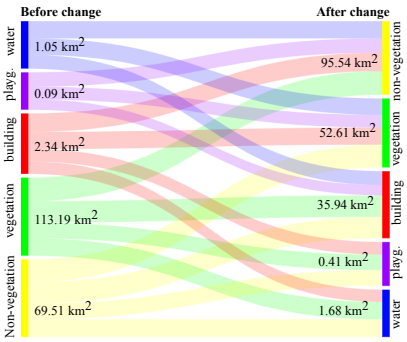

Figure 5: The basic change flow in DVL-Bench.

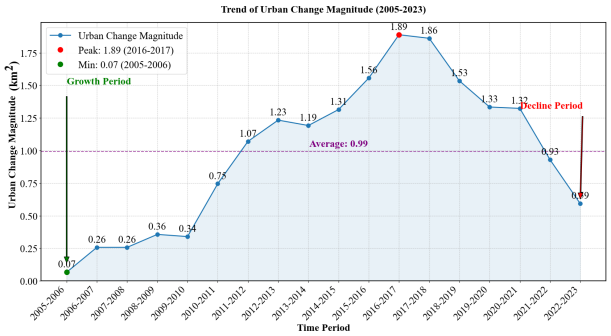

Figure 6: Trend in change magnitude per period, showing non-linear development speed across the U.S.

Referring change detection (RCD). Analysis of change scales in Figure 7 reveals distinct patterns across the five primary land-cover types. Changes in vegetation and non-vegetation areas show high variability, with a clear asymmetric distribution favoring smaller spatial extents. Buildings, water bodies, and recreational areas predominantly undergo small-scale changes, consistent with their limited spatial footprint in urban landscapes. These diverse change scales across land-cover types test MLLMs' ability to detect and segment temporal changes at varying spatial scales.

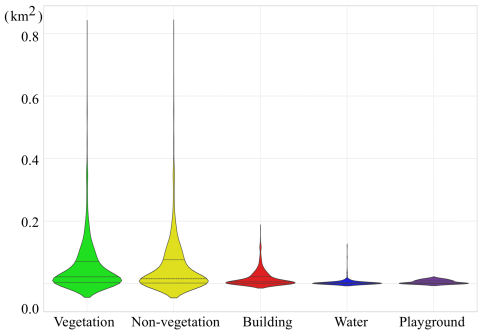

Figure 7: The change scale distributions in referring change detection.

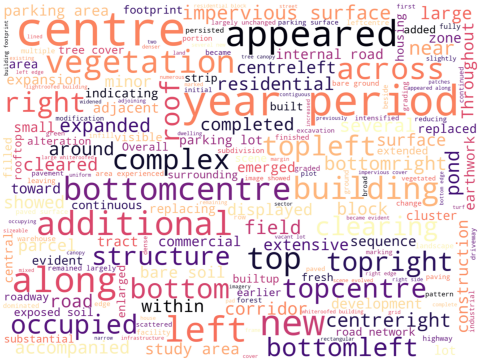

Figure 8: Word cloud in the dense temporal and regional change captioning tasks.

Dense temporal and region-level change caption (DTC and RCC). The word cloud visualization in Figure 8 illustrates the distribution of the 300 most frequent terms in designed long-form captions. While existing geospatial vision-language datasets primarily focus on spatial descriptors (topright, left, etc.), DVL-Bench distinctively features rich temporal references (year, period, etc.) and transition vocabulary (replaced, added, completed, etc.), enabling comprehensive documentation of urban dynamics.

5 Benchmark Experiments

In this section, we present the experimental setup, introduce DVLChat as our baseline model for dynamic urban understanding, and comprehensively evaluate 18 state-of-the-art MLLMs on DVL-Bench.

https://doi.org/10.52202/085713-1114

5.1 Implementation Details

Benchmark methods. We evaluate 18 widely-adopted MLLMs across different capabilities: (1) open-source MLLMs, including domain-specific models like TEOChat [13] and EarthDial [37], state-of-the-art MLLMs with multi-image and video perception abilities (Video-LLaVA [24], LLaVA-OneVision [20], InternVL3 [55], and Qwen2.5-VL [2]), and referring segmentation models (LISA [19], PSALM [53]); (2) commercial MLLMs, including o4-mini [34], GPT4.1 [33], GPT4o [31], and Gemini 2.5 Flash [9]. Unless otherwise specified, all experiments using open-source models were conducted on 8 H100 GPUs. For question-answering tasks, we utilized each model's native multi-image inference capability. For referring change detection tasks, LISA and PSALM were evaluated using a different approach, concatenating two temporal images into a single composite image as input. The detailed training and testing methods can be found in the Appendix § A.

