SURD-BENCH: BENCHMARKING UNIFIED MULTIMODAL MODELS FOR URBAN SPATIOTEMPORAL COGNITION

Anonymous authors
Paper under double-blind review

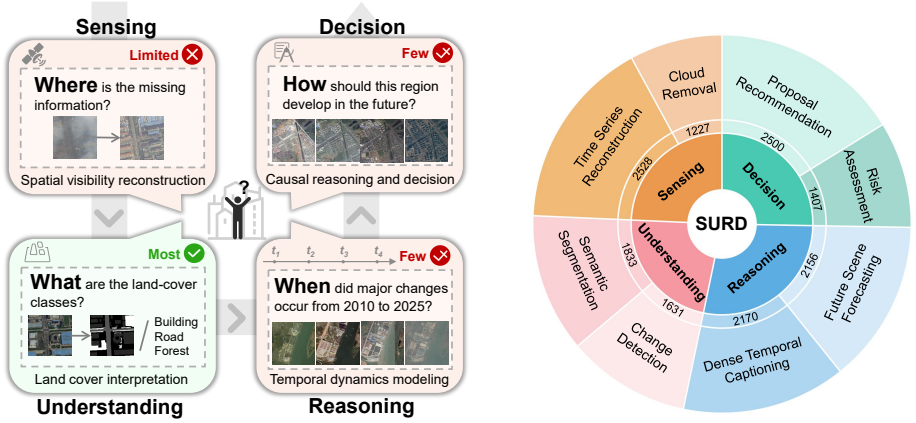
ABSTRACT

Understanding and modeling urban dynamics from Earth observation data is critical for urban remote sensing. Achieving this goal requires models to perceive urban objects in their spatial context, understand how they evolve over time, and predict their future states. Recent advances in Unified Multimodal Models (UMMs) have provided new opportunities to advance urban remote sensing from task-specific recognition toward comprehensive spatiotemporal cognition. However, existing benchmarks are mainly designed for general visual understanding or specific remote sensing tasks, and are not suitable for evaluating the unified understanding and generation capabilities of UMMs in urban remote sensing. The capability boundaries of UMMs in urban spatiotemporal cognition remain insufficiently characterized. Inspired by human cognitive processes, we propose SURD-Bench around four cognitive stages: Sensing, Understanding, Reasoning, and Decision, for evaluating UMMs in urban spatiotemporal cognition. In constructing SURD-Bench, we integrate in-house and public datasets, develop eight comprehensive tasks through an AI-assisted curation workflow with human verification, and design a unified evaluation framework for diverse outputs. Using SURD-Bench, we evaluate 18 representative commercial and open-source multimodal models. The results show that even the best-performing commercial model, Gemini-3.1-image, achieves an average score of only 41.85%, indicating that current UMMs still face substantial challenges in urban spatiotemporal cognition. Further analysis identifies limitations in fine-grained pixel-level semantic understanding, spatiotemporal relationship modeling, and the integration of understanding and generation capabilities. These findings offer insights into the future development of remote sensing UMMs.

1 INTRODUCTION

Urban environments are dynamic systems characterized by complex spatial structures and continuous evolution (Frolking et al., 2024). Facing complex urban environments, humans do not comprehend them through a single observation but gradually develop a comprehensive view through a continuous cognitive process (Peelen et al., 2024; Bengio, 2017). Specifically, humans first complete partial observations to identify **where** targets are located, then interpret **what** they represent, further analyze **when** changes occur based on historical observations, and ultimately infer **how** the urban environment may develop in the future. This cognitive capability, evolving from spatial perception and semantic understanding to temporal reasoning and future inference, is essential for modeling dynamic urban environments. With the rapid development of high-resolution and multitemporal Earth observation data, urban remote sensing (RS) is also transitioning from target-specific information extraction toward a new stage of urban spatiotemporal cognition (Wu et al., 2025b). Such spatiotemporal cognition contributes to revealing urban development patterns, supporting urban planning, and improving resource management (Wang et al., 2026a; Zheng et al., 2025). Previous RS tasks are typically designed for specific objectives, primarily answering the question of what exists in the city (Zhu et al., 2017; Zhang et al., 2020). Rapidly changing urban systems exhibit pronounced spatiotemporal complexity, and relying solely on isolated tasks is insufficient for establishing a comprehensive cognition of urban states, evolutionary processes, and future trends.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Cognitive Gaps. (b) SURD-Bench Tasks.

Figure 1: Motivation and overview of SURD-Bench for urban spatiotemporal cognition. (a) Gaps in existing remote sensing benchmarks across complementary stages of urban spatiotemporal cognition. (b) Task taxonomy of SURD-Bench covering sensing, understanding, reasoning, and decision capabilities.

With the breakthroughs of Multimodal Large Language Models (MLLMs) in visual perception and the advances of diffusion models in high-quality generation, researchers have begun to break the boundary between understanding and generation, developing Unified Multimodal Models (UMMs) that seamlessly integrate both capabilities within a unified framework (Yang et al., 2025). Their capabilities in visual perception, language understanding, and long-context modeling open up new possibilities for uncovering urban evolution patterns from Earth observation data (Zhao et al., 2025; Ning et al., 2025). As UMMs capabilities advance rapidly, a critical question arises: *what are the capability boundaries of UMMs in urban spatiotemporal cognition?*

A primary obstacle in answering this question is the lack of a suitable benchmark. Although recent studies (Xie et al., 2026; Shi et al., 2026) have begun to evaluate UMMs' capabilities in understanding and generation, and several benchmarks (An et al., 2025; Wang et al., 2025a; Xuan et al., 2025; Luo et al., 2026) have also been developed for RS scenarios, existing evaluation frameworks still suffer from two major limitations (Fig. 1):

- **Incomplete coverage of urban spatiotemporal cognition.** Although existing RS benchmarks have covered diverse task forms, such as scene description, object recognition, and visual question answering, simply expanding the task set cannot capture the continuous spatiotemporal evolution of urban environments. These tasks are typically organized around single observations or isolated change events, overlooking the fine-grained spatial structures and complex temporal relationships embedded in RS data. Such designs are insufficient to provide a comprehensive evaluation covering the full scope of urban spatiotemporal cognition, from spatial perception and state understanding to evolution reasoning and future prediction.
- **Misaligned evaluation of unified understanding and generation capabilities.** Existing RS benchmarks mainly focus on model understanding capabilities and lack systematic evaluation of the synergy between understanding and generation. Moreover, current evaluation paradigms heavily rely on discriminative tasks, such as multiple-choice questions and simple VQA, which are susceptible to language priors and answer biases, thereby failing to accurately characterize models' comprehensive multimodal capabilities.

To address the above limitations, we propose SURD-Bench, a benchmark for evaluating UMMs on urban spatiotemporal cognition. Inspired by human cognition of urban environments (Collins et al., 2024; Hassabis et al., 2017), we construct a four-stage cognitive framework consisting of Sensing, Understanding, Reasoning, and Decision. This framework conceptualizes urban spatiotemporal cognition as a progressive cognition chain, from fine-grained spatial perception and state under-

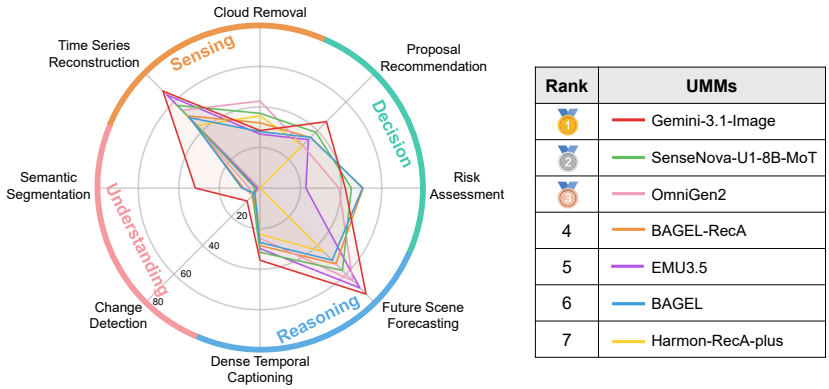


Figure 2: Performance comparison of seven UMMs on SURD-Bench.

standing to spatiotemporal evolution reasoning and decision-oriented generation. Guided by this cognition chain, SURD-Bench develops eight task-specific evaluations with clear objectives, enabling a systematic assessment of UMM capabilities at different stages under a unified taxonomy and evaluation protocol. To accommodate diverse output formats across different tasks, we further establish a unified evaluation protocol that supports heterogeneous tasks within a common framework and provides a comprehensive assessment of UMMs' capabilities in urban spatiotemporal cognition (Fig. 2). SURD-Bench is designed to facilitate the evolution of RS foundation models from task-specific visual recognition toward urban spatiotemporal cognition and world modeling.

Our main contributions are summarized as follows:

- We propose SURD-Bench, the first benchmark for evaluating UMMs on urban spatiotemporal cognition, covering four cognitive stages from spatial perception and understanding to reasoning and decision-making.
- We design a benchmark construction and evaluation framework tailored to urban spatiotemporal cognition. By integrating multi-source RS data, AI-assisted task construction, and human verification, we build high-quality evaluation data and establish a unified evaluation scheme.
- We evaluate 18 popular multimodal models on SURD-Bench. By analyzing performance differences across cognitive stages, task types, and model paradigms, we characterize the capabilities and limitations of current models in urban spatiotemporal cognition and identify directions for future development.

2 Related Work

2.1 Unified Multimodal Models

UMMs achieve end-to-end multimodal understanding and generation through unified representation spaces and training paradigms. Models such as EMU3 (Wang et al., 2026b) and Janus (Wu et al., 2025a) adopt autoregressive approaches to achieve unified modeling of text and images, while BAGEL (Deng et al., 2025) further explores the application of diffusion mechanisms in UMMs based on the Mixture-of-Transformers (MoT) architecture. Sensenova U1 (Diao et al., 2026) explores a native unified multimodal architecture that moves from modality integration toward true modality unification by removing the Visual Encoder (VE) and Variational Auto-Encoder (VAE). In the RS domain, early studies (Muhtar et al., 2024; Kuckreja et al., 2024; Zhan et al., 2025) mainly integrated MLLMs with Earth observation data to support intelligent RS interpretation tasks. Furthermore, Uni-RS (Zhang et al., 2026a) addresses the spatial reversal curse of UMMs in RS scenarios, while RS-WorldModel (Xu et al., 2026) explores the application of UMMs in RS change understanding and future scene prediction. With the continuous advancement of UMMs, their unified

Table 1: Comparison of SURD-Bench with existing multimodal benchmarks, where TF=True/False, MCQ=Multiple Choice Questions, BBox=Bounding Box, FF=Free-Form, Seg=Segmentation Mask, Img=Image Generation, Temp.=Temporal Modeling, Recon.=Fine-grained Reconstruction, Forec.=Future Forecasting, Open=Open-ended, and U&G=Unified Understanding and Generation, ✓=Supported, ✗=Unsupported, and ❍=Partially Supported.

