RS-WORLDMODEL: A UNIFIED BENCHMARK FOR FINE-GRAINED REMOTE SENSING WITH A STRONG BASELINE

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote sensing world modeling requires models that both explain observed Earth-surface changes and generate future scenes under specified conditions at fine granularity, yet existing methods typically decouple interpretation from forecasting and operate at a coarser granularity. To bridge this gap, we construct RSWBench-1.1M, a unified **1.1 million**-sample benchmark whose fine-grained language annotations tie each described change to rich geographic and acquisition metadata, for spatiotemporal change question answering and text-guided future scene forecasting. As a strong baseline, we present RS-WorldModel, a unified autoregressive world model that combines geo-aware generative pre-training, shared instruction tuning, and reference-based verifiable reward alignment to unify understanding and forecasting in *a single model*. With only 2B parameters, RS-WorldModel surpasses open-source models up to 120× larger on most spatiotemporal change question-answering metrics and, on text-guided future scene forecasting, achieves the best PSNR of 17.20 and the best FID of 43.13 among all compared methods, including the closed-source Gemini-2.5-Flash Image.

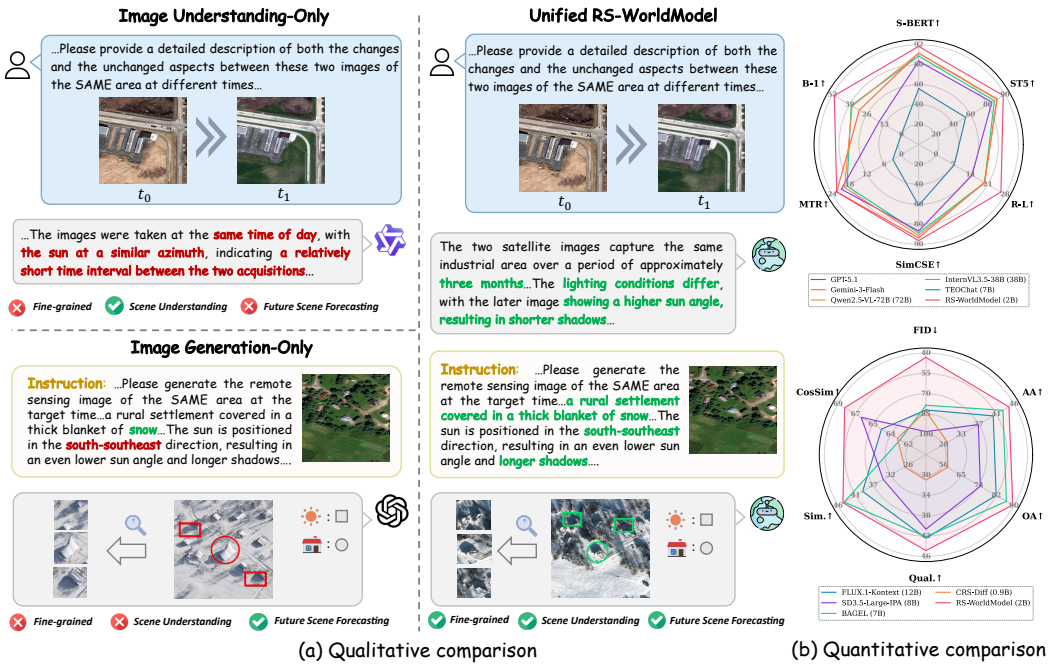
1 INTRODUCTION

Cross-modal remote sensing image understanding aims to interpret observed satellite imagery and produce textual descriptions of scene semantics and changes (Kuckreja et al., 2024; Hu et al., 2025; Zhan et al., 2025), while generative modeling supports satellite image synthesis and future forecasting, providing a complementary capability for scenario analysis and decision support (Khanna et al., 2023; Zhao et al., 2025; Guo et al., 2026). For example, GeoChat grounds natural-language interpretation in remote sensing scenes (Kuckreja et al., 2024), whereas DiffusionSat and TAMMs respectively study satellite image generation and time-series forecasting (Khanna et al., 2023; Guo et al., 2026).

