

A Training and Evaluation Details

A.1 Training Details

We train the parameters of Low-Rank Adaptation (LoRA) modules for DVLChat’s segmentation and visual question answering branches independently based on the Qwen2.5-VL 7B model [2]. Subsequently, we employ the [SE] and [QA] prefixes within the input text to selectively activate specific branches, thereby enabling segmentation and question answering capabilities. The detailed training procedures for each branch are outlined below.

A.1.1 Segmentation Module

The images requiring segmentation are interleaved with temporal information (years) and processed through the vision encoder of Qwen2.5-VL to generate an embedding sequence. This sequence is then fed into the Large Language Model (LLM) component using teacher-forcing for next-token prediction. Following LISA [19], we apply LoRA fine-tuning only to the LLM while keeping Qwen2.5-VL’s vision encoder frozen. This design choice enables better integration with the visual question-answering module through prefix-based selection. During this process, the final hidden states corresponding to segmentation tokens in our instruction training data serve as text prompt embeddings. These embeddings are combined with features from the Segment Anything Model’s (SAM) frozen visual encoder and fed into the fine-tuned mask decoder of SAM [16] to generate the final segmentation results. The hyperparameters utilized in our experiments are detailed in Table 4.

Table 4: Training configuration for the segmentation branch across two stages.

Configuration	Stage 1	Stage 2
<i>Data Configurations</i>		
Dataset	LISA	DVL-Instruct (Seg Data)
<i>Model Configurations</i>		
Base Model	Qwen2.5-VL 7B	Qwen2.5-VL 7B
SAM Version	ViT-H	ViT-H
LoRA Rank	8	8
LoRA Alpha	16	16
LoRA Dropout	0.05	0.05
<i>Training Configurations</i>		
Hardware	8 H100 GPUs	8 H100 GPUs
Global Batch Size	128	128
Gradient Accumulation	4	4
Training Steps	5000	1000
<i>Optimizer Configurations</i>		
Optimizer	AdamW	AdamW
Learning Rate	3×10^{-4}	3×10^{-4}
β_1	0.9	0.9
β_2	0.95	0.95
Weight Decay	0.0	0.0
LR Scheduler	Cosine	Cosine
<i>Loss Configurations</i>		
Next-Token Loss Weight	1.0	1.0
Dice Loss Weight	0.5	0.5
BCE Loss Weight	2.0	2.0

A.1.2 Question-Answering Module

For the question-answering branch, we employ a standard LoRA fine-tuning strategy to optimize the LLM of Qwen2.5-VL. During this process, we freeze the vision encoder and fine-tune only the LLM, which enables alignment with the segmentation branch. The detailed parameters used in this branch are shown as follows.

Table 5: Training configuration for the question-answering branch.

Configuration	Details
<i>Data Configurations</i>	
Dataset	DVL Instruct (excluding Seg Data)
<i>Model Configurations</i>	
Base Model	Qwen2.5-VL 7B
LoRA Rank	64
LoRA Alpha	16
LoRA Dropout	0.05
<i>Training Configurations</i>	
Hardware	4 H100 GPUs
Global Batch Size	256
Training Duration	1 epoch
<i>Optimizer Configurations</i>	
Optimizer	AdamW
Learning Rate	2×10^{-4}
β_1	0.9
β_2	0.95
LR Scheduler	Cosine

A.2 Evaluation Details

Our evaluation uses the vLLM [18] library to accelerate model inference. We employ different image-text input formats for model inference based on the characteristics of each model. For the LLaVA-OneVision [20] model, since its AnyRes multi-image encoding approach generates excessively long context sequences on our test data, we treat multi-temporal images as video frames for single-frame encoding while placing temporal information as prefixes before the video input. For other models, we interleave temporal information with images, feed them to the tokenizer for tokenization, and then pass them to the models for inference. Except for the Qwen2.5-VL model, all other models use the default resizing solution for input images. For Qwen2.5-VL, since directly inputting original remote sensing images would generate excessively long context sequences, we uniformly resize all images to 504×504 pixels for memory efficiency.

B Prompts

In this section, we provide all the prompts used in our study. To standardize the output format, we employ prompts with relatively strict formatting constraints. To mitigate the significant performance variations that models exhibit across different prompts for the same type of question, we design our prompting strategy around distinct question categories. Each question is systematically fed into its corresponding prompt template, ultimately generating our final task prompts.

Additionally, we utilize evaluation prompts specifically designed to assess caption tasks. These evaluation prompts feed both the ground truth captions and the model-generated reports into GPT4.1-mini [32], which then compares them and assigns scores from 0 to 5, along with detailed reasoning for each score. This automated evaluation approach ensures consistent and objective assessment of caption quality across different models.

B.1 Task Prompts

Basic Change Analysis QA (Multi)

Instruction Template

Your task is to examine a chronological series of remote sensing images over a consistent geographical area, recorded across different years.
 Your objective is to unearth the changes and continuities to answer the question below. Review the provided choices and select all that align with your findings.
 $\langle \text{first_year} \rangle$: $\langle \text{image} \rangle$
 $\langle \text{second_year} \rangle$: $\langle \text{image} \rangle$

...
 <last_year>: <image>
Question: <question_prompt>
Choices: <options>
 Submit ONLY the capital letter(s) for your confirmed findings.
 • If a single finding is confirmed, answer its letter, such as “B”
 • If multiple findings are confirmed, list their letters separated by a comma and a space, such as “C, D, E”
 No further commentary or annotations are permitted in your answer.