Benchmark	Domain	Answer Type	Temp.	Recon.	Forec.	Open	U&G
MME (Fu et al., 2025)	General	TF	✗	✗	✗	✗	✗
MMBench (Liu et al., 2024)	General	MCQ	✗	✗	✗	✗	✗
MMMU (Yue et al., 2024)	General	MCQ + FF	✗	✗	✗	❍	✗
SEED-Bench (Li et al., 2024)	General	MCQ	✗	✗	✓	✗	✗
SEED-Bench-2 (Li et al., 2023)	General	MCQ	✗	✗	✓	✗	✗
MME-RealWorld (Zhang et al., 2025b)	General	MCQ	✗	✗	✓	✗	✗
MME-Unify (Xie et al., 2026)	General	MCQ + Img	✗	✓	✓	❍	✓
RealUnify (Shi et al., 2026)	General	MCQ + Img	✗	✗	✗	❍	✓
GEOBench-VLM (Danish et al., 2025)	RS	MCQ, BBox, Seg, FF	✓	✗	✗	❍	✗
CHOICE (An et al., 2025)	RS	MCQ, BBox, Seg	❍	✗	✗	❍	✗
XLRS-Bench (Wang et al., 2025a)	RS	MCQ, BBox, FF	❍	✗	✗	❍	✗
VLRS-Bench (Luo et al., 2026)	RS	MCQ, FF, TF	✓	✗	✓	❍	✗
UrBench (Zhou et al., 2025)	RS	MCQ, FF	✗	✗	✗	❍	✗
DVL-Bench (Xuan et al., 2025)	RS	MCQ, FF, Seg	✓	✗	✗	❍	✗
SURD-Bench (Ours)	RS	FF, Img, Seg	✓	✓	✓	✓	✓

understanding and generation capabilities open new opportunities for discovering urban evolution patterns from high-resolution, multi-temporal Earth observation data.

2.2 Multimodal Benchmarks

Early efforts including MME (Fu et al., 2025), MMBench (Liu et al., 2024), and MMMU (Yue et al., 2024) mainly focus on models' perception and reasoning capabilities. Subsequently, SEED-Bench (Li et al., 2024) and MME-RealWorld (Zhang et al., 2025b) further extend evaluation toward multi-image understanding and complex real-world scenarios. With the development of UMMs, MME-Unify (Xie et al., 2026) introduces an evaluation framework covering understanding, generation, and unified understanding and generation tasks. UniEval (Li et al., 2025a) introduces a self-evaluation paradigm that leverages UMMs' own capabilities to assess generated outputs. RealUnify (Shi et al., 2026) further explores the bidirectional synergy between UMMs' understanding and generation capabilities. In the RS domain, studies such as GEOBench-VLM (Danish et al., 2025), CHOICE (An et al., 2025), XLRS-Bench (Wang et al., 2025a), and DVL-Bench (Xuan et al., 2025) explore multimodal evaluation in Earth observation scenarios, gradually extending from scene understanding toward high-resolution and multi-temporal imagery analysis. However, although existing RS benchmarks cover various understanding and reasoning tasks, they typically assess these capabilities separately, without a unified framework for systematically characterizing the diverse cognitive requirements of urban evolution, from recovering incomplete spatial observations to interpreting changing urban states, inferring temporal dynamics, and informing decisions. They also lack evaluation paradigms tailored to the combined understanding and generation capabilities of UMMs.

3 SURD-Bench

3.1 Overview

SURD-Bench establishes a unified evaluation framework for urban spatiotemporal cognition to assess UMMs' capabilities in modeling dynamic urban environments. The framework characterizes urban spatiotemporal cognition along four complementary stages: Sensing, Understanding, Reasoning, and Decision, corresponding to spatial information recovery, semantic interpretation, temporal reasoning, and decision-oriented analysis. Together, these stages support a systematic characterization of UMMs' cognitive capabilities in complex urban environments. Based on high-resolution, multi-temporal, and multi-source urban observation data, SURD-Bench covers eight tasks across

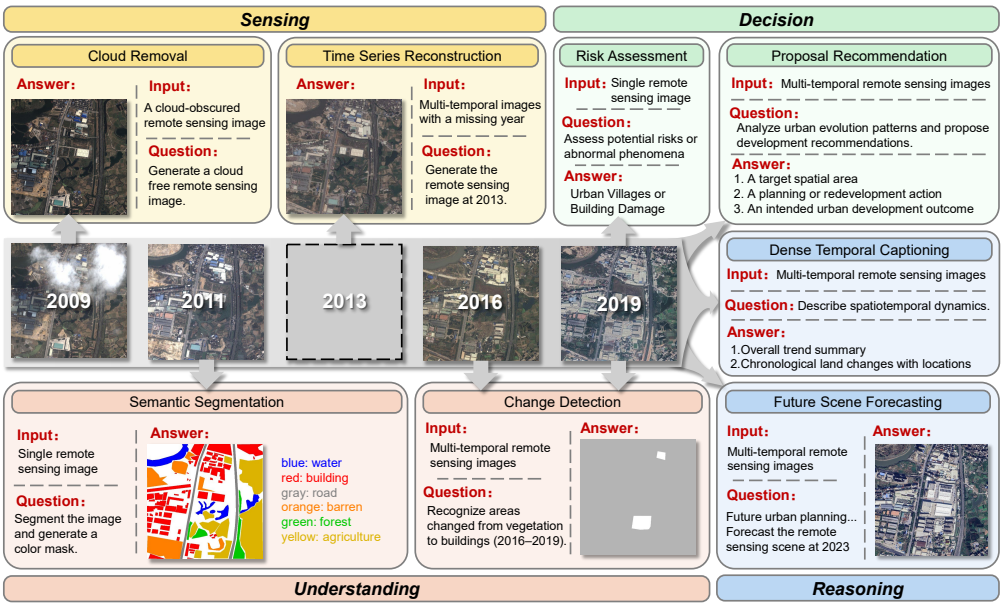


Figure 3: Illustration of SURD-Bench tasks through an urban remote sensing time-series example, covering single-image, multi-temporal, and temporal-missing cases across four cognitive stages: sensing, understanding, reasoning, and decision.

these four stages (Fig. 1(b)) and introduces unified evaluation protocols to accommodate diverse task outputs. SURD-Bench contains 3,685 test samples and provides an additional 11,767 training samples to facilitate model fine-tuning by the research community and encourage future research on remote sensing UMMs. Table 1 presents a comparison between SURD-Bench and existing multi-modal benchmarks.

3.2 Benchmark Tasks

SURD-Bench develops eight tasks across four cognitive stages: Sensing, Understanding, Reasoning, and Decision, as illustrated in Fig. 3. Tasks are assigned according to the primary cognitive objective represented by their target outputs. Sensing recovers incomplete spatial or temporal observations to establish an observable urban scene. Understanding identifies the semantic composition of urban scenes and localizes their changes. Reasoning organizes multi-temporal evidence to describe urban evolution and infer future states. Decision transforms urban observations into risk-aware assessments and action-oriented planning recommendations. Detailed task definitions and example questions are provided in the Appendix.

Sensing. This stage evaluates models' capability to reconstruct fine-grained visual information from urban RS imagery. As the foundation of urban spatiotemporal cognition, this stage examines whether models can recover meaningful spatial information from incomplete Earth observation data while preserving the structural characteristics of high-resolution RS imagery. Accordingly, this stage includes two task types: spatial inpainting and temporal inpainting. The spatial inpainting task focuses on Cloud Removal (CR) in RS, covering both thin and thick cloud scenarios. It requires models to reconstruct underlying surface information in cloud-covered regions. Inspired by UniTS (Zhang et al., 2026b), the temporal inpainting task introduces Time Series Reconstruction (TSR), where models receive an incomplete temporal sequence of images (3–5 timestamps) and are required to reconstruct missing observations, evaluating their ability to model spatiotemporal continuity and latent states dynamics.

Understanding. Moving beyond visual information recovery, this stage evaluates models' ability to interpret spatial semantics and their changes in urban RS scenarios. This stage comprises two tasks: semantic segmentation (SS) and change detection (CD). For the SS task, in addition to requiring models to perform pixel-level classification of major land-cover categories in RS imagery, we

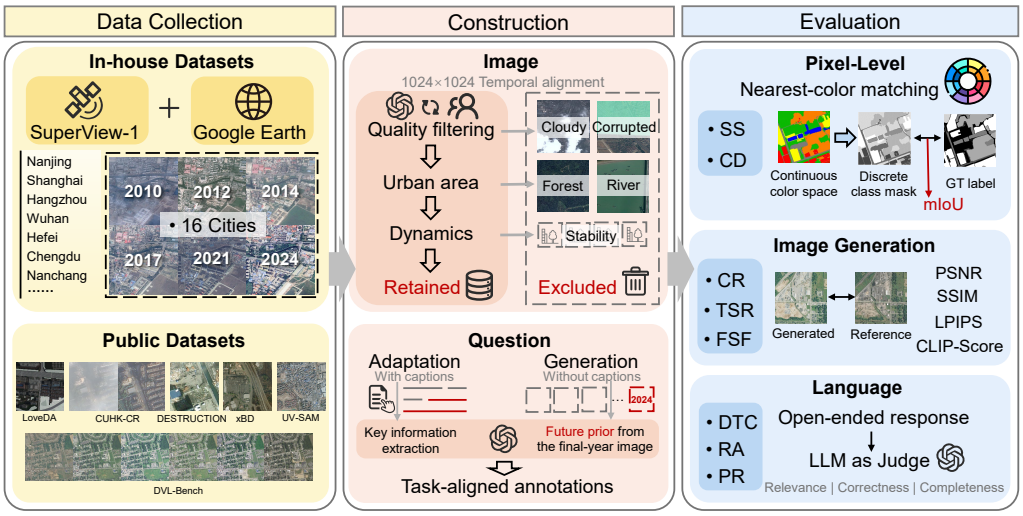


Figure 4: The curation pipeline of SURD-Bench, including data collection, task-specific construction, and comprehensive evaluation.

introduce a simplified binary semantic segmentation (BSS) task, where models extract foreground regions based on specified object categories to analyze the boundaries of their segmentation capabilities. The CD task requires models to identify changed regions of specified object categories at specific time points from multi-temporal RS imagery to assess their understanding of urban spatial change patterns and temporal differences.

Reasoning. This stage evaluates models' ability to capture urban evolution patterns from temporal observations and forecast future states. This stage comprises two tasks: dense temporal captioning (DTC) and future scene forecasting (FSF). In the DTC task, models generate descriptions of regional dynamic changes based on continuous temporal RS imagery, including major change types, occurrence times, spatial locations, and evolution trends. In the FSF task, models generate future RS scenes at future timestamps based on historical temporal imagery and future scene constraints, evaluating their ability to extrapolate urban evolution trends.

Decision. This stage targets urban management and planning applications, evaluating whether models can transform RS visual information into high-level decision support from the perspective of "how we should act". In the Risk Assessment (RA) task, models are required to identify potential urban risks in RS imagery, covering two representative scenarios: building damage assessment and urban village identification. Unlike existing discriminative tasks that rely on predefined risk categories or candidate options, this task adopts an open-ended evaluation setting, requiring models to autonomously identify and describe potential risks based on observations. In the Proposal Recommendation (PR) task, models are required to generate targeted urban planning recommendations by considering current urban states, historical evolution trends, and planning objectives, providing decision support for urban renewal and spatial optimization.