Evaluation metrics. Our evaluation framework employs multiple metrics to assess model performance across different tasks comprehensively. For the evaluations presented in Table 2, we measure both basic change analysis and change speed estimation using two approaches. First, we calculate accuracy percentages for single and multiple-choice questions. Second, for open-ended generation tasks, we evaluate Basic Change Reports using three metrics: Land Cover Type Identification (LCT), Time Period Accuracy (TPA), and Change Quantification Accuracy (CQA). Similarly, Change Speed Reports are assessed using Change Rate Precision (CRP), Time Period Accuracy (TPA), and Change Pattern Accuracy (CPA). For the long-form captioning tasks shown in Table 3, Regional Change Captioning is evaluated using Temporal Coverage (TC), Spatial Accuracy (SA), Process Fidelity (PF), and Region Containment (RC), while Dense Temporal Captioning uses TC, SA, and PF. All captioning metrics are scored on a 0-5 scale, with higher scores indicating better performance. These scores are determined by GPT4.1-mini [32] through comparison with reference captions. The detailed evaluation prompts can be found in the Appendix § B.2.

DVLChat design. As dynamic urban understanding necessitates both semantic comprehension and fine-grained pixel-level understanding, we followed the main architecture of LISA [19] to develop DVLChat. However, the original LISA model was unable to perform pixel-level segmentation while maintaining high open-ended capabilities due to optimization conflicts. Furthermore, LISA, designed for single-image analysis, lacked the ability to analyze changes across multiple images, whereas the DVL-Instruct data enables these capabilities. Therefore, leveraging DVL-Instruct, we develop DVLChat as a baseline model for multi-temporal urban understanding tasks. As shown in Figure 9, DVLChat employs a task-specific routing mechanism through dedicated prefix tokens from users. The system routes user queries to specialized modules based on their prefixes: inputs with [QA] activate the VQA LoRA module for generating textual responses, while those with [SE] engage the change detection LoRA module. DVLChat addresses multi-temporal analysis by interleaving image features from multiple temporal images before decoding. For referring change detection tasks, the system processes this interleaved representation and decodes the <SEG> token embedding using SAM's [16] frozen vision backbone and unfrozen decoder to generate precise segmentation masks. DVLChat effectively isolates question-answering and change detection functionalities, preventing task interference in the original LISA algorithm while maintaining model efficiency. While our implementation uses Qwen2.5-VL as the MLLM, the architecture is MLLM-agnostic and can accommodate other multimodal language models.

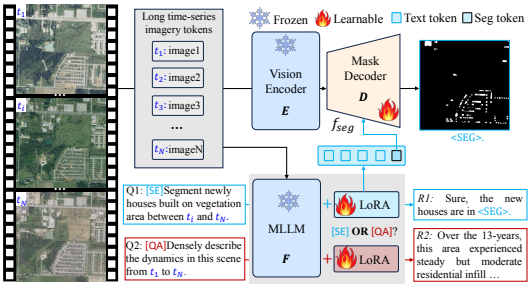

Figure 9: Detailed illustration of DVLChat.We separate question-answering and referring change detection by implementing two distinct LoRA [11] modules, enabling the model to possess independent VQA and segmentation capabilities while preventing interference between their respective data streams in the training.

5.2 Benchmark Results

Overall benchmark performance. As shown in Table 2, quantitative results reveal significant challenges in understanding urban dynamics through remote sensing imagery, with all current models

https://doi.org/10.52202/085713-1114

Table 2: Quantitative evaluation results of various vision-language models on basic change analysis (BCA), change speed estimation (CSE), and environmental assessment (EA) tasks. Performance is measured by accuracy percentages for single/multiple-choice questions and a scale of 0-5 for captioning tasks.

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

demonstrating limited capabilities. The highest averaged accuracy of multiple-choice questions reaches merely 34.1% with o4-mini, while Qwen2.5-VL 32B and GPT4.1 achieve 31.4% and 32.5% respectively. Notably, TEOChat, despite being specifically designed for multi-temporal remote sensing vision-language tasks, achieves only 17.2% overall accuracy, struggling significantly with the benchmark's larger and city-level understanding compared to its native 256×256 input size. In contrast, our DVLChat 7B, leveraging the proposed DVL-Instruct dataset, demonstrates competitive performance at 33.3% while maintaining referring change detection capabilities.