? Question Prompts

- Which surface cover designations evolved from <start_year> through <end_year>?
- Which land-cover types shift during the period of <start_year>-<end_year>?
- Which categories of land surface have been transformed from <start_year> to <end_year>?
- Which types of ground cover experienced modifications from <start_year> to <end_year>?

Change Speed Estimation QA (Multi)

Instruction Template

Your task is to examine a chronological series of remote sensing images over a consistent geographical area, recorded across different years.
 Your objective is to analyze the speed and timing of changes to answer the question below. Review the provided choices and select all that align with your findings.
 <first_year>: <image>
 <second_year>: <image>
 ...
 <last_year>: <image>
Question: <question_prompt>
Choices: <options>
 Submit ONLY the capital letter(s) for your confirmed findings.
 • If a single finding is confirmed, answer its letter, such as “B”
 • If multiple findings are confirmed, list their letters separated by a comma and a space, such as “C, D, E”
 No further commentary or annotations are permitted in your answer.

? Question Prompts

- During what intervals did the expansion of <landuse_category> areas stop?
- During which time periods did <landuse_category> area expansion cease?
- Which eras showed a cessation in <landuse_category> area growth?
- In which timeframes did <landuse_category> development growth come to a halt?
- In what periods did the increase of <landuse_category> area pause?

Basic Change Analysis QA (Single)

Instruction Template

Your task is to examine a chronological series of remote sensing images over a consistent geographical area, recorded across different years.
 Your objective is to unearth the changes and continuities to answer the question below. Review the provided choices and select the one that aligns with your finding.
 <first_year>: <image>
 <second_year>: <image>
 ...
 <last_year>: <image>
Question: <question_prompt>
Choices: <options>

Submit ONLY the capital letter for your confirmed finding. Directly answer its letter, such as “B”.
No further commentary or annotations are permitted in your answer.

? Question Prompts

- What feature manifested the most considerable change over the years <start_year> to <end_year>?
- Which category underwent the most minimal modification between <start_year> and <end_year>?
- What showed the most significant transformation between <start_year> and <end_year>?
- Which element showed the least significant alteration from <start_year> to <end_year>?
- What transformations occurred in the <landuse_category> from <start_year> to <end_year>?
- How did the <landuse_category> evolve between <start_year> and <end_year>?

Change Speed Estimation QA (Single)

Instruction Template

Your task is to examine a chronological series of remote sensing images over a consistent geographical area, recorded across different years.

Your objective is to analyze the speed and timing of changes to answer the question below. Review the provided choices and select the one that aligns with your finding.

<first_year>: <image>

<second_year>: <image>

...

<last_year>: <image>

Question: <question_prompt>

Choices: <options>

Submit ONLY the capital letter for your confirmed finding. Directly answer its letter, such as “B”.
No further commentary or annotations are permitted in your answer.

? Question Prompts

- During which timeframe did the most substantial transformations occur?
- During which span of time were the most considerable modifications observed?
- How extensive was the overall transformation in this area between <start_year> and <end_year>, considering all changes?
- What was the cumulative extent of transformation in the whole region from <start_year> through <end_year>?
- What interval showed the most dramatic shifts?
- Which chronological period displayed the most profound changes?

Environmental Assessment

Instruction Template

Your task is to interpret the provided image and answer the question below.

From the potential choices offered, please identify the one that most accurately answers the question.

Image: <image>

Question: <question_prompt>

Choices: <options>

Please provide ONLY the capital letter corresponding to your answer choice, such as “C”, with no additional explanations.

? Question Prompts

- How much CO2 was emitted per day in this location in the year <year>?
- What was the average level of CO2 emissions in this area during <year>?
- Determine the average CO2 emissions in this area for the year <year>.
- What was the mean temperature in this region during <year>?
- What was the average temperature in this area during <year>?
- Determine the average temperature in this region for <year>.

Basic Change Analysis Report

Task Description

Your task is to examine a chronological series of remote sensing images over a consistent geographical area, recorded across different years. Your objective is to identify and report basic land cover changes.

<first_year>: <image>

<second_year>: <image>

...

<last_year>: <image>

Generate a concise summary (1-3 sentences) that adheres to the following:

1. **Time Period:** Begin by stating the full observation period (e.g., “Over the YYYY–YYYY period,” or “Over the X-year period from YYYY to YYYY,”).
2. **Key Transitions:** Detail the primary land cover transitions observed, focusing on changes between major land cover types, such as vegetated areas, non-vegetated surfaces, and built-up areas/buildings.
3. **Quantification:** Report these transitions using percentage ranges (e.g., X-Y%) or approximate percentages (e.g., ~X%).

Change Speed Estimation Report

Task Description

Your task is to examine a chronological series of remote sensing images over a consistent geographical area, recorded across different years. Your objective is to determine and report the building area changes between consecutive image dates, including expansions, shrinkages, and areas with no change.

<first_year>: <image>

<second_year>: <image>

...