3.3 Benchmark Curation

The curation pipeline of SURD-Bench is illustrated in Fig. 4. It consists of data collection, AI-assisted filtering and task construction with human verification, and task-specific evaluation.

Data Collection. SURD-Bench consists of in-house and public datasets. The in-house datasets are constructed from SuperView-1 high-resolution satellite imagery and historical Google Earth imagery across 16 cities, containing multi-temporal observations spanning 4–6 timestamps to characterize urban spatial structures and dynamic changes. To cover diverse urban RS scenarios, SURD-Bench further integrates several public datasets, including LoveDA (Wang et al., 2021) for high-resolution RS image semantic segmentation, CUHK-CR (Sui et al., 2024) for cloud removal, DESTRUCTION (Ali et al., 2020) and xBD (Gupta et al., 2019) for building damage assessment, UV-

Table 2: Performance comparison of multimodal models on SURD-Bench across four cognitive levels and eight tasks. The evaluated models include open-source models, commercial models*, and *remote sensing-specific models*. The bold and underlined values denote best and second-best performance for each task, respectively. "-" indicates that the model does not support the corresponding task. Avg represents the average score across the eight tasks. Task abbreviations: CR = Cloud Removal; TSR = Time Series Reconstruction; SS = Semantic Segmentation; CD = Change Detection; DTC = Dense Temporal Captioning; FSF = Future Scene Forecasting; RA = Risk Assessment; PR = Proposal Recommendation.

Model	Sensing		Understanding		Reasoning		Decision		Avg
	CR	TSR	SS	CD	DTC	FSF	RA	PR	
Understanding Models									
TEOChat	-	-	-	-	10.04	-	7.96	22.47	5.06
VHM	-	-	-	-	18.13	-	22.11	28.44	8.59
LHRS-Bot-nova	-	-	-	-	16.76	-	26.13	30.93	9.23
ScoreRS w/ RL	-	-	-	-	17.93	-	35.42	29.68	10.38
InternVL3.5-8B	-	-	-	-	32.55	-	50.56	35.67	14.85
Qwen3-VL-8B-Instruct	-	-	-	-	<u>36.44</u>	-	57.61	<u>43.85</u>	17.24
GPT-4o*	-	-	-	-	**43.46**	-	**77.25**	41.64	20.29
Generative Models									
Qwen-Image-2.0-Pro*	31.62	-	14.79	-	-	-	-	-	5.80
FLUX.2	33.15	<u>66.34</u>	4.17	5.09	-	<u>71.72</u>	-	-	22.56
GPT-Image-2*	28.00	66.19	<u>24.08</u>	<u>6.37</u>	-	71.11	-	-	24.47
Unified Models									
Janus-Pro	-	-	-	-	10.92	-	<u>62.54</u>	32.20	13.21
Harmon-RecA-plus	35.58	43.78	1.97	3.45	22.83	44.45	0.00	29.61	22.71
BAGEL	27.83	49.09	8.97	3.58	26.84	50.31	50.56	35.39	31.57
EMU3.5	26.57	64.98	1.30	5.09	29.81	69.66	22.61	33.81	31.73
Bagel-RecA	32.17	50.34	8.14	4.74	28.37	52.89	50.85	35.20	32.84
OmniGen2	**42.87**	54.22	5.29	5.57	24.76	64.68	38.94	29.79	33.27
SenseNova-U1-8B-MoT	<u>36.86</u>	57.63	2.48	5.12	31.65	57.49	45.07	38.91	<u>34.40</u>
Gemini-3.1-image*	28.35	**67.65**	**31.89**	**9.00**	35.60	**73.93**	42.04	**46.32**	**41.85**

SAM (Zhang et al., 2024) for urban village identification, and DVL-Bench (Xuan et al., 2025) for urban dynamic understanding.

Construction. In this stage, we further perform image filtering and task annotation based on the collected multi-source RS data. During image filtering, we adopt an AI-assisted curation workflow, in which GPT-4o (Hurst et al., 2024) performs primary filtering and spatiotemporal standardization of the RS imagery with manual verification. The filtering criteria include image quality, urban area coverage, and temporal dynamics. All images are uniformly cropped to 1024×1024 pixels. Samples with severe cloud occlusion, image corruption, or insufficient temporal variation are excluded, while those with meaningful visual information and spatiotemporal change characteristics are retained. Subsequently, during task annotation, we adopt different question construction strategies for different data sources. For data with existing annotations, key information is extracted to generate task-specific descriptions. For data without original annotations, the final timestamp is used as a future prior to construct descriptions for prediction tasks. Detailed construction procedures for DTC, RA, and PR are provided in the Appendix. These processes ultimately produce multimodal evaluation samples aligned with different task objectives.

Evaluation. We design three evaluation strategies for diverse output formats. (1) Pixel-level evaluation. For SS and CD tasks, models are required to generate segmentation results corresponding to the target categories. Since generative models typically produce images in a continuous space whereas ground-truth labels are represented as discrete category masks, the two cannot be directly compared. Therefore, we propose a nearest-color matching method that computes the Euclidean distance between the generated output and each category label color in a three-dimensional color space, and maps each pixel to the nearest category to convert it into a discrete mask. The resulting mask is then compared with the ground-truth label to compute the mean Intersection over Union (mIoU). (2) Image generation evaluation. For CR, TSR, and FSF tasks, models are required to

generate images representing a target state based on existing observations. According to the requirements of each task, we select evaluation metrics along two dimensions, visual fidelity and semantic consistency, including PSNR, SSIM, LPIPS (Zhang et al., 2018), and CLIP-Score (Hessel et al., 2021). (3) Language evaluation. For DTC, RA, and PR tasks, models generate open-ended natural language responses. We adopt the LLM-as-Judge paradigm and use GPT-5-mini (OpenAI, 2025) as the evaluator. The evaluation scheme scores model outputs from three dimensions: relevance, correctness, and completeness. It also generates explanations for each score to improve interpretability. More details of the evaluation protocol are provided in the Appendix.

4 EXPERIMENTS

4.1 EVALUATION SETUP

Evaluated Models. We evaluate 18 MLLMs and UMMs, including RS specific models, general understanding models, image generation models, and unified multimodal models. RS specific models include TEOChat (Irvin et al., 2025), VHM (Pang et al., 2025), LHRS-Bot-nova (Li et al., 2025b), and ScoreRS fine-tuned with RL (Muhtar et al., 2025). Understanding models include InternVL3.5 (Wang et al., 2025b), Qwen3-VL-Instruct (Bai et al., 2025), and GPT-4o (Hurst et al., 2024). Generative models include FLUX.2 (Black Forest Labs, 2025), Qwen-Image-2.0-Pro (Zhao et al., 2026), and GPT-Image-2 (OpenAI, 2026). UMMs include BAGEL (Deng et al., 2025), Janus-Pro (Chen et al., 2025), EMU3.5 (Cui et al., 2025), OmniGen2 (Wu et al., 2026), SenseNova-U1 (Diao et al., 2026), Gemini-3.1-image (Google DeepMind, 2026), as well as BAGEL-RecA and Harmon-RecA-plus, two models fine-tuned with RecA (Xie et al., 2025).

Evaluation Protocols. All models are evaluated using the LMMs-Eval (Zhang et al., 2025a) framework with temperature set to 0, employing greedy decoding and the same zero-shot prompts across all tasks. For task-specific evaluation metrics, we normalize all metrics to a range of 0–100 and convert them into unified positive scores. The final score for each task is obtained by averaging all metrics, ensuring comparability across different tasks. Detailed score calculation procedures are provided in Appendix.

4.2 MAIN RESULTS

Overall Challenges in Urban Remote Sensing Cognition. Table 2 summarizes the overall performance of different models on SURD-Bench. We find that current multimodal models still exhibit limited capabilities in urban spatiotemporal cognition. The best-performing commercial model, Gemini-3.1-image (Google DeepMind, 2026), achieves an average score of only 41.85%, while the best open-source model, SenseNova-U1-8B-MoT (Diao et al., 2026), reaches 34.40%. Nearly all models perform substantially worse on Understanding tasks than on other tasks, indicating that current models still struggle with fine-grained semantic understanding in high-resolution RS scenarios. Furthermore, for image generation tasks, some models fail to produce high-quality images consistent with RS characteristics and may even generate content inconsistent with textual conditions, suggesting that current generation modules still face challenges in following textual constraints. Overall, although current MLLMs and UMMs have demonstrated strong general visual understanding and generation capabilities, these abilities do not fully transfer to urban spatiotemporal cognition, leaving a substantial gap between general-purpose foundation models and practical urban applications.

Semantic Segmentation Capability. We further compare the performance of Generative Models and UMMs on multi-class SS and BSS, as shown in Fig. 5(a). The results show that although current models achieve relatively low performance on SS, all models obtain substantial improvements when the task is simplified to BSS, with SenseNova-U1-8B-MoT (Diao et al., 2026) and Gemini-3.1-image (Google DeepMind, 2026) improving by 11.18 and 10.86 points respectively. These gains suggest that reducing the label space alleviates cross-class ambiguity, allowing models to focus on identifying a single target category. In multi-class SS, models must distinguish land-cover classes with similar appearances while preserving accurate boundaries and local spatial layouts. Qualitative results in the Appendix show that some models can preserve plausible local boundaries and spatial structures, while their land-cover assignments still deviate from the reference masks. Overall, the low SS scores reflect the joint challenge of fine-grained semantic discrimination and spatially precise pixel-level generation in complex urban RS scenes.

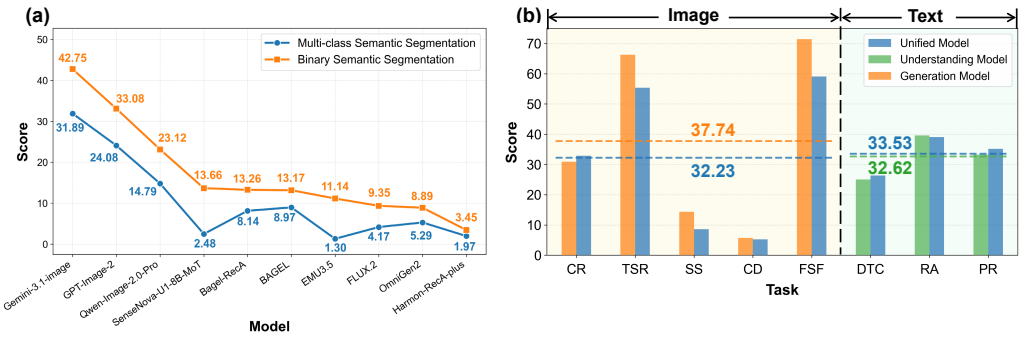


Figure 5: Performance comparison of multimodal models across diverse task categories. (a) Performance comparison on multi-class and binary semantic segmentation tasks. (b) Average performance comparison of different model paradigms on image and language generation tasks.