Despite this progress, existing research faces two intertwined challenges. *First*, the field remains fragmented, with current methods treating cross-modal understanding and generation/forecasting as separate problems, as illustrated in Figure 1. This separation leads to disjointed workflows, where interpretation and prediction are optimized independently, without a shared representation of Earth-observation dynamics (Kuckreja et al., 2024; Zhan et al., 2025; Khanna et al., 2023; Zhao et al., 2025; Guo et al., 2026). In contrast, unified multimodal models have emerged as a promising direction for coupling perception and generation within general-purpose model interfaces (Wu et al., 2025; Chen et al., 2025b;a; Xie et al., 2025; Cui et al., 2025). *Second*, there is a lack of fine-grained modeling across both understanding and generation tasks. On the understanding side, models often rely on coarse descriptions, limiting their ability to capture detailed spatiotemporal changes and subtle semantic variations (Xuan et al., 2025; Wang et al., 2025b). On the generation side, existing approaches struggle to produce geographically and semantically precise outputs, particularly under detailed instructions or metadata (Khanna et al., 2023; Zhao et al., 2025; Guo et al., 2026).

To address these challenges, we first construct RSWBench-1.1M, a large-scale benchmark of **1.1M** samples drawn from high-resolution fMoW imagery (Christie et al., 2018), covering both spatiotemporal change understanding and text-guided future scene forecasting across the globe. Its annotations are fine-grained in the language sense: long-form descriptions (about 246 and 110 tokens for the two tasks) that tie each change to geographic and acquisition metadata. As a strong baseline on this

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) Qualitative comparison (b) Quantitative comparison

Figure 1: (a) We present RS-WorldModel, a unified remote sensing world model that jointly supports spatiotemporal change understanding and text-guided future scene forecasting in one model, where red marks inaccurate or hallucinated content from single-task baselines and green marks correctly captured changes and forecasts. (b) Compared with state-of-the-art models, RS-WorldModel performs better on both understanding (top) and forecasting (bottom) metrics on RSWBench-1.1M.

benchmark, we propose RS-WorldModel, a unified autoregressive remote sensing world model that combines geographic conditioning, joint understanding-generation training, and reference-based verifiable reward alignment in a single framework. We use *world model* in a restricted sense: predicting the future observation of a scene from its current observation and the target acquisition conditions, and describing observed change with the same weights, without actions, interventions, or multi-step rollouts. This design lets one model with one set of weights handle both tasks while keeping its outputs consistent with the metadata. With only **2B** parameters, RS-WorldModel surpasses open-source models up to 120× larger on most spatiotemporal change QA metrics and, on text-guided future scene forecasting, achieves the best PSNR of **17.20** and the best FID of **43.13** among all compared methods, including the closed-source Gemini-2.5-Flash Image. Our contributions are:

- We construct RSWBench-1.1M, a benchmark of 1.1M samples from high-resolution imagery with fine-grained, metadata-rich language annotations, covering spatiotemporal change understanding and text-guided future scene forecasting.
- We propose RS-WorldModel, a unified autoregressive world model that handles fine-grained change description and metadata-conditioned future scene forecasting with a single set of weights.
- Extensive experiments with both semantic and reference-based metrics demonstrate that a compact 2B-parameter model achieves strong performance on both change understanding and future forecasting tasks, showing that a single model can handle both tasks.

2 RSWBench-1.1M Dataset

Training a unified remote sensing world model requires data that simultaneously support Spatiotemporal Change Question-Answering (ST-CQA) and Text-Guided Future Scene Forecasting (TFSF). We contribute a scalable annotation pipeline together with a 1.1M training corpus and a 6.6K evaluation benchmark, both derived exclusively from the fMoW archive under its official train/test split. The split is applied before temporal pairing and language annotation, so no evaluation image, caption, or generation prompt enters any training stage. As an additional check, we compare every evaluation