<last_year>: <image>

Generate a report string that adheres to the following:

1. Calculate the building area change rates between consecutive years (including expansion, shrinkage, and no change periods)
2. Output the results in the following format (choose one of the following for each time period):
 - “The changes were as follows: X% expansion from [Start Year] to [End Year]” OR
 - “Y% shrinkage from [Start Year] to [End Year]” OR
 - “no significant change from [Start Year] to [End Year]”

Referring Change Detection

Task Prompts

- Please help me identify and segment the areas that changed from `<source_category>` to `<target_category>` between `<start_year>` and `<end_year>`.
- Please help me recognize and partition the territories that shifted from `<source_category>` to `<target_category>` during `<start_year>`-`<end_year>`.
- Please help me detect and delineate the regions that transformed from `<source_category>` to `<target_category>` within the `<start_year>`-`<end_year>` period.
- Please help me outline the areas converted from `<source_category>` to `<target_category>` within the `<start_year>`-`<end_year>` period.
- Please help me find and outline the sections where `<source_category>` were developed into `<target_category>` during `<start_year>`-`<end_year>`.

Regional Change Captioning

Task Description

Your task is to examine a chronological series of remote sensing images over a consistent geographical area, recorded across different years. Your objective is to densely describe all change events within the red-boxed area over the time period covered by these images.

`<first_year>`: `<image>`

`<second_year>`: `<image>`

...

`<last_year>`: `<image>`

Generate a description string that adheres to the following:

1. Segment descriptions by distinct year ranges (e.g., YYYY-YYYY).
2. For each period, detail the initial land cover, specific changes observed (construction, clearing, feature additions/removals), and explicitly note any stability.
3. Use concise, factual language focusing on visible features like vegetation, buildings, infrastructure, and land states, including spatial references where relevant.

Dense Temporal Captioning

Task Description

Your task is to examine a chronological series of remote sensing images over a consistent geographical area, recorded across different years. Your objective is to describe the dynamics of this area over the time period covered by these images.

`<first_year>`: `<image>`

`<second_year>`: `<image>`

...

`<last_year>`: `<image>`

Generate a description string that adheres to the following:

1. Start with a general summary statement outlining the main trend over the entire period.
2. Detail specific land feature changes chronologically, using clear spatial references (e.g., top-left, center-right).
3. Conclude with an optional overall summary of the changes or final state.

B.2 Evaluation Prompts

BCA-Report Evaluation

You are an advanced intelligent chatbot specialized in evaluating remote sensing image change detection results for basic land cover changes.

Your primary task is to meticulously compare the predicted change description with the ground truth description and assess their accuracy.

To accomplish this, you will evaluate the results across three key dimensions:

1. **Land Cover Type Identification Accuracy:**

Evaluate how accurately the predicted result identifies the main types of land cover changes, including:

- Correctly identifying the initial and final land cover types involved in the change
- Correctly describing the direction of transitions between different land cover types
- Capturing all major land cover transition patterns mentioned in the ground truth

2. Time Period Accuracy:

Assess how accurately the predicted result captures the correct time period mentioned in the ground truth result. This measures whether the overall start year and end year are correctly identified.

3. Change Quantification Accuracy:

Assess how well the predicted result quantifies the magnitude of changes, including:

- Accuracy of the reported percentage changes
- Alignment between predicted quantitative changes and ground truth transition percentages
- Completeness of quantitative information compared to ground truth

Please assign a score for each of these three dimensions, using an integer from 0 to 5, where 5 indicates perfect performance and 0 signifies poor performance. Accompany your assessments with brief explanations to clarify your scoring rationale.

Predicted Result: {predicted_result}

Ground Truth Result: {ground_truth_result}

CSE-Report Evaluation

You are an advanced intelligent chatbot specialized in evaluating remote sensing image change detection results.

Your primary task is to meticulously compare the predicted change detection results with the ground truth results and assess their accuracy. To accomplish this, you will evaluate the results across three key dimensions:

1. Change Rate Precision:

Evaluate how accurately the percentage value in the predicted result matches the actual change rate described in the ground truth result. This measures whether the predicted rate is correct (without considering other aspects).

2. Time Period Accuracy:

Assess how accurately the predicted result captures the correct time period mentioned in the ground truth result. This measures whether the start year and end year are correctly identified.

3. Change Pattern Accuracy:

Evaluate how accurately the predicted result describes the specific pattern and nature of changes, including:

- Correctly identifying the type of change (e.g., expansion, decrease, conversion)
- Accurately describing the spatial distribution or pattern of change
- Properly capturing the change dynamics mentioned in the ground truth

Please assign a score for each of these three dimensions, using an integer from 0 to 5, where 5 indicates perfect performance and 0 signifies poor performance. Accompany your assessments with brief explanations to clarify your scoring rationale.

Predicted Result: {predicted_result}

Ground Truth Result: {ground_truth_result}

DTC Evaluation

You are an advanced intelligent chatbot specialized in evaluating dense temporal captioning for remote sensing image time series.

Your primary task is to meticulously compare the predicted dense temporal caption with the ground truth caption and assess their accuracy. To accomplish this, you will evaluate the captions across three key dimensions:

1. Temporal Coverage:

Evaluate how well the predicted caption captures all significant time points and periods of change throughout the entire temporal range. This includes identifying key temporal milestones, maintaining a logical sequence of events, providing appropriate temporal context, and capturing the complete temporal narrative without significant gaps.