When Temporal Generation Meets Limited Understanding. SURD-Bench covers tasks involving both spatial understanding and temporal evolution. As shown in Table 2, some models achieve strong performance on TSR and FSF tasks. For example, Gemini-3.1-image (Google DeepMind, 2026) achieves 67.65% and 73.93% on TSR and FSF, respectively, while BAGEL (Deng et al., 2025) reaches 50.31% on FSF, indicating that current models have acquired preliminary capabilities in simulating temporal evolution. However, model performance drops substantially on CD, which requires explicit understanding of temporal changes. All models achieve scores below 10%, with Gemini-3.1-image obtaining the highest score of only 9.00%. This discrepancy suggests that current models can capture temporal patterns but struggle to associate temporal changes with spatial semantics. In other words, models can generate visually plausible states at a given time, but struggle to answer fine-grained questions such as "where did the changes occur". We attribute this phenomenon to the different requirements of temporal generation and dynamic understanding. Temporal generation mainly relies on learning statistical patterns of appearance variations over time, whereas dynamic understanding requires integrating spatial structures and change contexts for semantic reasoning, as well as localizing change regions with pixel-level precision. Therefore, the key challenge for current UMMs lies not only in modeling temporal information, but more importantly in effectively linking it to spatial semantics and the pixel-level locations of changes.

Does Unified Modeling Enable Comprehensive Multimodal Cognition? As shown in Fig. 5(b), UMMs do not demonstrate clear advantages over specialized models on SURD-Bench. For image generation tasks, UMMs generally underperform specialized Generation Models, while for language generation tasks, UMMs only slightly outperform Understanding Models with limited improvements. These results indicate that although unified modeling expands the range of supported tasks, simply integrating understanding and generation capabilities does not necessarily lead to stronger comprehensive cognition. Achieving effective synergy between generation and understanding capabilities remains a key challenge for current UMMs.

5 Conclusion

We introduce SURD-Bench, a benchmark for evaluating UMMs' urban spatiotemporal cognition. By integrating in-house and public datasets, SURD-Bench establishes a unified evaluation framework spanning four cognitive stages: Sensing, Understanding, Reasoning, and Decision, with evaluation protocols designed for different outputs. Extensive experiments on 18 models reveal key limitations of current models in urban remote sensing scenarios. Although existing models demonstrate visual perception and generation capabilities, they remain limited in complex scene understanding, spatiotemporal relationship modeling, and the integration of generation and understanding, revealing a substantial gap toward comprehensive urban cognition. We hope that SURD-Bench can facilitate the development of multimodal models for urban remote sensing and advance multimodal intelligence from general visual understanding toward geospatial cognition.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. The study does not involve human participants or animal experiments. All datasets used in this work were obtained and used in accordance with their applicable usage guidelines. No personally identifiable information was used in this work. We took care to minimize potential sources of bias throughout the research process and are committed to maintaining transparency, rigor, and research integrity.

SURD-Bench is intended for research evaluation. Its outputs should not be treated as operational risk assessments or planning decisions without further validation and appropriate human oversight.

REPRODUCIBILITY STATEMENT

The Appendix provides dataset statistics, task definitions, prompt templates, language-reference construction procedures, evaluation metrics, model settings, detailed results, and qualitative examples. Public data sources are identified through their original citations. All evaluated models use standardized zero-shot prompts and task-specific evaluation protocols.

LLM USAGE

Large Language Models (LLMs) were used as auxiliary tools during the preparation and evaluation of this work. Their use included editorial assistance, such as language refinement, grammar checking, sentence rephrasing, and improving clarity and readability. Following the widely adopted LLM-as-a-Judge paradigm, GPT-5-mini was used for the evaluation of language-oriented tasks.

LLMs were also used to assist dataset construction, including AI-assisted image filtering, spatiotemporal standardization, and the generation of planning-oriented reference proposals.

LLMs were not used to determine the research objectives, formulate the core benchmark framework, or draw the scientific conclusions of this work.

The authors take full responsibility for the manuscript, benchmark construction, experimental results, and conclusions presented in this work.

REFERENCES

Muhammad Usman Ali, Waqas Sultani, and Mohsen Ali. Destruction from sky: Weakly supervised approach for destruction detection in satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162: 115–124, 2020.

Xiao An, Jiaxing Sun, Zihan Gui, and Wei He. Choice: Benchmarking the remote sensing capabilities of large vision-language models. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-4396.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report, 2025. arXiv:2511.21631.

Yoshua Bengio. The consciousness prior, 2017. arXiv:1709.08568.

Black Forest Labs. Flux.2 klein: Towards interactive visual intelligence. `https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence`, 2025. Accessed: 2025-07-25.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. arXiv:2501.17811.

Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, et al. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10):1851–1863, 2024.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3.5: Native multimodal models are world learners, 2025. arXiv:2510.26583.

Muhammad Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7132–7142, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining, 2025. arXiv:2505.14683.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture, 2026. arXiv:2605.12500.

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.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-4899.

Google DeepMind. Gemini 3.1 flash image. `https://deepmind.google/models/model-cards/gemini-3-1-flash-image/`, 2026. Accessed: 2026-07-26.

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, 2019. arXiv:1911.09296.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7514–7528, 2021.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card, 2024. arXiv:2410.21276.

Jeremy Irvin, Emily Liu, Joyce Chen, Ines Dormoy, Jinyoung Kim, Samar Khanna, Zhuo Zheng, and Stefano Ermon. Teochat: A large vision-language assistant for temporal earth observation data. In *International Conference on Learning Representations (ICLR)*, pp. 68883–68911, 2025.

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 (CVPR)*, pp. 27831–27840, 2024.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models, 2023. arXiv:2311.17092.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308, 2024.

Yi Li, Haonan Wang, Qixiang Zhang, Boyu Xiao, Chenchang Hu, Hualiang Wang, and Xiaomeng Li. Unieval: Unified holistic evaluation for unified multimodal understanding and generation, 2025a. arXiv:2505.10483.

Zhenshi Li, Dilxat Muhtar, Feng Gu, Yanglangxing He, Xueliang Zhang, Pengfeng Xiao, Guangjun He, and Xiaoxiang Zhu. Lhrs-bot-nova: Improved multimodal large language model for remote sensing vision-language interpretation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227:539–550, 2025b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*, pp. 216–233. Springer, 2024.

Zhiming Luo, Di Wang, Haonan Guo, Jing Zhang, and Bo Du. Vlrs-bench: A vision-language reasoning benchmark for remote sensing, 2026. arXiv:2602.07045.

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 (ECCV)*, pp. 440–457. Springer, 2024.

Dilxat Muhtar, Enzhuo Zhang, Zhenshi Li, Feng Gu, Yanglangxing He, Pengfeng Xiao, and Xueliang Zhang. Quality-driven curation of remote sensing vision-language data via learned scoring models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 171088–171133. Curran Associates, Inc., 2025. doi: 10.52202/085713-5699.

Junzhi Ning, Wei Li, Cheng Tang, Jiashi Lin, Chenglong Ma, Chaoyang Zhang, Jiyao Liu, Ying Chen, Shujian Gao, Yuandong Pu, et al. Unimedvl: Unifying medical multimodal understanding and generation through observation-knowledge-analysis. In *International Conference on Machine Learning (ICML)*, 2025.

OpenAI. Gpt-5 system card, 2025. arXiv:2601.03267.

OpenAI. Introducing chatgpt images 2.0. `https://openai.com/index/introducing-chatgpt-images-2-0/`, 2026. Accessed: 2026-07-26.

Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, et al. Vhm: Versatile and honest vision language model for remote sensing image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6381–6388, 2025.

Marius V Peelen, Eva Berlot, and Floris P de Lange. Predictive processing of scenes and objects. *Nature Reviews Psychology*, 3(1):13–26, 2024.

Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, et al. Realunify: Do unified models truly benefit from unification? a comprehensive benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22488–22497, 2026.

Jialu Sui, Yiyang Ma, Wenhan Yang, Xiaokang Zhang, Man-On Pun, and Jiaying Liu. Diffusion enhancement for cloud removal in ultra-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14325–14336, 2025a.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Junjue Wang, Weihao Xuan, Heli Qi, Zihang Chen, Hongruixuan Chen, Zhuo Zheng, Junshi Xia, Yanfei Zhong, and Naoto Yokoya. Cityvlm: Towards sustainable urban development via multi-view coordinated vision–language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 232:62–74, 2026a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. arXiv:2508.18265.

Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Zhen Li, Yuqi Wang, et al. Multimodal learning with next-token prediction for large multimodal models. *Nature*, pp. 1–7, 2026b.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977, 2025a.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, et al. Omnigen2: Towards instruction-aligned multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21964–21975, 2026.

Kang Wu, Yingying Zhang, Lixiang Ru, Bo Dang, Jiangwei Lao, Lei Yu, Junwei Luo, Zifan Zhu, Yue Sun, Jiahao Zhang, et al. A semantic-enhanced multi-modal remote sensing foundation model for earth observation. *Nature Machine Intelligence*, 7(8):1235–1249, 2025b.

Ji Xie, Trevor Darrell, Luke Zettlemoyer, and Xudong Wang. Reconstruction alignment improves unified multimodal models, 2025. arXiv:2509.07295.

Wulin Xie, YiFan Zhang, Chaoyou Fu, Yang Shi, Jianshu Zeng, Bingyan Nie, Hongkai Chen, Zhang Zhang, and Liang Wang. MME-unify: A comprehensive benchmark for unified multimodal understanding and generation models. In *International Conference on Learning Representations (ICLR)*, 2026.

Linrui Xu, Zhongan Wang, Fei Shen, Gang Xu, Huiping Zhuang, Ming Li, and Haifeng Li. Rs-worldmodel: A unified model for remote sensing understanding and future sense forecasting, 2026. arXiv:2603.14941.

Weihao Xuan, Junjue Wang, Heli Qi, Zihang Chen, Zhuo Zheng, Yanfei Zhong, Junshi Xia, and Naoto Yokoya. Dynamicvl: Benchmarking multimodal large language models for dynamic city understanding. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-1114.

Yan Yang, Haochen Tian, Yang Shi, Wulin Xie, Yi-Fan Zhang, Yuhao Dong, Yibo Hu, Liang Wang, Ran He, Caifeng Shan, et al. A survey of unified multimodal understanding and generation: Advances and challenges. *Authorea Preprints*, 2025.

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 (CVPR)*, pp. 9556–9567, 2024.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221: 64–77, 2025.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025a.

Mingyang Zhang, Tong Li, Yue Yu, Yong Li, Pan Hui, and Yu Zheng. Urban anomaly analytics: Description, detection, and prediction. *IEEE Transactions on Big Data*, 8(3):809–826, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Weiyu Zhang, Yuan Hu, Yong Li, and Yu Liu. Uni-rs: A spatially faithful unified understanding and generation model for remote sensing, 2026a. arXiv:2601.17673.

Xin Zhang, Yu Liu, Yuming Lin, Qingmin Liao, and Yong Li. Uv-sam: Adapting segment anything model for urban village identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 22520–22528, 2024.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations (ICLR)*, pp. 89655–89701, 2025b.