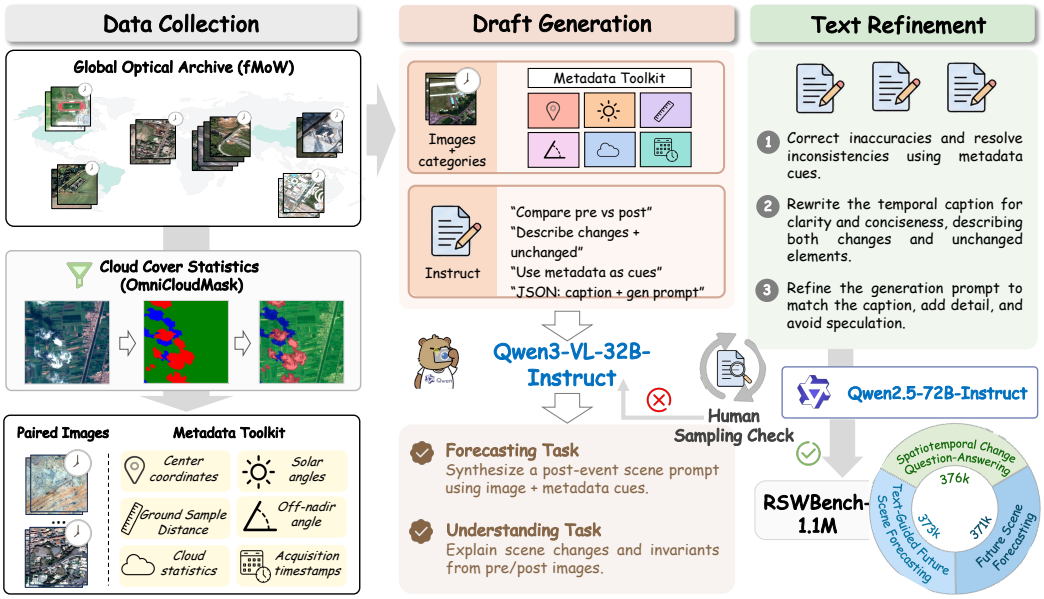


Figure 2: **Data construction pipeline and dataset composition.** A scalable pipeline turns multi-temporal observations into instruction data for understanding and forecasting, with training and evaluation samples drawn from the official fMoW train and test splits.

sample with the training corpus on its full metadata, namely geographic coordinates, ground sampling distance, both acquisition timestamps, sun azimuth and elevation, off-nadir angle, and cloud-cover ratio, and remove any sample that exactly matches a training sample.

Construction Pipeline. We build the data with a three-stage pipeline that unifies physical filtering, generate-and-refine semantic annotation, and a closed-loop human audit for quality control (Figure 2).

Physical Standardization. For each location, we pair multi-temporal observations and normalize their acquisition metadata (sun azimuth and elevation, off-nadir angle, timestamps) to disentangle land-cover changes from sensor-induced variations. Every sample is processed by OmniCloudMask (Wright et al., 2025) to estimate the cloud ratio $\rho_{\text{Cloud}}(I)$, retaining samples with $\rho_{\text{Cloud}}(I) \leq 0.9$. Unlike conventional datasets enforcing $\leq$5–10% clear-sky filters (Almar et al., 2025), we keep partially cloudy scenes so that cloud cover becomes a controllable condition.

Semantic Refinement. We adopt a generate-and-refine strategy for scalable language supervision. Qwen3-VL-32B-Instruct drafts structured JSON annotations from each image pair and its metadata, producing a temporal caption and a future-scene generation prompt. A larger LLM (Qwen2.5-72B-Instruct) then refines the drafts for factual consistency and metadata alignment, correcting VLM hallucinations. Critically, raw numeric values are translated into natural linguistic cues (e.g., low solar elevation as long shadows) to prevent overfitting to numerical tokens. Since the prompt is written with the target image in view, TFSF evaluates description-conditioned forecasting, in which the model must realize the described change while preserving the rest of the current scene.

Human Sampling Audit. All generated samples first pass an automatic JSON format check. To control semantic quality, a randomly sampled subset is then manually inspected. Drafts with correct content but malformed format are repaired in place; drafts with incorrect content are returned to the annotation step for regeneration, forming a closed feedback loop that progressively reduces annotation errors. In addition, two remote sensing experts audited 20,000 training records, with a mean pass rate of 90.03% (Appendix D).