2. Spatial Accuracy:

Assess how accurately and comprehensively the predicted caption describes the spatial aspects of changes. This includes correctly identifying all regions where significant changes occurred, accurately

describing the spatial relationships between different areas, using precise spatial referencing, and ensuring comprehensive coverage of all spatially relevant changes in the image.

3. Process Fidelity:

Evaluate how accurately and completely the predicted caption describes the nature and processes of change. This includes correctly identifying initial and final land cover/use states, describing intermediate stages of development, capturing the complexity of multiple change processes, and accurately describing the specific features that changed.

Please assign a score for each of these three dimensions, using an integer from 0 to 5, where 5 indicates perfect performance and 0 signifies poor performance. Accompany your assessments with brief explanations to clarify your scoring rationale.

Predicted Caption: {predicted_caption}

Ground Truth Caption: {ground_truth_caption}

RCC Evaluation

You are an advanced intelligent chatbot specialized in evaluating regional captions for remote sensing images.

Your primary task is to meticulously compare the predicted regional caption with the ground truth regional caption and assess their accuracy. To accomplish this, you will evaluate the captions across four key dimensions:

1. Temporal Coverage:

Evaluate how well the predicted caption captures all significant time points and periods of change within the specified region. This includes identifying key temporal milestones, maintaining a logical sequence of events, and capturing the complete temporal narrative without significant gaps.

2. Spatial Accuracy:

Assess how accurately the predicted caption describes the spatial aspects of changes within the region. This includes correctly identifying sub-areas where changes occurred and accurately describing the spatial relationships between different features.

3. Process Fidelity:

Evaluate how accurately the predicted caption describes the nature and processes of change. This includes correctly identifying initial and final land cover/use states, describing intermediate stages of development, and accurately describing the specific features that changed.

4. Region Containment:

Assess whether the caption strictly focuses on changes within the specified region box only, without including irrelevant information about areas outside the designated region. This measures the ground truth caption's precision in adhering to the spatial boundaries defined by the region box.

Please assign a score for each of these four dimensions, using an integer from 0 to 5, where 5 indicates perfect performance and 0 signifies poor performance. Accompany your assessments with brief explanations to clarify your scoring rationale.

Predicted Caption: {predicted_caption}

Ground Truth Caption: {ground_truth_caption}

C Prompt Sensitivity

Prompt sensitivity analysis is crucial for reliable evaluation of multimodal language models, as different prompt formulations can significantly impact model performance and lead to inconsistent benchmarking results.

[Figure 11](#) shows the prompt sensitivity analysis conducted on five models: InternVL 78B, Qwen2.5-VL 72B, Qwen2.5-VL 7B, DVLChat, and GPT4.1. We evaluate these models across five tasks: BCA (single), BCA (multi), CSE (single), CSE (multi), and EA. Based on our original task prompts, we use GPT4.1 to generate four additional close but different prompt templates for each task, resulting in five prompt variations per task. We then test each model with all five prompt templates and compute the variance across these results to quantify prompt sensitivity for each model-task pair.

The results reveal several important patterns in prompt sensitivity across models and tasks. Most models demonstrate relatively low sensitivity (variance < 1%) for simpler single-choice tasks like BCA (single) and CSE (single), indicating robust performance regardless of prompt wording. However, the multiple-choice scenarios show dramatically different behavior. BCA (multi) exhibits the highest sensitivity overall, with Qwen2.5-VL 7B showing large variance (about 4.9%), suggesting this model is highly susceptible to prompt variations in complex classification tasks for such a small model. In comparison, our DVLChat, based upon Qwen2.5-VL 7B, shows improved robustness in most tasks, particularly demonstrating significantly lower variance in the challenging BCA (multi) task compared to its base model. While DVLChat exhibits slightly higher sensitivity in the EA task, it maintains more consistent performance across the majority of evaluation scenarios.

GPT4.1 demonstrates relatively consistent low sensitivity across all tasks, indicating its high robustness to prompt variance. This analysis highlights the importance of careful prompt design in model evaluation, particularly for complex multi-choice scenarios, and emphasizes the need to report prompt sensitivity when benchmarking multimodal language models. Overall, the relatively small performance variations across prompt templates for most model-task combinations demonstrate that our selected prompt templates provide an objective and fair evaluation of model capabilities.

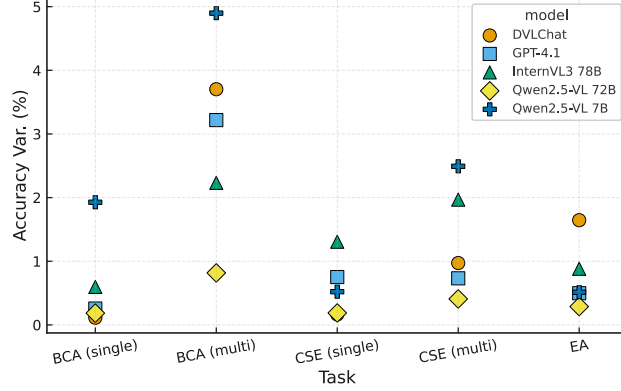


Figure 11: Prompt sensitivity analysis.

D Data Sources

D.1 Environmental Assessment

To enable a comprehensive environmental impact assessment, we incorporated five key urban indicators: population density, nighttime light intensity, CO₂ and O₃ emissions, and temperature. These indicators were sourced from authoritative databases: population data from WorldPop⁴, nighttime illumination from VIIRS Nighttime Day/Night Annual Band Composites V2.1⁵, carbon emissions from the Open-source Data Inventory for Anthropogenic CO₂⁶, and temperature from the GloUTCI-M product^[47].