Task-specific challenges. Diverse tasks evaluate MLLMs' performances from different aspects. While `[BCA]` with single-choice questions shows promising results, where Qwen2.5-VL 72B achieves 65.4% accuracy, performance degrades substantially in multi-choice settings, where even the leading model GPT4.1 only reaches 39.7%. The challenges become more pronounced in detailed analytical tasks. `[BCA]` report metrics expose fundamental limitations in both land cover type identification (LCT) and change quantification (CQA), with LCT scores maxing at 2.85 and CQA not exceeding 1.93 across existing models. Notably, by leveraging DVL-Instruct's comprehensive training data, DVLChat achieves a significant breakthrough in LCT with a score of 3.41, enhancing the recognition of changed land-cover types. `[CSE]` results reveal a critical limitation of current MLLMs in pixel-level change perception, with multi-choice accuracy peaking at merely 13.6% and Change Rate Precision (CRP) consistently below 1.21, indicating models' inability to capture and quantify fine-grained temporal variations. `[EA]` results are similarly concerning: except for our DVLChat 7B, other models achieve accuracies ranging from 14.6% to 26.8%, with many performing at or even below random chance (20% for 5-option questions).

Captioning capabilities. Table 3 reveals a substantial gap between commercial and open-source models in detailed captioning tasks. For the regional change captioning task, commercial models demonstrate superior performance with o4-mini achieving an average score of 4.58, while the best open-source model, InternVL3 14B, reaches only 3.96. Our DVLChat, incorporating DVL-Instruct, demonstrates strong performance with an average score of 3.98, approaching the performance of commercial models with 7B parameters. The disparity becomes even more pronounced in dense temporal captioning, where commercial models maintain strong performance with o4-mini reaching an average score of 4.14, while open-source alternatives struggle considerably with scores below 3.40. Notably, TEOChat achieves only 1.45, revealing severe limitations in handling complex temporal dynamics beyond bi-temporal comparisons.

Scaling parameters. Analysis of model size scaling reveals inconsistent improvements within different model families, which is distinguished from most generic computer vision tasks [52, 12]. Particularly in basic change analysis and change speed estimation tasks, the Qwen2.5-VL series shows notable improvements as model size increases to 32B, reaching 31.4% average accuracy, but performance declines to 29.7% with the 72B counterpart. Similarly, while LLaVA-OneVision

Table 3: Performance comparison of different models on regional change captioning (RCC) and dense temporal captioning (DTC) tasks, evaluated using Temporal Coverage (TC), Spatial Accuracy (SA), Process Fidelity (PF), and Region Containment (RC) metrics on a 0-5 scale.

Method	RCC					DTC			
	AVG	TC	SA	PF	RC	AVG	TC	SA	PF
Commercial models									
o4-mini [34]	4.58	4.79	4.21	4.35	4.97	4.14	4.64	4.04	3.73
GPT4.1 [33]	4.46	4.74	3.99	4.16	4.97	3.98	4.53	3.75	3.65
GPT4o [31]	4.32	4.66	3.78	3.89	4.96	3.87	4.45	3.65	3.49
Gemini 2.5 Flash [9]	4.34	4.66	3.84	3.87	4.99	3.61	4.15	3.41	3.28
Open-source models									
TEOChat [13]	1.66	1.02	0.45	0.29	4.87	1.45	1.65	1.14	1.57
EarthDial [37]	1.53	0.68	0.39	0.17	4.86	0.90	0.80	1.16	0.75
Video-LLaVA [24]	2.49	1.21	1.93	2.04	4.81	1.76	2.38	1.57	1.34
LLaVA-OneVision 7B [20]	3.07	3.12	2.37	2.17	4.63	2.17	2.36	2.09	2.08
LLaVA-OneVision 72B [20]	3.60	3.91	2.77	2.80	4.93	2.87	3.51	2.56	2.54
InternVL3 8B [55]	3.69	3.99	3.02	2.99	4.76	2.97	3.57	2.71	2.64
InternVL3 14B [55]	3.96	4.33	3.25	3.36	4.91	3.22	3.84	2.96	2.85
InternVL3 78B [55]	3.92	4.18	3.34	3.18	4.97	3.33	3.98	3.01	2.99
Qwen2.5-VL 3B [2]	2.76	2.77	1.82	1.52	4.92	2.38	2.38	2.66	2.11
Qwen2.5-VL 7B [2]	3.21	3.30	2.42	2.20	4.92	2.85	3.47	2.57	2.51
Qwen2.5-VL 32B [2]	3.90	4.28	3.18	3.23	4.92	2.91	3.39	2.77	2.57
Qwen2.5-VL 72B [2]	3.89	4.24	3.24	3.17	4.90	3.28	3.94	2.95	2.95
Ours									
DVLChat 7B	3.98	4.33	3.28	3.41	4.92	3.40	4.04	3.13	3.02