Yuxiang Zhang, Shunlin Liang, Wenyuan Li, Han Ma, Jianglei Xu, Yichuan Ma, Jiangwei Xie, Wei Li, Mengmeng Zhang, Ran Tao, and Xiang-Gen Xia. Units: Unified spatio-temporal generative model for remote sensing, 2026b. arXiv:2512.04461.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report, 2026. arXiv:2605.10730.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1702–1713, 2025.

Yu Zheng, Xu Fengli, Yuming Lin, Paolo Santi, Carlo Ratti, Li Yong, et al. Urban planning in the era of large language models. *Nature Computational Science*, 5(9):727–736, 2025.

Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10707–10715, 2025.

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.

APPENDIX

A. OVERVIEW

This appendix provides supplementary materials for SURD-Bench. Section B presents benchmark details, including dataset statistics, task definitions, and prompt design. Section C describes the evaluation protocol, including metrics and model settings. Section D reports additional experimental results, while Section E provides visualization and case studies.

B. MORE DETAILS OF SURD-BENCH

B.1 DATASET STATISTICS

Table 3 summarizes the statistics of SURD-Bench across the four cognitive stages and eight tasks.

Table 3: Dataset statistics of SURD-Bench.

Stage	Task	Image Size	Output Type	Train	Test	Total
Sensing	CR	512	Image	982	245	1227
	TSR	1024	Image	2022	506	2528
Understanding	SS	1024	Mask	1156	677	1833
	CD	1024	Mask	1142	489	1631
Reasoning	DTC	1024	Text	1678	492	2170
	FSF	1024	Image	1664	492	2156
Decision	RA	1024	Text	1123	284	1407
	PR	1024	Text	2000	500	2500
Total				**11767**	**3685**	**15452**

B.2 TASK DEFINITIONS

Each task in SURD-Bench is formulated as an instruction-following problem:

$$(I, [C], P) \rightarrow O,$$

where I denotes the input remote sensing observations, C is an optional condition, P represents the task-specific prompt, and O denotes the expected output.

Sensing Tasks

Cloud Removal (CR). Given a cloud-contaminated remote sensing image and a task prompt, the model generates the corresponding cloud-free image:

$$(I_{cloud}, P) \rightarrow I_{clear}.$$

Time Series Reconstruction (TSR). Given incomplete multi-temporal remote sensing observations, together with a reconstruction prompt, the model predicts the target temporal scene:

$$(\{I_{t_1}, I_{t_2}, ..., I_{t_n}\}, P) \rightarrow I_{t_k}.$$

Understanding Tasks

Semantic Segmentation (SS). Given a remote sensing image and a segmentation prompt, the model generates a pixel-level semantic mask of urban land-cover categories:

$$(I, P) \rightarrow M_{sem}.$$

For Binary Semantic Segmentation (BSS), the semantic mask is simplified into a foreground-background representation:

$$(I, P) \rightarrow M_{binary}.$$

Change Detection (CD). Given multi-temporal remote sensing observations over a specified time period and a prompt describing the target change type, the model detects regions that experienced the specified change:

$$(\{I_{t_1}, I_{t_2}, ..., I_{t_n}\}, C_{type}, P) \rightarrow M_{change}.$$

Reasoning Tasks

Dense Temporal Captioning (DTC). Given multi-temporal remote sensing observations and a temporal reasoning prompt, the model describes urban evolution processes:

$$(\{I_{t_1}, I_{t_2}, ..., I_{t_n}\}, P) \rightarrow T_{evolution}.$$

Future Scene Forecasting (FSF). Given historical observations, an optional planning condition, and a forecasting prompt, the model generates the projected future urban scene:

$$(\{I_{t_1}, I_{t_2}, ..., I_{t_n}\}, C_{plan}, P) \rightarrow I_{future}.$$

Decision Tasks

Risk Assessment (RA). Given urban remote sensing observations and a risk assessment prompt, the model produces risk identification and reasoning results:

$$(I, P) \rightarrow (R, T_{reasoning}),$$

where R denotes the identified risk category and $T_{reasoning}$ denotes the generated reasoning explanation.

Proposal Recommendation (PR). Given urban scenes and a planning prompt, the model generates spatially grounded development recommendations:

$$(I, P) \rightarrow T_{recommendation}.$$

B.3 Reference Construction for Language Tasks

This section describes the construction of reference supervision for the three language-oriented tasks: Dense Temporal Captioning (DTC), Risk Assessment (RA), and Proposal Recommendation (PR).

DTC. We reuse the reference temporal captions and associated annotations from DVL-Bench (Xuan et al., 2025). These references cover major change events, their temporal order, and their spatial locations.

RA. RA uses the original risk labels from the corresponding public datasets as reference risk types. Specifically, building-damage labels are obtained from DESTRUCTION (Ali et al., 2020) and xBD (Gupta et al., 2019), while urban-village labels are obtained from UV-SAM (Zhang et al., 2024). These labels serve as the reference basis for risk-type identification.

PR. For PR, we use an existing detailed description of the final-timestamp image and prompt GPT-5-mini to rewrite it into a planning-oriented reference proposal. Each proposal specifies a target spatial area or functional zone, a recommended planning action, and an expected development outcome. The final-timestamp image is used only to construct the reference proposal and is not provided to the evaluated models. The generated proposals are manually verified against the source descriptions to remove unsupported, inconsistent, or ambiguous recommendations.

B.4 Task Formulation and Prompt Design

We provide all prompts used in SURD-Bench to ensure reproducibility. Task-specific prompt templates are designed for different task categories, with explicit instructions and output constraints to standardize model responses. Each prompt is generated by integrating the task input into the corresponding template.

CR Instruction Template

Your task is to remove cloud contamination from a remote sensing image.
Cloud type: <cls_description>
Input image: <image>
Generate a cloud-free remote sensing image.

TSR Instruction Template

Your task is to reconstruct a missing observation from a chronological series of remote sensing images.
Given a sequence of remote sensing images with timestamps:
<timestamp_1>: <image>
<timestamp_2>: <image>
...
<timestamp_n>: <image>
Target timestamp: <target_time>
Generate the remote sensing image corresponding to the target timestamp.

SS Instruction Template

SEMANTIC SEGMENTATION

Your task is to segment the input remote sensing image into semantic categories.
Input image: <image>
Generate a color PNG segmentation mask.
Use only the following fixed RGB colors:
black=(0,0,0): nodata
white=(255,255,255): background
red=(255,0,0): building
gray=(128,128,128): road
blue=(0,0,255): water
orange=(255,128,0): barren
green=(0,200,0): forest
yellow=(255,255,0): agriculture
Each semantic category must use only its assigned RGB color.

BINARY SEMANTIC SEGMENTATION

Your task is to segment the input remote sensing image for the specified category.
Input image: <image>
Target category: <cls_description>
Generate a binary segmentation mask.
Use: black (0): other areas white (255): <cls_description>

CD Instruction Template

Your task is to identify specific land-cover transitions from a chronological series of remote sensing images.
Given a sequence of remote sensing images with timestamps:
<timestamp_1>: <image>
<timestamp_2>: <image>
...
<timestamp_n>: <image>
Detect the regions that experienced the following change during <start_year>-<end_year>: <change_description>
Generate a binary change map image:
black (0) for unchanged areas, white (255) for changed areas.

DTC 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 describe the temporal evolution and major changes of this area over the entire observation period.
Given a sequence of historical remote sensing images with timestamps:
<timestamp_1>: <image>
<timestamp_2>: <image>
...
<timestamp_n>: <image>
Generate a description that follows these requirements:
1. Start with a general summary statement describing the overall evolution trend.
2. Describe specific land feature changes chronologically across different timestamps.
3. Include clear spatial references when describing changes (e.g., top-left, center-right).
4. Conclude with an overall summary of the final landscape state.

DTC Evaluation Prompt

You are an evaluator specialized in dense temporal captioning for remote sensing image time series.
Compare the predicted caption with the ground-truth caption and evaluate the prediction from the following three aspects:
1. Temporal Coverage: Evaluate whether the caption captures important temporal milestones, periods of change, and the complete evolution sequence.
2. Spatial Accuracy: Evaluate whether the caption correctly identifies changed regions and spatial relationships.
3. Process Fidelity: Evaluate whether the caption accurately describes the initial state, intermediate processes, and final state of land-cover changes.
Assign a score from 0 to 5 for each dimension, where 5 indicates perfect agreement and 0 indicates poor performance.
Ground Truth Caption: <ground_truth_caption>
Predicted Caption: <predicted_caption>
Return only a valid JSON object in the following format:
{
"Temporal Coverage": { "score": <0-5>, "explanation": "brief explanation" },
"Spatial Accuracy": { "score": <0-5>, "explanation": "brief explanation" },
"Process Fidelity": { "score": <0-5>, "explanation": "brief explanation" }
}

FSF Instruction Template

Your task is to forecast the future remote sensing scene based on historical observations and future urban planning information.
Given a sequence of historical remote sensing images with timestamps:
<timestamp_1>: <image>
<timestamp_2>: <image>
...
<timestamp_n>: <image>
Future urban planning description: <planning>
Forecast the remote sensing scene at the target timestamp: <target_time>
Generate a realistic future RGB remote sensing image with spatially coherent structures and plausible land-cover transitions.

RA Instruction Template

You are a remote sensing image risk assessment expert.
Your task is to identify potential urban risks or abnormal phenomena from the input remote sensing image.
Input image: <image>
Analyze the scene based on visible visual evidence and provide a risk assessment.
Your response should:
1. Start with the most likely risk type.
2. Describe the main scene characteristics related to the risk.
3. Explain the visual evidence supporting the judgment.
4. Assess the severity of the potential risk.
Provide the assessment in one concise paragraph.

RA Evaluation Prompt

You are an impartial evaluator for remote sensing risk assessment.
Given a ground-truth risk label and a candidate model response, evaluate whether the model correctly identifies the risk type.
Ground Truth Risk Type: <ground_truth>
Candidate Response: <predicted_answer>
Evaluation criteria:
1. Risk type consistency: Determine whether the predicted risk is semantically consistent with the ground-truth label.
2. Evidence quality: Evaluate whether the explanation is supported by relevant visual evidence.
3. Assessment quality: Evaluate whether the response provides a clear and professional risk assessment.
Assign a score from 0 to 5, where 5 indicates a correct and well-supported assessment.
Return only a valid JSON object:
{
"is_correct": true or false,
"score": 0-5 integer,
"explanation": "brief explanation"
}

PR Instruction Template

You are an expert in urban planning and remote sensing interpretation.
Your task is to analyze multi-temporal remote sensing images of the same urban area and propose reasonable urban development or renewal recommendations.
Given a sequence of remote sensing images with timestamps:
<timestamp_1>: <image>
<timestamp_2>: <image>
...
<timestamp_n>: <image>
Generate planning recommendations based on observed spatial evolution patterns.
Your recommendations should be grounded in visible spatial evidence, including:
- land-use changes,
- transportation development,
- building expansion,
- open spaces,
- industrial areas,
- and environmental structures.
Each recommendation should include:
1. A target spatial area or functional zone;
2. A planning or redevelopment action;
3. An expected urban development outcome or functional benefit.
Provide 2–4 distinct spatial interventions in a concise and professional paragraph.