D.2 Selection of NAIP Data

We compare available remote sensing imagery sources for long-term urban analysis in Table 6. NAIP uniquely satisfies four critical requirements: (1) public accessibility with redistribution rights, (2) extended temporal coverage exceeding 20 years, (3) high spatial resolution (0.3 to 1 m), and (4) sufficient geographic scope with diverse environmental contexts.

Commercial satellites, while offering global coverage and high resolution, impose licensing restrictions that prevent public dataset creation and redistribution. SpaceNet provides openly licensed imagery from multiple cities but lacks consistent temporal coverage at individual locations, preventing the tracking of long-term urban changes. Existing semantic change detection datasets such as SECOND^[46] and Hi-UCD^[38] provide only bi-temporal or tri-temporal snapshots, which are insufficient for comprehensive temporal analysis of urban dynamics.

Within NAIP’s coverage area, we selected 42 major cities distributed across the continental United States to maximize geographic and environmental diversity. These cities span different climatic zones, urban development patterns, and geographic contexts, from coastal to inland metropolitan areas. This selection provides broad representation while maintaining the temporal continuity necessary for analyzing urban evolution.

Table 6: Comparison of satellite imagery sources. NAIP offers the unique combination of public licensing, high resolution, and extended temporal coverage required for long-term urban analysis.

Data Source	Time Span	Temporal Depth	Resolution	Geographic Scope	License
Maxar/DigitalGlobe	2001–present	Continuous	31–60 cm	Global	Commercial
Pléiades/Neo	2012–present	Continuous	30–70 cm	Global	Commercial
SPOT-6/7	2012–present	Continuous	1.5 m	Global	Commercial
PlanetScope/Dove	2014–present	Continuous	3–4 m	Global	Commercial
SkySat	2013–present	Continuous	50 cm	Global	Commercial
RapidEye	2009–2020	Static	5 m	Global	Commercial
Jilin-1	2015–present	Continuous	0.5–0.75 m	Global	Commercial
Gaofen-2	2014–present	Programmatic	0.8 m	Limited access	Policy-managed
SpaceNet	2016–Present	Static	Varied	Multiple cities, limited temporal continuity per location	Varied
NAIP	2003–present	Continuous	0.3–1 m	Continental US	Public Domain

⁴<https://www.worldpop.org>

⁵https://developers.google.com/earth-engine/datasets/catalog/NOAA_VIIRS_DNB_ANNUAL_V21

⁶https://db.cger.nies.go.jp/dataset/ODIAC/DL_odiac2020.html

E More Results

More fine-tuning results. Table 7 presents the performance of models fine-tuned on DVL-Instruct. DVLChat consistently outperforms its base models across all tasks, with improvements spanning multiple-choice questions, report generation, captioning, and segmentation. The 7B model exemplifies these gains: DVLChat-7B achieves 33.3% average accuracy on multiple-choice questions (+10.0% over base model), 2.99 average score on report generation tasks (+0.66), 3.69 average score on caption tasks (+0.66), and 29.1 cIoU on segmentation. Performance scales with model size, as DVLChat-32B achieves the strongest results across most metrics (42.1% MCQ accuracy, 4.22 RCC score, 3.68 DTC score). These results demonstrate the effectiveness of DVL-Instruct for training comprehensive dynamic remote sensing analysis capabilities.

Table 7: Performance of models fine-tuned on DVL-Instruct and evaluated on DVL-Bench. MCQ: Average accuracy across five multiple-choice question tasks. BCA-Report and CSE-Report: Averaged report generation scores (0-5 scale) for Basic Change Analysis and Change Speed Estimation tasks, respectively. RCC and DTC: Averaged caption scores (0-5 scale) for the Regional Change Captioning and Dense Temporal Captioning tasks, respectively. cIoU (Seg): Cumulative Intersection over Union for referring change detection segmentation.

Models	MCQ	BCA-Report	CSE-Report	RCC	DTC	cIoU (Seg)
Qwen2.5-VL 3B	24.7	2.99	1.72	2.76	2.38	-
DVLChat 3B (Qwen2.5-VL)	31.6	3.43	2.25	3.78	3.19	26.9
Qwen2.5-VL 7B	23.3	2.94	1.73	3.21	2.85	-
DVLChat 7B (Qwen2.5-VL)	33.3	3.47	2.51	3.98	3.40	29.1
InternVL3 8B	23.9	2.99	2.15	3.69	2.97	-
DVLChat 8B (InternVL3)	37.6	3.65	2.41	3.83	3.42	30.7
InternVL3 14B	27.2	3.02	2.36	3.96	3.22	-
DVLChat 14B (InternVL3)	41.3	3.68	2.61	3.99	3.48	31.3
Qwen2.5-VL 32B	31.4	3.04	2.60	3.90	2.91	-
DVLChat 32B (Qwen2.5-VL)	42.1	3.60	2.65	4.22	3.68	34.7

Transferability to other benchmarks. We investigate whether improvements of DVLChat observed on DVL-Bench extend to other vision-language tasks. As shown in Table 8, DVLChat 7B consistently outperforms its base model (Qwen2.5-VL 7B) across multiple benchmarks: remote sensing visual question answering (VRS-Bench), image classification (GeoChat with UC Merced and AID datasets), and temporal reasoning (TEOChatlas). The improvements are particularly notable on the TEOChat RTQA task (+4.3%), demonstrating that training on DVL-Instruct not only enhances long-term urban understanding but also strengthens general remote sensing and temporal reasoning capabilities.