improves from 19.3% to 25.0% when scaling from 7B to 72B, InternVL3 peaks at 27.2% with its 14B variant before slightly declining to 27.1% with the 78B model. These non-monotonic scaling patterns in analytical tasks contrast sharply with the consistent improvements observed in captioning tasks, where larger models consistently achieve better performance in both regional and dense temporal captioning. This divergence in scaling behavior suggests that while model size benefits language generation and temporal narrative abilities, merely increasing parameters is insufficient for enhancing precise change detection and quantification capabilities. This is further evidenced by our DVLChat 7B outperforming larger models (up to 78B parameters) across multiple tasks when trained with domain-specific data. This highlights that incorporating strategies into domain-specific data is crucial for advancing model capabilities in understanding the analytical aspects of urban dynamics. We provide more analysis on scaling patterns by incorporating domain-specific data in Appendix § E.

Referring change detection analysis. We compare DVLChat with the specialist change-detection model ChangeMamba [3] and MLLM-based referring-segmentation models LISA [19] and PSALM [53], all fine-tuned on our dataset. As shown in Figure 10, ChangeMamba attains the highest IoU (32.41%) as a task-specific model trained on a fixed target ("new buildings"). Among MLLM-based methods, PSALM (26.93%) outperforms LISA (13.85%). DVLChat reaches 29.06% IoU, within 3.35% of the specialist. The qualitative results show DVLChat's tighter alignment with the ground truth than LISA/PSALM, especially around building boundaries and the spatial extent of new constructions.

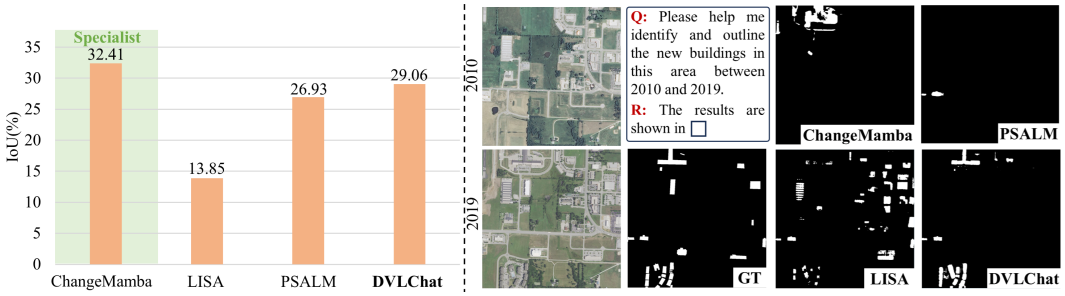

Figure 10: Performance comparison of specialist and generalist models on referring change detection.

https://doi.org/10.52202/085713-1114

6 Limitations and Future Directions

Several key directions remain for future exploration. First, DVL-Suite includes near-infrared band information from NAIP imagery, but the limited capabilities of current MLLMs in processing these spectral bands prevent their potential utilization, particularly for economic assessment tasks. Second, while DVLChat provides a unified baseline, it does not yet leverage pixel-level segmentation data to enhance numerical quantification across tasks. Finally, while DVLChat outperforms existing open-source models on most metrics, it still falls behind commercial models. Future work will focus on developing specialized algorithms and scaling up parameter size to bridge this performance gap.