PR Evaluation Prompt

You are an impartial evaluator for urban planning proposal generation.
You will be given:
1. A ground-truth urban planning proposal.
2. A candidate model-generated proposal.
Evaluate the candidate proposal from three aspects:
1. Spatial Grounding Accuracy: Evaluate whether the proposal correctly identifies the target spatial regions, functional zones, or geographic areas relevant to the reference proposal.
2. Planning Action Fidelity: Evaluate whether the proposed planning or redevelopment actions are consistent with the reference planning objectives.
3. Development Outcome Relevance: Evaluate whether the expected outcomes, functions, or benefits align with the reference proposal.
Ground-truth proposal: <ground_truth>
Candidate proposal: <predicted_proposal>
Assign a score from 0 to 5 for each dimension, where 5 indicates excellent agreement.
Return only a valid JSON object:
{
"Spatial Grounding Accuracy": { "score": <0-5>, "explanation": "brief explanation" },
"Planning Action Fidelity": { "score": <0-5>, "explanation": "brief explanation" },
"Development Outcome Relevance": { "score": <0-5>, "explanation": "brief explanation" }
}

C. Evaluation Protocol Details

C.1 Evaluation Metrics

To evaluate the spatiotemporal cognitive capabilities of UMMs, SURD-Bench adopts task-specific metrics and scoring functions for the eight tasks. Individual metrics are aligned and aggregated according to task characteristics, and the resulting Task Scores (TS) are normalized to a unified range of 0–100 for fair comparison.

For a task with multiple complementary metrics, the Task Score is generally computed as:

$$\mathrm{TS}_i = 100 \times f_i(m_{i1}, m_{i2}, ..., m_{iM_i}), \tag{1}$$

where M_i denotes the number of evaluation metrics in the i-th task, m_{ij} represents the j-th metric after aligning the optimization direction, and $f_i(\cdot)$ denotes the task-specific aggregation function.

Detailed scoring functions for each task are described in Sections C.1.1–C.1.4. Finally, the overall benchmark performance is measured by Avg. TS, calculated as the average of the Task Scores across the eight tasks:

$$\text{Avg. TS} = \frac{1}{8}\sum_{i=1}^{8} \mathrm{TS}_i. \tag{2}$$

C.1.1 Sensing Tasks

For the Cloud Removal (CR) task, we evaluate image reconstruction quality using three complementary metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). PSNR and SSIM measure pixel-level fidelity and structural consistency, while LPIPS captures perceptual similarity. After aligning the optimization direction and normalizing each metric, the Task Score is computed as:

$$\mathrm{TS}_{\mathrm{CR}} = 100 \times \frac{1}{3}\left(\frac{\mathrm{PSNR} - 0}{40 - 0} + \mathrm{SSIM} + (1 - \mathrm{LPIPS})\right). \tag{3}$$

For the Time Series Reconstruction (TSR) task, we use LPIPS to measure perceptual differences and CLIP-I to evaluate image-level semantic similarity between reconstructed and reference images. The Task Score is calculated as:

$$\mathrm{TS}_{\mathrm{TSR}} = 100 \times \frac{1}{2}\left((1 - \mathrm{LPIPS}) + \text{CLIP-I}\right). \tag{4}$$

C.1.2 UNDERSTANDING TASKS

For the Semantic Segmentation (SS) task, we introduce two evaluation settings: multi-class semantic segmentation and binary semantic segmentation (BSS). The SS setting evaluates the model's overall semantic parsing capability across six urban land-cover categories, including buildings, roads, water bodies, bare land, forest, and farmland. In contrast, BSS simplifies the task into a foreground-background segmentation problem to assess the model's ability to identify target regions. Intersection over Union (IoU) and mean Intersection over Union (mIoU) are adopted as evaluation metrics. The Task Scores are calculated as:

$$\mathrm{TS}_{\mathrm{SS}} = 100 \times \mathrm{mIoU}, \tag{5}$$

$$\mathrm{TS}_{\mathrm{BSS}} = 100 \times \mathrm{IoU}. \tag{6}$$

For the Change Detection (CD) task, we employ Intersection over Union (IoU) to evaluate the model's ability to identify changed regions from long-term multi-temporal observations. The Task Score is calculated as:

$$\mathrm{TS}_{\mathrm{CD}} = 100 \times \mathrm{IoU}. \tag{7}$$

C.1.3 REASONING TASKS

For the Dense Temporal Captioning (DTC) task, we evaluate the generated captions from three dimensions: Temporal Coverage, Spatial Accuracy, and Process Fidelity. Temporal Coverage measures whether the generated captions capture key temporal changes; Spatial Accuracy evaluates the correctness of spatial references; and Process Fidelity assesses the consistency between the described evolutionary process and the observed urban dynamics. Since each dimension is scored on a five-point scale, the Task Score is calculated as:

$$\mathrm{TS}_{\mathrm{DTC}} = 100 \times \frac{S_{\mathrm{TC}} + S_{\mathrm{SA}} + S_{\mathrm{PF}}}{3 \times 5}, \tag{8}$$

where S_{TC}, S_{SA}, and S_{PF} denote the scores of the three evaluation dimensions, respectively.

For the Future Scene Forecasting (FSF) task, we employ perceptual and semantic metrics to evaluate the generated future scenes. LPIPS measures the perceptual similarity between the predicted and reference scenes, while CLIP-I evaluates their semantic consistency. The Task Score is calculated as:

$$\mathrm{TS}_{\mathrm{FSF}} = 100 \times \frac{(1 - \mathrm{LPIPS}) + \mathrm{CLIP\text{-}I}}{2}. \tag{9}$$

C.1.4 DECISION TASKS

For the Risk Assessment (RA) task, we evaluate model performance from two aspects: Identification Accuracy and Assessment Score. Identification Accuracy (S_{IA}) indicates whether the model correctly identifies the risk category, while Assessment Score (S_{AS}) measures the reliability of the generated reasoning process on a five-point scale. To emphasize that reliable assessment requires both correct identification and trustworthy reasoning, the Task Score is calculated as:

$$\mathrm{TS}_{\mathrm{RA}} = 100 \times S_{\mathrm{IA}} \times \frac{S_{\mathrm{AS}}}{5}, \tag{10}$$

where S_{IA} is the identification accuracy and S_{AS} denotes the reasoning reliability score.

For the Proposal Recommendation (PR) task, we evaluate the generated recommendations from three dimensions: Spatial Grounding Accuracy, Planning Action Fidelity, and Development Outcome Relevance. Each dimension is scored on a five-point scale, and the Task Score is calculated as:

$$\mathrm{TS}_{\mathrm{PR}} = 100 \times \frac{S_{\mathrm{SGA}} + S_{\mathrm{PAF}} + S_{\mathrm{DOR}}}{3 \times 5}. \tag{11}$$

C.2 MODEL EVALUATION SETTINGS

We evaluate a diverse set of multimodal models, including understanding-oriented models, image generation models, and unified multimodal models. The evaluated models cover both proprietary and open-source models,

with different architectures and language model backbones. As summarized in Table 4, these models are selected to represent the current capabilities of multimodal systems across different categories.

All models are evaluated under a unified zero-shot setting without additional fine-tuning or task-specific optimization. For each task, models receive the same input data and task-specific prompts described in Section B.4, and their outputs are evaluated using the corresponding metrics defined in Section C.1. For models supporting image generation or multimodal reasoning, we directly use their official inference interfaces or publicly released checkpoints.

Table 4: Details of evaluated multimodal models.

Category	Model	LLM Backbone	Open-source
Understanding	TEOChat (Irvin et al., 2025)	LLaMA2-7B	✓
	VHM (Pang et al., 2025)	Vicuna-7B	✓
	LHRS-Bot-nova (Li et al., 2025b)	LLaMA3-8B	✓
	ScoreRS w/ RL (Muhtar et al., 2025)	Qwen2-7B	✓
	InternVL3.5-8B (Wang et al., 2025b)	Qwen3-8B	✓
	Qwen3-VL-8B-Instruct (Bai et al., 2025)	Qwen3-8B	✓
	GPT-4o (Hurst et al., 2024)	–	✗
Generation	Qwen-Image-2.0-Pro (Zhao et al., 2026)	–	✗
	FLUX.2 (Black Forest Labs, 2025)	–	✓
	GPT-Image-2 (OpenAI, 2026)	–	✗
Unified	Janus-Pro (Chen et al., 2025)	DeepSeek-LLM-7B-base	✓
	Harmon-RecA-plus (Xie et al., 2025)	Qwen2.5-1.5B	✓
	EMU3.5 (Cui et al., 2025)	–	✓
	Bagel-RecA (Xie et al., 2025)	Qwen2.5-7B	✓
	OmniGen2 (Wu et al., 2026)	Qwen2.5-3B	✓
	SenseNova-U1-8B-MoT (Diao et al., 2026)	Qwen3-8B	✓
	BAGEL (Deng et al., 2025)	Qwen2.5-7B	✓
	Gemini 3.1-image (Google DeepMind, 2026)	–	✗

D. Additional Experimental Results and Analysis

To provide a more comprehensive analysis of model performance, we report detailed evaluation results for each task in SURD-Bench. While the main paper presents the overall performance based on Avg. TS, this section provides the original task-level metrics before aggregation, enabling a fine-grained comparison across different cognitive abilities.

Tables 5–12 summarize the detailed results of the eight tasks.

Table 5: Detailed results of cloud removal.

Model	LPIPS↓	PSNR↑	SSIM↑
Qwen-Image-2.0-Pro	0.6061	12.2632	0.2482
FLUX.2	0.6674	13.2788	0.3299
GPT-Image-2	0.6229	9.3968	0.2281
Harmon-RecA-plus	0.6696	15.1455	0.3584
EMU3.5	0.6789	9.9489	0.2272
Bagel-RecA	0.6363	11.4936	0.3141
OmniGen2	0.5379	15.0104	0.4486
SenseNova-U1-8B-MoT	0.5651	11.8523	0.3747
BAGEL	0.7172	10.6813	0.2850
Gemini-3.1-image	0.5967	9.8627	0.2006

Table 6: Detailed results of Time Series Reconstruction.

Model	LPIPS↓	CLIP-I↑
FLUX.2	0.5808	0.9075
GPT-Image-2	0.5764	0.9001
Harmon-RecA-plus	0.6783	0.5540
EMU3.5	0.5858	0.8854
Bagel-RecA	0.7385	0.7453
OmniGen2	0.6742	0.7585
SenseNova-U1-8B-MoT	0.6882	0.8408
BAGEL	0.7363	0.7181
Gemini-3.1-image	0.5571	0.9101

Table 7: Detailed results of Semantic Segmentation.

Model	IoU-Binary	mIoU
Qwen-Image-2.0-Pro	0.2312	0.1479
FLUX.2	0.0935	0.0417
GPT-Image-2	0.3308	0.2408
Harmon-RecA-plus	0.0345	0.0197
EMU3.5	0.1114	0.0130
Bagel-RecA	0.1426	0.0814
OmniGen2	0.0889	0.0529
SenseNova-U1-8B-MoT	0.1366	0.0248
BAGEL	0.1317	0.0897
Gemini-3.1-image	0.4275	0.3189

Table 8: Detailed results of Change Detection.