Table 8: Performance on other vision-language benchmarks. RQA in TEOChat comprises two subtasks from xBD and S2Looking, with results separated by slash (/).

Method	VRS-Bench [23]		GeoChat [17]		TEOChat [13]	
	VQA (Acc %)	Captioning (CLAIR)	UC Merced (Acc %)	AID (Acc %)	RQA (Acc %)	RTQA (Acc %)
Qwen2.5-VL 7B	43.5	0.6374	72.3	64.6	76.7/87.3	59.9
DVLChat 7B	44.7	0.6790	74.4	66.0	91.5/90.0	64.2

F More Discussions

F.1 Data

Evaluation Methodology and Precision. Our evaluation methodology accounts for potential uncertainties in pixel-level annotations through metrics that emphasize overall accuracy rather than pixel-perfect precision. We employ two primary metrics. Change Quantification Accuracy (BCA reports) measures the alignment between predicted and ground truth transition percentages, as well as the completeness of quantitative information. Change Rate Precision (CSE reports) evaluates how closely predicted percentage values match the actual change rates in ground truth results. The choice of two decimal precision is motivated by the spatial characteristics of our dataset. Given the 1-meter ground sampling distance (GSD) and 1024×1024 pixel images, each image covers over 1 square kilometer of land area. At this scale, a 0.01% area change represents more than 100 square meters of actual land area. This level of precision enables the detection of small-scale land use changes that are relevant for practical applications. Our evaluation approach provides a consistent comparative assessment across different models by applying uniform conditions and annotation standards to all algorithms. This design allows

for reliable measurement of quantitative reasoning capabilities while accommodating the inherent variability in pixel-level annotations.

Dataset Design and Compatibility. DVL-Suite adopts land cover type conventions that are widely used in remote sensing research, providing compatibility with existing datasets such as SECOND [46]. The categorical scheme differs from SECOND in one aspect: "tree" and "low vegetation" are consolidated into a unified "vegetation" class to reduce potential annotation ambiguity in high-resolution imagery. All other category definitions align with established conventions. This categorical design facilitates cross-dataset research applications. The compatibility enables domain adaptation studies between different datasets and geographic regions. Researchers can leverage the shared categorical framework to investigate model transferability across diverse spatial and temporal contexts. DVL-Suite employs a multi-temporal framework that differs from SECOND's bi-temporal structure. The dataset includes image sequences that support flexible temporal combinations, such as adjacent pairs or first-last frame comparisons. For studies requiring bi-temporal configurations, the first and last frames can be extracted to create maximum temporal separation while maintaining compatibility with bi-temporal benchmarks. The dataset encompasses a larger scale and greater temporal diversity compared to existing bi-temporal semantic change detection datasets. While DVL-Suite is designed primarily as a vision-language benchmark, its scale and temporal richness also support traditional semantic change detection research.

Economic Assessment Task. Economic assessment using remote sensing imagery represents an established research area with demonstrated feasibility. Previous studies have shown that satellite imagery can be used to predict economic indicators when combined with machine learning approaches. Early work demonstrated poverty prediction from satellite data [14], while subsequent research has extended these methods to assess economic well-being across different geographic regions [48, 35, 35]. The underlying imagery in our dataset is derived from NAIP, which captures near-infrared (NIR) bands in addition to RGB channels. NIR data enhances the distinction between land cover types such as farmland, forests, water bodies, and urban structures, supporting more detailed analysis of land use changes and regional development patterns. The Economic Assessment task in DVL-Suite serves two purposes. First, it evaluates current multimodal large language models on a task that connects semantic change detection with economic interpretation. Second, it establishes a benchmark for future work that may incorporate additional spectral information beyond the RGB channels, ensuring compatibility with existing vision-language models. The task results reveal current capabilities and limitations in this domain, providing direction for subsequent research.

Potential Applications. DVL-Suite provides long-term temporal coverage spanning at most 18 years, high-resolution imagery at 1024×1024 pixels, and multi-level analysis capabilities. These characteristics support applications in urban planning, disaster response, and environmental monitoring. The dataset enables analysis of historical development patterns, infrastructure vulnerability assessment, and urban sustainability factors, including heat island effects and green space development. The vision-language model paradigm differs from traditional segmentation-then-analysis workflows by enabling natural language interactions with remote sensing data. This approach reduces the technical barrier for non-specialist users, including policymakers and community stakeholders, to access and interpret complex spatial analysis results. The natural language interface also facilitates communication of remote sensing insights across different user groups with varying levels of technical expertise.