7 Conclusion

In this paper, we present DVL-Suite, a large-scale vision-language benchmark for analyzing long-term urban dynamics through remote sensing imagery. Featuring 14,871 high-resolution multi-temporal images across 42 U.S. cities with detailed annotations spanning six urban understanding tasks, DVL-Suite enables systematic evaluation of MLLMs' capabilities from pixel-precise change detection to comprehensive temporal reasoning. Through extensive evaluation of 18 state-of-the-art models, we reveal critical insights: current MLLMs struggle significantly with long-term temporal understanding and quantitative analysis, while scaling model parameters alone proves insufficient without domain-specific training data. To address these limitations, we introduce DVL-Instruct, a specialized instruction-tuning dataset, and develop DVLChat as a baseline model that demonstrates substantial improvements, showcasing the potential of domain-specific data for advancing multi-temporal urban understanding capabilities.

Acknowledgements

This work was supported in part by the Council for Science, Technology and Innovation (CSTI) and the Cross-ministerial Strategic Innovation Promotion Program (SIP) "Development of a Resilient Smart Network System against Natural Disasters" (funding agency: NIED), KAKENHI (25K03145). This work was also supported by NVIDIA Academic Grant. This work used computational resources on the Miyabi supercomputer, provided by The University of Tokyo through the Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures and High Performance Computing Infrastructure in Japan (Project ID: jh250017). Weihao Xuan is supported by RIKEN Junior Research Associate (JRA) Program.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. Changemamba: Remote sensing change detection with spatio-temporal state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[4] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[6] Pei Deng, Wenqian Zhou, and Hanlin Wu. Changechat: An interactive model for remote sensing change analysis via multimodal instruction tuning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[7] Hosam Elgendy, Ahmed Sharshar, Ahmed Aboeitta, Yasser Ashraf, and Mohsen Guizani. Geollava: Efficient fine-tuned vision-language models for temporal change detection in remote sensing. *arXiv preprint arXiv:2410.19552*, 2024.

[8] Steve Frolking, Richa Mahtta, Tom Milliman, Thomas Esch, and Karen C Seto. Global urban structural growth shows a profound shift from spreading out to building up. *Nature Cities*, 1(9):555–566, 2024.

[9] Google DeepMind. Start building with gemini 2.5 flash. https://developers.googleblog.com/en/start-building-with-gemini-25-flash/, 2025.

[10] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019.

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[12] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

[13] Jeremy Andrew Irvin, Emily Ruoyu Liu, Joyce Chuyi Chen, Ines Dormoy, Jinyoung Kim, Samar Khanna, Zhuo Zheng, and Stefano Ermon. Teochat: A large vision-language assistant for temporal earth observation data. *arXiv preprint arXiv:2410.06234*, 2024.

[14] Neal Jean, Marshall Burke, Michael Xie, W Matthew Alampay Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[15] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[17] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024.

[18] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[19] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[21] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun MA, and Chunyuan Li. LLaVA-neXT-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[22] Ke Li, Fuyu Dong, Di Wang, Shaofeng Li, Quan Wang, Xinbo Gao, and Tat-Seng Chua. Show me what and where has changed? question answering and grounding for remote sensing change detection. *arXiv preprint arXiv:2410.23828*, 2024.

[23] Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[24] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024.

 https://doi.org/10.52202/085713-1114

[25] Chenyang Liu, Keyan Chen, Haotian Zhang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. Change-agent: Toward interactive comprehensive remote sensing change interpretation and analysis. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–1, 2024.

[26] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[27] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang W Chen. Et bench: Towards open-ended event-level video-language understanding. *Advances in Neural Information Processing Systems*, 37:32076–32110, 2024.

[28] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195.

[29] Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*, 2024.

[30] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *European Conference on Computer Vision*, pages 440–457. Springer, 2024.

[31] OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024.

[32] OpenAI. Gpt-4.1 mini. https://platform.openai.com/docs/models/gpt-4.1-mini, 2025.

[33] OpenAI. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/, 2025.

[34] OpenAI. Introducing openai o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, 2025.

[35] Markus B Pettersson, Mohammad Kakooei, Julia Ortheden, Fredrik D Johansson, and Adel Daoud. Time series of satellite imagery improve deep learning estimates of neighborhood-level poverty in africa. In *IJCAI*, pages 6165–6173, 2023.

[36] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024.

[37] Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, et al. Earthdial: Turning multi-sensory earth observations to interactive dialogues. *arXiv preprint arXiv:2412.15190*, 2024.

[38] Shiqi Tian, Ailong Ma, Zhuo Zheng, and Yanfei Zhong. Hi-ucd: A large-scale dataset for urban semantic change detection in remote sensing imagery. *arXiv preprint arXiv:2011.03247*, 2020.