Model	IoU	F1	Precision	Recall	OA
FLUX.2	0.0509	0.0876	0.0578	0.5589	0.5165
GPT-Image-2	0.0637	0.1094	0.0878	0.3125	0.8422
Harmon-RecA-plus	0.0345	0.0622	0.0500	0.2616	0.7054
EMU3.5	0.0509	0.0873	0.0582	0.5471	0.5238
Bagel-RecA	0.0474	0.0804	0.0515	0.5280	0.4624
OmniGen2	0.0557	0.0957	0.0607	0.6520	0.4594
SenseNova-U1-8B-MoT	0.0512	0.0891	0.0585	0.5358	0.5159
BAGEL	0.0358	0.0623	0.0420	0.4035	0.5663
Gemini-3.1-image	0.0900	0.1342	0.1343	0.2101	0.9132

Table 9: Detailed results of Dense Temporal Captioning.

Model	Temporal Coverage	Spatial Accuracy	Process Fidelity
TEOChat	0.6951	0.2459	0.5650
VHM	1.3008	0.6626	0.7561
LHRS-Bot-nova	0.8882	0.7256	0.9004
ScoreRS w/ RL	0.8923	0.9207	0.8760
InternVL3.5-8B	2.1972	1.3252	1.3598
Qwen3-VL-8B-Instruct	2.3150	1.4756	1.6748
GPT-4o	2.7744	1.8435	1.9004
Janus-Pro	0.7236	0.4593	0.4553
Harmon-RecA-plus	1.6565	0.9085	0.8598
EMU3.5	2.0874	1.1077	1.2764
Bagel-RecA	1.8435	1.1809	1.2317
OmniGen2	1.5528	1.0528	1.1077
SenseNova-U1-8B-MoT	2.0569	1.3720	1.3191
BAGEL	1.7195	1.1199	1.1870
Gemini-3.1-image	2.1809	1.5711	1.5874

Table 10: Detailed results of Future Scene Forecasting.

Model	CLIP-I↑	LPIPS↓
FLUX.2	0.9420	0.5076
GPT-Image-2	0.9296	0.5075
Harmon-RecA-plus	0.5743	0.6852
SenseNova-U1-8B-MoT	0.8460	0.6962
OmniGen2	0.8904	0.5968
EMU3.5	0.9323	0.5391
Bagel-RecA	0.7902	0.7325
BAGEL	0.7618	0.7556
Gemini-3.1-image	0.9471	0.4685

Table 11: Detailed results of Risk Assessment.

Model	Identification Accuracy	Assessment Score
TEOChat	0.0951	1.2465
VHM	0.2324	1.6725
LHRS-Bot-nova	0.3204	1.9437
ScoreRS w/ RL	0.4049	2.3380
InternVL3.5-8B	0.5317	3.0141
Qwen3-VL-8B-Instruct	0.6021	3.4577
GPT-4o	0.7817	4.2465
Janus-Pro	0.6901	3.5106
Harmon-RecA-plus	0	0.8239
EMU3.5	0.2606	2.1232
Bagel-RecA	0.5282	3.1092
OmniGen2	0.4366	2.5246
SenseNova-U1-8B-MoT	0.4718	2.8415
BAGEL	0.5282	3.1021
Gemini-3.1-image	0.4225	3.0317

Table 12: Detailed results of Proposal Recommendation.

Model	Spatial Grounding Accuracy	Planning Action Fidelity	Development Outcome Relevance
TEOChat	0.6600	0.9620	1.7480
VHM	1.1420	1.2080	1.9160
LHRS-Bot-nova	1.0520	1.3540	2.2340
ScoreRS w/ RL	1.0980	1.2760	2.0780
InternVL3.5-8B	1.4840	1.4640	2.4020
Qwen3-VL-8B-Instruct	2.1160	1.7380	2.7240
GPT-4o	1.9320	1.6340	2.6800
Janus-Pro	1.2360	1.4100	2.1840
Harmon-RecA-plus	0.9420	1.3480	2.1520
EMU3.5	1.2880	1.4520	2.3320
Bagel-RecA	1.3380	1.5280	2.4140
OmniGen2	1.0700	1.2680	2.1300
SenseNova-U1-8B-MoT	1.8800	1.5740	2.3820
BAGEL	1.3640	1.5100	2.4340
Gemini-3.1-image	2.2520	1.8480	2.8480

E. Visualization and Case Studies

To provide qualitative insights into the behaviors of UMMs on SURD-Bench, we present representative examples from all eight tasks. These cases cover different output formats, including generated images, segmentation masks, and textual responses.

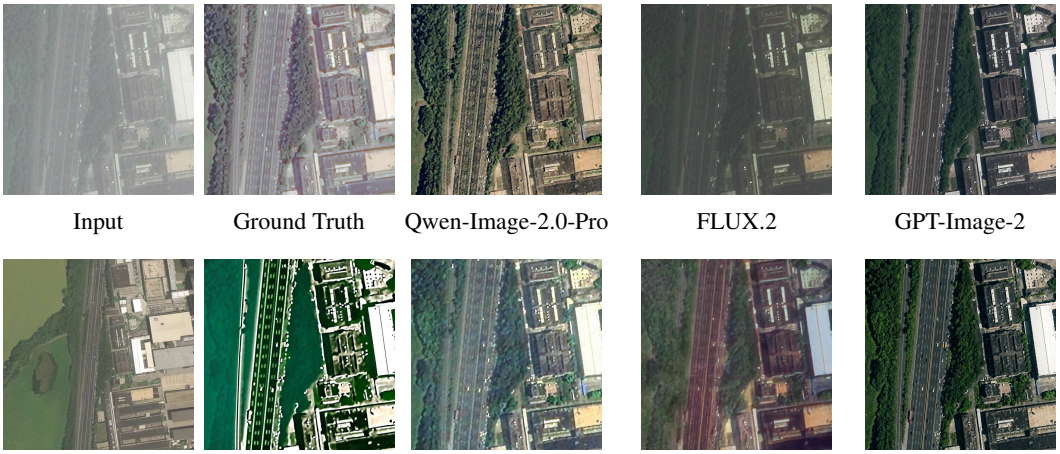


Figure 6: Qualitative results of the Cloud Removal task.

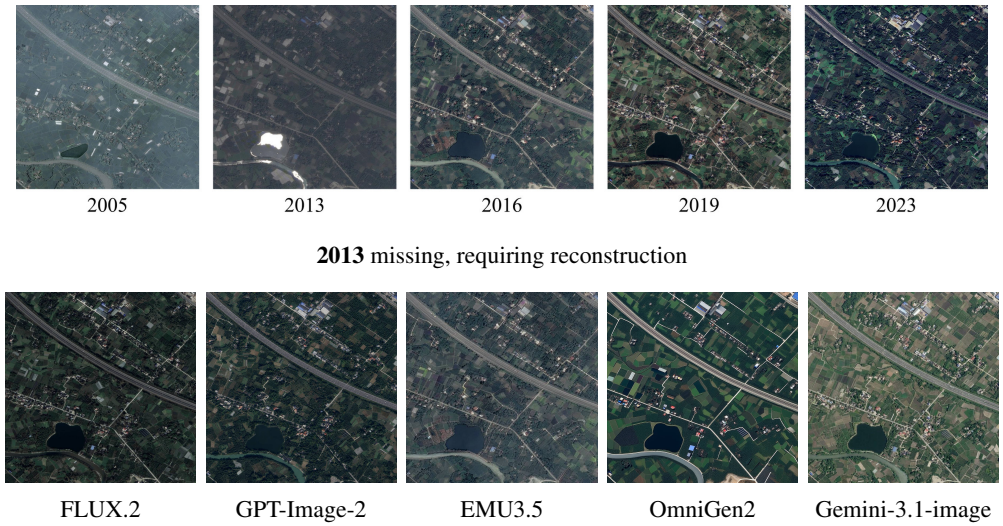


Figure 7: Qualitative results of the Time Series Reconstruction task. The image from 2013 is missing and requires reconstruction.

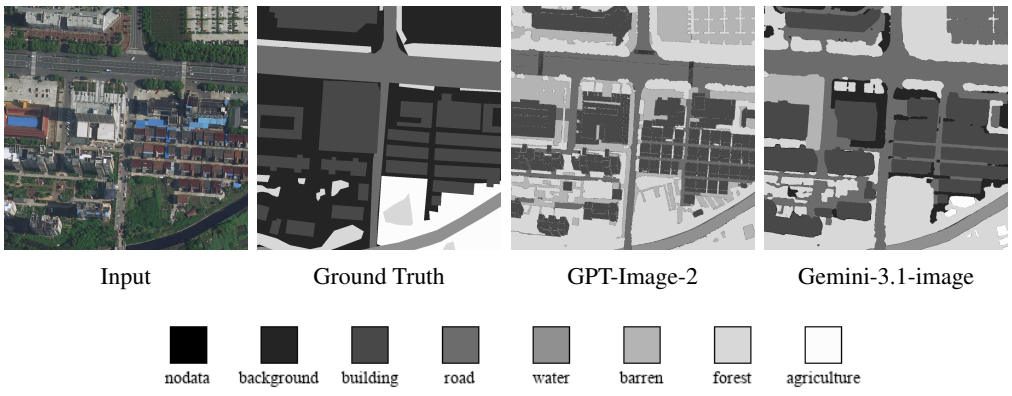


Figure 8: Qualitative results of the Semantic Segmentation task.

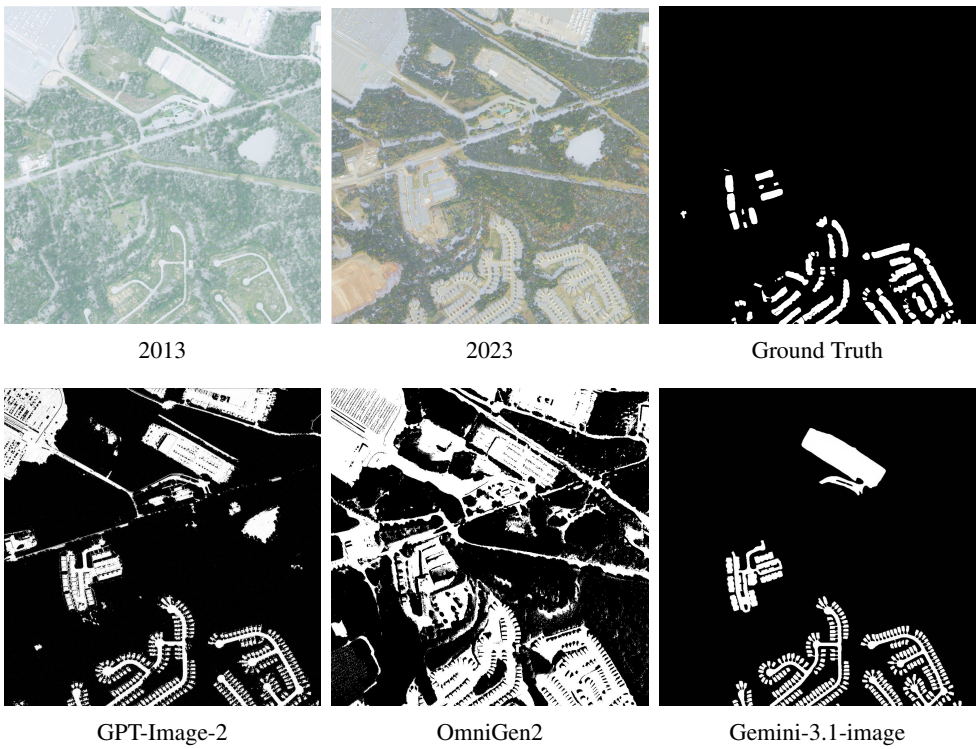


Figure 9: Qualitative results of the Change Detection task.