Statistics and Comparison. The resulting RSWBench-1.1M contains 1.1M training samples and 6.6K evaluation samples with rich metadata and detailed language annotations. As shown in Table 1, to our knowledge RSWBench-1.1M is the first million-scale remote sensing dataset that pairs open-ended change descriptions with metadata-conditioned future scene targets. The closest work, TAMMs (Guo et al., 2026), couples change captioning with time-series forecasting but without explicit acquisition metadata or open-ended descriptions (Appendix B).

Table 1: Comparison of dataset capabilities and scales for remote sensing understanding and generation. ✓: supported; ✗: not supported; –: not applicable (dataset does not target this task category).

Dataset	Scale	Understanding		Generation		
		Temporal	Earth Observation	Spatiotemporal Metadata	Observation Environment	Fine-grained Text
TEOChatlas	554K	✓	✓	–	–	–
MMRS-1M	1.0M	✗	✓	–	–	–
CVACT-Text	88K	–	–	✗	✓	✓
RSWBench-1.1M	1.1M	✓	✓	✓	✓	✓

3 RS-WorldModel

We present RS-WorldModel, a unified autoregressive framework that handles both fine-grained *Spatiotemporal Change Question-Answering* (ST-CQA) and *Text-Guided Future Scene Forecasting* (TFSF) within a single 2B-parameter model (Figure 3). It rests on three choices: (i) shared metadata conditioning; (ii) a shared discrete output vocabulary of text and visual tokens under one next-token objective; and (iii) a three-stage pipeline that builds physical priors, tunes both tasks, and refines outputs with task-specific rewards.

3.1 Problem Formulation

Two Tasks, One Objective. Let I denote a remote sensing image and m its geospatial metadata (geographic coordinates, ground sampling distance, timestamp, sun azimuth and elevation, off-nadir angle, and cloud-cover ratio). We cast both ST-CQA and TFSF as instruction-conditioned sequence generation: given a prompt P containing images and metadata written as text, the model factorizes the conditional likelihood of the output sequence y as

$$p_\theta(y \mid P, I, m) \;=\; \prod_{i=1}^{|y|} p_\theta(y_i \mid y_{<i}, P, I, m)\,, \tag{1}$$

where y is text tokens for ST-CQA and visual tokens for TFSF. The two tasks differ in inputs and output modality. For brevity, we hereafter absorb I and m into P; we extract its elements when needed.

Unified Output Tokenization. Input images are encoded as continuous features by the native vision encoder of the backbone; discretization is used only on the output side. We discretize each 256×256 target image I into visual tokens $z = \mathrm{Tok}(I)$ using a frozen MoVQGAN tokenizer (Zheng et al., 2022) with codebook size $K = 16{,}384$ and sequence length $L = 1{,}024$. Text and visual tokens share the same autoregressive head, trained with a single next-token prediction loss over the mixed-modality sequence s:

$$\mathcal{L}_{\mathrm{AR}}(\theta) \;=\; -\sum_{i=1}^{|s|} \log p_\theta(s_i \mid s_{<i}, P)\,, \tag{2}$$

where s_i is either a text or a visual token, and visual tokens are decoded as $\hat{I} = \mathrm{Dec}(z)$ at inference.

Prompt Templates. Each prompt P is a structured set whose elements include image content and metadata serialized as text. The three forms used throughout this work share the same metadata conditioning and differ only in their language and visual content:

$$P_{\mathrm{TFSF}} = \{I_{\mathrm{cur}},\ T_{\mathrm{ins}},\ m_t,\ m_{t'}\}, \tag{3}$$

$$P_{\mathrm{FSF}} = \{I_{\mathrm{cur}},\ m_t,\ m_{t'}\}, \tag{4}$$

$$P_{\text{ST-CQA}} = \{I_{\mathrm{pre}},\ I_{\mathrm{post}},\ Q,\ m_{\mathrm{pre}},\ m_{\mathrm{post}}\}. \tag{5}$$

P_{TFSF} provides the current observation, a natural-language instruction T_{ins} describing the desired changes, and target metadata $m_{t'}$; P_{FSF} is a text-free variant used during pre-training (Section 3.3); and $P_{\text{ST-CQA}}$ packages a bi-temporal pair with a question Q about observed changes. Conditioning every prompt on metadata lets the model later disentangle physical land-cover changes from sensor-induced variations such as illumination or off-nadir geometry. Target solar geometry and viewing angle are known before acquisition, as they follow from its time, location, and tasking plan. Target cloud cover, in contrast, must be estimated, for example from weather forecasts.