[39] Junjue Wang, Weihao Xuan, Heli Qi, Zhihao Liu, Kunyi Liu, Yuhan Wu, Hongruixuan Chen, Jian Song, Junshi Xia, Zhuo Zheng, et al. Disasterm3: A remote sensing vision-language dataset for disaster damage assessment and response. *arXiv preprint arXiv:2505.21089*, 2025.

[40] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[41] Zhiming Wang, Mingze Wang, Sheng Xu, Yanjing Li, and Baochang Zhang. Ccexpert: Advancing mllm capability in remote sensing change captioning with difference-aware integration and a foundational dataset. *arXiv preprint arXiv:2411.11360*, 2024.

[42] Chen Wu, Bo Du, and Liangpei Zhang. Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9774–9788, 2023.

[43] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

[44] Aoran Xiao, Weihao Xuan, Junjue Wang, Jiaxing Huang, Dacheng Tao, Shijian Lu, and Naoto Yokoya. Foundation models for remote sensing and earth observation: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 2025.

[45] Zhitong Xiong, Fahong Zhang, Yi Wang, Yilei Shi, and Xiao Xiang Zhu. Earthnets: Empowering artificial intelligence for earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 2024.

[46] Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Semantic change detection with asymmetric siamese networks. *arXiv preprint arXiv:2010.05687*, 2020.

[47] Zhiwei Yang, Jian Peng, Yanxu Liu, Song Jiang, Xueyan Cheng, Xuebang Liu, Jianquan Dong, Tiantian Hua, and Xiaoyu Yu. Gloutci-m: a global monthly 1 km universal thermal climate index dataset from 2000 to 2022. *Earth System Science Data*, 16(5):2407–2424, 2024.

[48] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):2583, 2020.

[49] Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, Jianke Zhu, and Lidong Bing. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[50] Zhenghang Yuan, Lichao Mou, Zhitong Xiong, and Xiao Xiang Zhu. Change detection meets visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[51] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 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*, pages 9556–9567, 2024.

[52] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

[53] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pages 74–91. Springer, 2025.

[54] Zhuo Zheng, Yanfei Zhong, Liangpei Zhang, and Stefano Ermon. Segment any change. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[55] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

 https://doi.org/10.52202/085713-1114

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the contributions and scope of the paper. The abstract clearly outlines the creation of DVL-Suite, a comprehensive framework for analyzing long-term urban dynamics, which consists of DVL-Bench (a benchmark with 3,469 high-resolution multi-temporal images) and DVL-Instruct (a specialized instruction-tuning dataset). The introduction further elaborates on these claims by detailing the six essential capabilities for urban understanding and the framework's applications in diverse urban scenarios. These claims are supported by systematic evaluations of 18 state-of-the-art MLLMs and the development of DVLChat as a baseline model, which are thoroughly documented in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses two main limitations of the work. First, despite access to near-infrared band data from the NAIP platform, we acknowledge that current MLLMs' limited perception ability in this spectrum prevents them from leveraging this data to enhance accuracy, particularly in economic assessment tasks. Second, we point out that while DVLChat serves as a unified framework for change detection and question answering, it cannot effectively leverage pixel-level segmentation data to improve numerical quantification across multiple tasks. Both limitations are framed constructively as opportunities for future research.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

   Justification: This paper focuses on the dataset and benchmark, so it does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include all detailed metrics in the main text, while the implementation details of DVLChat and comprehensive prompting templates are provided in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

https://doi.org/10.52202/085713-1114

Justification: The dataset is publicly available at https://huggingface.co/datasets/weihao1115/dvl_suite. Additionally, we provide detailed evaluation code at https://github.com/weihao1115/dynamicvl.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All details are included in Section 5 and the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We conducted sensitivity analyses of various MLLMs with respect to different prompting strategies, and included these analyses in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).

- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provided the information in Section 5.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We carefully followed the ethics guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We detailed the contribution of this benchmark to society in the Introduction. We will also discuss this in the Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

https://doi.org/10.52202/085713-1114

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: We discussed this in the Appendix.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly cited all data sources and complied with data redistribution requirements.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new assets, DVL-Suite and DVLChat, are well documented in the paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We do not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: We do not involve such research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes] .

    Justification: We used GPT-4.1/mini for language refinement of annotation text and for generating diverse task instructions and evaluation prompts. LLMs were not used to create raw sensor data or to produce gold labels. All LLM-generated content was reviewed by humans.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-1114