Ground_Truth: Over the 12-year period (2010 – 2022) the scene evolved from predominantly agricultural land to a mixed residential-commercial landscape. By 2012, the housing subdivision at the top had infilled with many new dwellings, and fresh earth disturbance emerged at the centre-right. In 2014, the subdivision expanded further across the top, while a wide, newly cleared tract dominated the centre and a large bare excavation appeared along the right. Earth-moving persisted in 2018, evidenced by freshly graded curvilinear roads threading through the centre and connecting to the prominent road at the bottom. During 2020, several large white-roofed buildings with parking areas were erected in a row along the bottom-centre, converting previously cleared ground to impervious surfaces. By 2022, similar structures extended this commercial strip toward the left, and adjacent agricultural parcels beneath the prominent road displayed lighter tones consistent with diminished cultivation. Throughout the sequence, the watercourse and its vegetated corridor at the right remained largely unchanged.
VHM: The area captured in these images exhibits a gradual transformation over the years. In 2010, the area is predominantly rural, characterized by open fields and scattered trees. By 2012, the area has undergone significant development, with the construction of new residential neighborhoods and the expansion of existing ones. The presence of a bridge in the center-right (×) area suggests the development of transportation infrastructure. In 2014, the area continues to grow, with the addition of new roads and the expansion of existing ones. By 2018, the area has become a suburban region, with a mix of residential and commercial development (√). The presence of a park in the center-left area indicates the importance of recreational spaces. In 2020, the area has further developed, with the construction of new schools and the expansion of existing ones (×). The presence of a sports field (×) in the center-right area suggests the importance of recreational facilities. In 2022, the area has reached a mature state, with a well-developed infrastructure and a mix of residential, commercial, and recreational spaces.
GPT-4o: Over the period from 2010 to 2022, the area underwent significant urban development, transitioning from primarily agricultural land to a suburban residential and commercial zone.\n\n1. **2010 (20100616)**: The area is predominantly agricultural, with fields occupying most of the space. A small residential neighborhood is visible in the top-center (√), with a few houses and roads already established. The rest of the area remains undeveloped.\n\n2. **2012 (20120626)**: The residential neighborhood in the top-center has expanded slightly, with more houses and roads added. The surrounding fields remain largely intact, though some land clearing is evident in the center-right, indicating early stages of development.\n\n3. **2014 (20140916)**: Significant land clearing is visible in the center and center-right areas, with large patches of bare soil replacing agricultural fields (√). The residential neighborhood in the top-center continues to grow, with additional houses and roads constructed.\n\n4. **2018 (20180711)**: The cleared areas in the center and center-right have been further developed, with new roads and infrastructure visible. A new road network extends southward, connecting to the main highway at the bottom. The residential neighborhood in the top-center is now well-established, with dense housing.\n\n5. **2020 (20200815)**: Commercial development becomes evident in the center-bottom area, with several large buildings constructed along the new road network. The residential neighborhood in the top-center remains stable, while the surrounding areas continue to transition from agricultural to developed land.\n\n6. **2022 (20220802)**: The commercial zone in the center-bottom has expanded, with additional buildings and infrastructure completed (√). The residential neighborhood in the top-center remains fully developed, and the surrounding areas show minimal remaining agricultural land.\n\nIn summary, the area has transformed from a primarily agricultural landscape in 2010 to a suburban residential and commercial hub by 2022, with significant development concentrated in the center and bottom portions of the region.

Figure 10: Qualitative results of the Dense Temporal Captioning task.

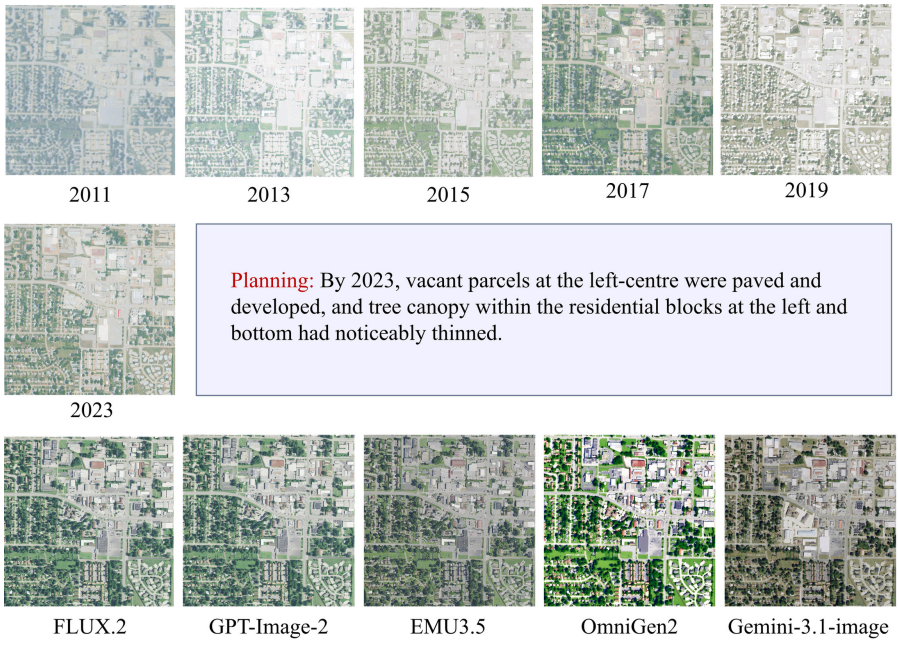
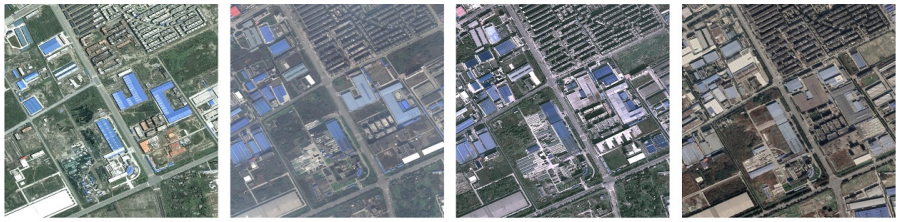


Figure 11: Qualitative results of the Future Scene Forecasting task.



Ground_Truth: Urban Villages.
Cls_Description: Urban Villages or Building Damage.
TEOChat: Based on the input remote sensing image, the most likely risk type in the urban area is the presence of a tsunami. The main scene characteristics include a coastal area with multiple buildings, harbors, and a golf field. Potential risks arise from the proximity of these structures to the coast and the possibility of a tsunami impact. The judgment basis is the observed visual evidence of the coastal area and the potential for tsunami activity in the region. The risk severity is high due to the large number of structures and the potential for significant damage in the event of a tsunami.
Score: 1
Explanation: The candidate predicts a tsunami risk (coastal hazard), while the ground-truth label is 'Urban Villages' (a built-environment / socio-urban risk). This is a mismatched risk type; although the model notes buildings and coast, it fails to identify the correct urban-village risk. Evidence and reasoning do not support the ground-truth label.
Janus-Pro: The most likely risk type in this urban area is infrastructure-related risks, such as inadequate drainage systems or insufficient road networks. The main scene characteristics include densely packed buildings, irregular spatial structures, and visible water bodies, which may indicate potential flooding or waterlogging issues. The judgment basis is the observed density and layout of the buildings, as well as the proximity of water bodies to the urban area. The risk severity is moderate, as the infrastructure appears to be functional, but the lack of clear separation between residential and industrial zones may pose long-term risks to water management and flood resilience.
Score: 4
Explanation: The response describes densely packed, irregular building layouts and related infrastructure problems, which are characteristic of urban villages. The focus on drainage/road issues and flood risk is a reasonable interpretation, but it omits some common urban-village concerns (e.g., informal construction, sanitation, fire hazards), so the match is good but not complete.

Figure 12: Qualitative results of the Risk Assessment task.



Ground_Truth: In the central arterial corridor, demolish fragmented sheds and construct larger consolidated multi-bay factory blocks to create continuous industrial frontage and larger, more efficient production floorplates; in the eastern roadside zone, upgrade peripheral vacant plots and add modular logistics parks with covered warehousing and dedicated loading yards to increase storage capacity and improve freight access; along the northern residential strip, connect housing clusters to nearby workplaces by construct pedestrian-priority streets and add formal bus stops to produce direct safe access and reduced travel distances; on local vacant land and inner-courtyard green patches, open up and convert low-lying plots into linear stormwater-retention parks and add tree-lined greenways to deliver flood mitigation, passive recreation, and visual separation between industrial and residential areas.

TEOChat: Based on the provided sequence of remote sensing images, it appears that there has been no significant change in the urban area's overall structure. However, there have been some minor developments in transportation infrastructure.\n\nTo improve the urban planning and development of this area, I recommend the following interventions:\n\n1. Expand the existing road network to improve connectivity and accessibility within the city.\n2. Develop a new public transportation hub, such as a train station or bus terminal, to enhance mobility and reduce traffic congestion.\n3. Implement green spaces and parks in strategic locations to improve air quality, provide recreational areas, and promote a healthier lifestyle for the residents.\n4. Encourage sustainable urban development by promoting mixed-use zoning, which combines residential, commercial, and industrial areas to create vibrant and livable communities.\n\nThese interventions will help to create a more balanced and sustainable urban environment, improving the overall quality of life for the city's inhabitants.
Score: Spatial Grounding Accuracy: 1, Planning Action Fidelity: 1, Development Outcome Relevance: 2

GPT-4o: The sequence of images reveals significant spatial evolution, including industrial expansion, residential densification, and the reduction of open spaces. First, the industrial zone in the southern area should undergo modernization by integrating green technologies and optimizing land use to enhance environmental sustainability and economic productivity. Second, the residential area in the northern section, which has become increasingly dense, should be redeveloped with mixed-use buildings and improved public amenities to support livability and reduce urban congestion. Third, the central open spaces, which have diminished over time, should be revitalized into urban parks and recreational zones to improve air quality and provide community spaces. Lastly, the transportation corridors should be upgraded with multimodal transit systems to improve connectivity and reduce traffic congestion, fostering efficient urban mobility.
Score: Spatial Grounding Accuracy: 2, Planning Action Fidelity: 2, Development Outcome Relevance: 3

Figure 13: Qualitative results of the Proposal Recommendation task.