F.2 Experiments

Optimization conflicts between VQA and RCD. Preliminary experiments with LISA [19] on multi-temporal urban understanding tasks revealed instruction-following inconsistencies across different task types. For single-choice questions, the model sometimes generated responses outside the provided option set (e.g., "O." instead of valid options A-E). For multiple-choice questions, outputs usually included sequences such as "O, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19" rather than coherent answer combinations. For open-ended reporting tasks, the model often defaulted to segmentation-style responses (e.g., "Sure, it is [SEG]") instead of generating descriptive analysis. These behaviors indicate an optimization conflict between segmentation and language generation objectives in the unified training framework. The segmentation branch influences the optimization process in ways that affect instruction-following behavior across diverse question types. This interference pattern is particularly evident in tasks requiring free-form text generation rather than mask prediction. DVLChat addresses this challenge through task-specific LoRA [11] modules with explicit routing mechanisms. This architecture separates the optimization paths for question-answering and segmentation tasks, reducing task interference. The design enables the model to perform both multi-temporal urban analysis and change detection within a single framework while maintaining appropriate output formats for each task type.

G Qualitative Visualizations

Figure 12 presents representative qualitative samples from the DVL-Bench dataset, focusing on two critical tasks: referring change detection and dense temporal captioning. The figure demonstrates a typical urban transformation from 2010 to 2021, where agricultural land evolves into residential areas. The DTC responses reveal significant

disparities in model capabilities. LLaVA-OneVision 7B fails to pinpoint specific years accurately and provides overly broad temporal descriptions, lacking the fine-grained analysis required for precise urban monitoring. TeoChat 7B, limited by its training data TEOChatlas, which lacks captioning capabilities, essentially loses the ability to generate long-context outputs. Its brief response reveals constraints imposed by its instruction tuning dataset, with limited land-cover change categories and poor spatial localization. While Qwen2.5-VL 7B can generate longer contextual descriptions, it exhibits weak temporal awareness and imprecise localization abilities. In contrast, our DVLChat demonstrates superior performance in both tasks: the segmentation results show the most accurate change detection masks, while the DTC output precisely identifies temporal boundaries and accurately recognizes land-cover types with specific spatial references (e.g., "top-right-side field"). This comprehensive superiority validates how our DVL-Instruct dataset effectively enhances models' capabilities in multi-temporal understanding, fine-grained temporal reasoning, and precise spatial-semantic understanding of urban dynamics.

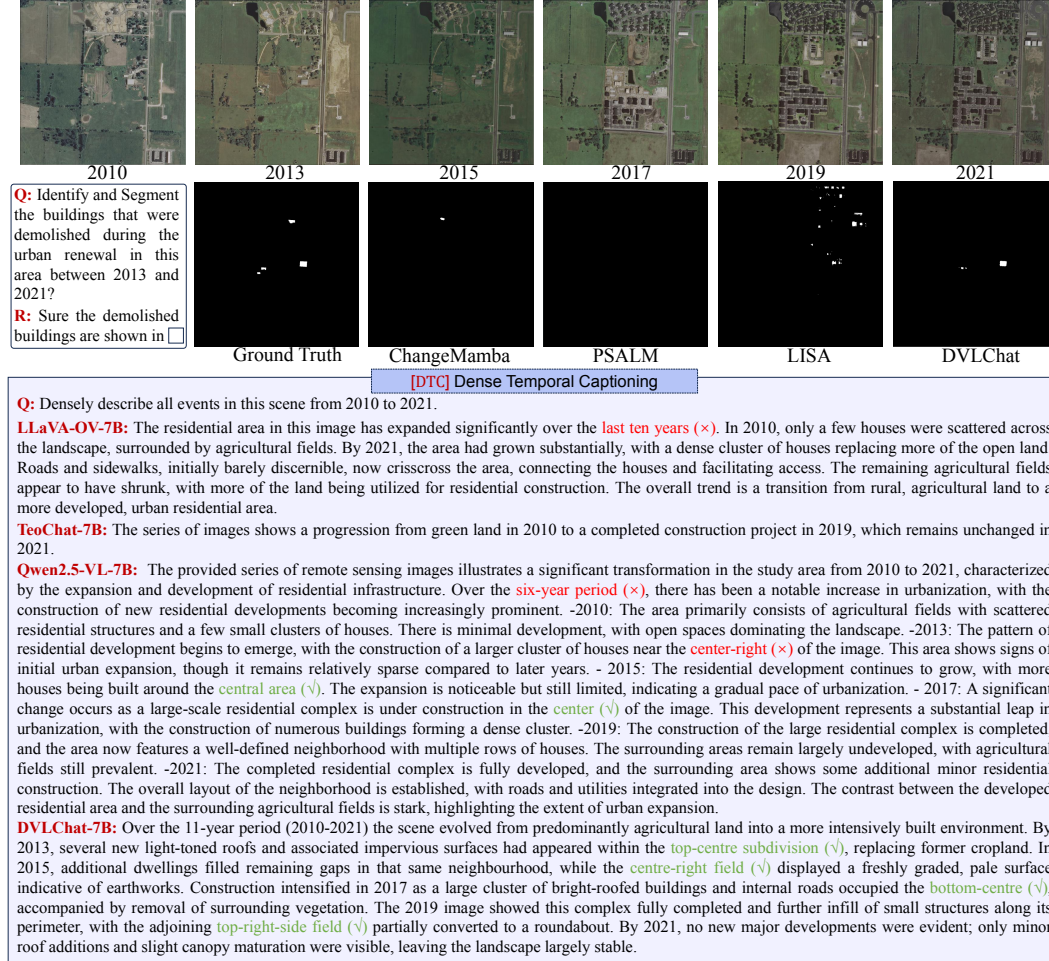


Figure 12: Typical answering samples in an urban renovation scene.