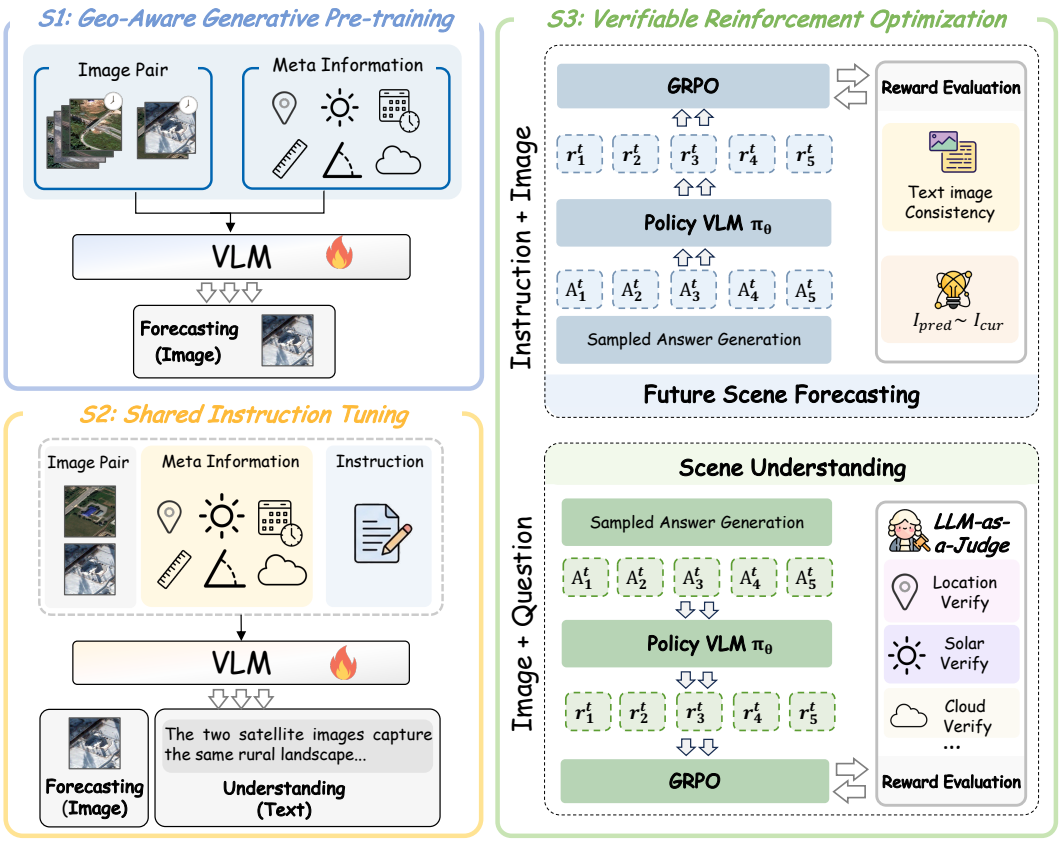


Figure 3: **Overview of RS-WorldModel.** A vision-language world model trained in three stages: **S1:** geo-aware generative pre-training with metadata-conditioned forecasting, **S2:** shared instruction tuning on both tasks, and **S3:** verifiable reinforcement optimization with task-specific rewards.

3.2 Unified Architecture

RS-WorldModel is built on Qwen3-VL-2B-Instruct (Bai et al., 2025) and functions as a world model for satellite imagery: a single autoregressive backbone that interprets bi-temporal scenes and generates future scenes conditioned on a change description and target metadata. As illustrated in the left half of Figure 3, three input streams form one mixed-modality sequence: continuous features of the input images from the Qwen3-VL vision encoder, acquisition metadata (Section 3.1) serialized as text, and textual tokens that carry the task instruction or question. The MoVQGAN tokenizer (Zheng et al., 2022) is used only on the output side, to tokenize target images for training and to decode predicted visual tokens at inference. The model processes this sequence with a single transformer stack and emits its output through a shared autoregressive head that produces either text tokens (ST-CQA) or visual tokens (TFSF) depending on the prompt. Interpreting current states and predicting future states therefore reuse the same parameters, metadata serialization, and loss, in line with recent visual world models that unify scene understanding with future-state prediction under a single autoregressive backbone (Zeng et al., 2025; Wang et al., 2026).