As shown in [Figure 13](#), there are critical differences in model capabilities for fine-grained urban change analysis across multiple tasks. The multiple-choice questions reveal significant performance disparities between models. For environmental assessment, general models struggle with quantitative CO2 emission estimates, and their answers vary widely across options. In contrast, DVLChat's selections align more closely with the ground truth, demonstrating its effectiveness in assessing environments using the remote sensing imagery.

The change speed estimation task further illustrates these differences in challenging fine-grained urban understanding. Different from dense temporal captioning tasks, TeoChat-7B produces well-formatted responses that follow the required structure, but it struggles with accuracy and makes numerical errors when interpreting changes. Qwen2.5-VL 7B performs even worse, failing to detect any changes in the image sequence. Despite clear visual transformations, this model consistently reports "no significant change" across all time periods. This behavior highlights a fundamental weakness: general-purpose vision-language models cannot detect small-scale changes at the pixel level, which severely limits their use in urban analysis applications. Our DVLChat model,

however, successfully identifies these subtle changes and provides numerical estimates that closely match the actual values.

These consistent advantages across both answer selection and text generation tasks demonstrate that DVL-Instruct successfully connects high-level understanding with the precise quantitative analysis needed for urban monitoring applications.

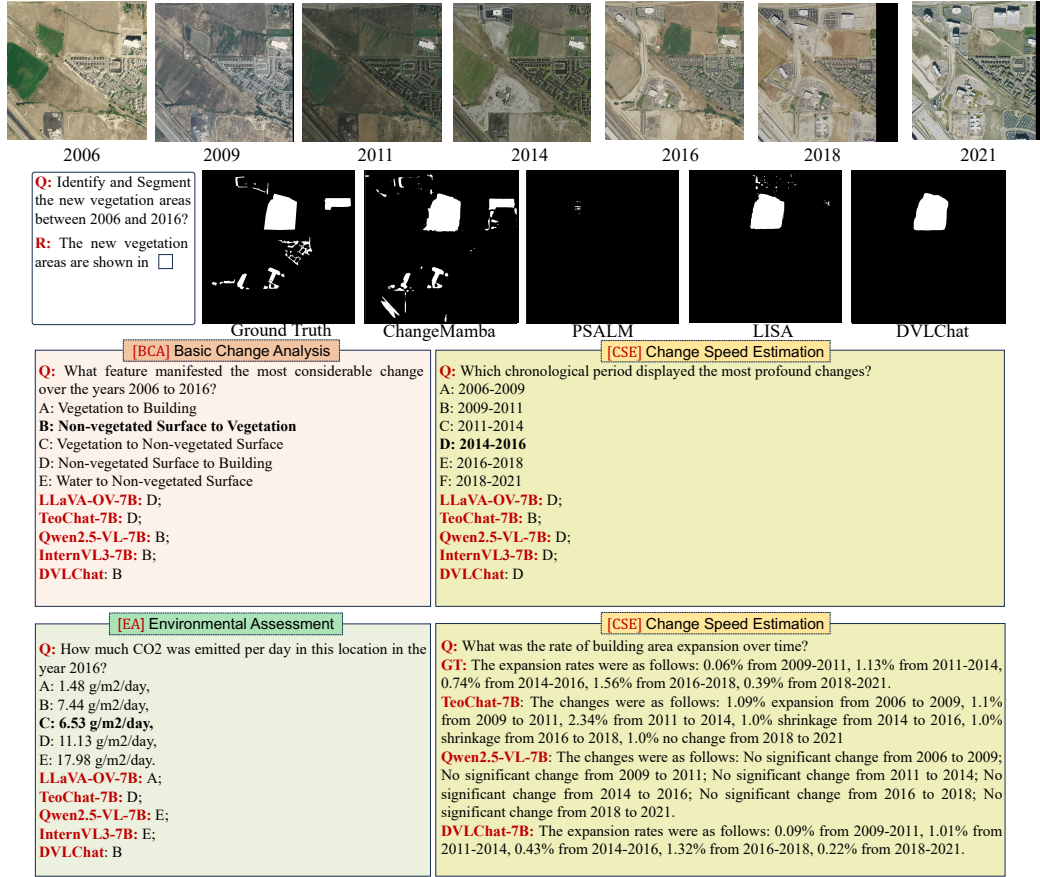


Figure 13: Typical answer examples in residential greening scene.

H Broader Impacts

The DVL-Suite has significant potential for positive societal impact by advancing sustainable urban development through comprehensive long-term urban dynamics analysis. By enabling multimodal AI systems to understand complex urban changes over 18-year periods, this work could help urban planners and policymakers make more informed decisions about sustainable development, climate adaptation, and disaster resilience, directly supporting UN Sustainable Development Goal 11 for sustainable cities and communities. The framework's ability to analyze environmental indicators, urban expansion patterns, and infrastructure changes could accelerate evidence-based urban planning and enhance cities' capacity to address climate change impacts. However, there are risks associated with high-resolution satellite imagery analysis, including potential privacy and surveillance concerns, as well as the possibility that over-reliance on automated urban analysis could lead to misguided planning decisions that overlook important social and cultural contexts. We emphasize that DVL-Suite should be used as a complementary tool that enhances human expertise in urban planning, with careful attention to incorporating community perspectives and local knowledge in urban development processes.