3.3 Three-Stage Training

RS-WorldModel is trained in three stages that build world-modeling capability progressively: first acquiring physical priors that link metadata to visual appearance (**S1: GAGP**), then training perception and forecasting jointly under language supervision (**S2: SIT**), and finally aligning outputs with task-specific verifiable rewards (**S3: VRO**). All three stages share the architecture and forward pass of Section 3.2; they differ only in the data they consume and the training objective they optimize. Stages are connected by checkpoint initialization, with Figure 3 illustrating the full pipeline.

Stage 1: Geo-Aware Generative Pre-training (GAGP). The first stage instills a metadata-to-image prior without language supervision, so that the model learns how geographic and acquisition context shape visual appearance (lighting, shadows, seasonality, viewing geometry, and cloud cover) before being asked to follow textual instructions. For each location, we sample a source observation (I_{cur}, m_t) and a target observation $(I_{t'}, m_{t'})$, and condition the model on the prompt P_{FSF} (Equation (4)) to autoregressively predict the target visual tokens $z_{t'} = \text{Tok}(I_{t'})$:

$$\mathcal{L}_{\text{GAGP}}(\theta) \;=\; -\mathbb{E}\left[\sum_{i=1}^{|z_{t'}|} \log p_\theta(z_{t',i} \mid z_{t',<i},\, P_{\text{FSF}})\right]. \tag{6}$$

This stage grounds future-frame prediction in physical conditions before instruction tuning.

Stage 2: Shared Instruction Tuning (SIT). Initialized from the GAGP checkpoint, SIT introduces language supervision and jointly trains both tasks on $\mathcal{D}^{\text{SIT}} = \mathcal{D}_{\text{ST-CQA}} \cup \mathcal{D}_{\text{TFSF}}$. Each sample provides a prompt P, drawn from $\{P_{\text{TFSF}}, P_{\text{ST-CQA}}\}$ according to its task, together with a target sequence y. Regardless of output modality, the model is optimized under a unified next-token prediction objective:

$$\mathcal{L}_{\text{SIT}}(\theta) \;=\; -\mathbb{E}_{(P,y)\sim\mathcal{D}^{\text{SIT}}}\left[\sum_{i=1}^{|y|} \log p_\theta(y_i \mid y_{<i},\, P)\right]. \tag{7}$$

Joint training lets both tasks share one backbone: TFSF prompts add textual control to generation, and ST-CQA is learned on the same metadata-conditioned scene representation.

Stage 3: Verifiable Reinforcement Optimization (VRO). The final stage refines the SIT policy using Group Relative Policy Optimization (GRPO) (Guo et al., 2025; Shao et al., 2024) without a separate value network. For each prompt P we sample $G = 5$ rollouts, score them with task-specific verifiable rewards, and update the policy by maximizing the group-relative advantage with KL regularization back to the SIT policy. Here *verifiable* means that each reward is checked against fixed references rather than produced by a separately trained reward model: the instruction and the current observation for TFSF, and the reference caption together with the prompt metadata for ST-CQA. This limits reward hacking, but the rewards remain proxies for, not measurements of, the true change.

TFSF Reward. For a TFSF prompt P_{TFSF}, we extract the textual instruction T_{ins} and the current observation I_{cur}, and decode the predicted visual tokens z_{pred} into $\hat{I} = \text{Dec}(z_{\text{pred}})$. Two cosine similarities are then computed under a frozen vision-language embedding $f(\cdot)$ (Li et al., 2026):

$$s_{\text{it}} \;=\; \cos\Big(f(\hat{I}),\, f(T_{\text{ins}})\Big), \qquad s_{\text{ir}} \;=\; \cos\Big(f(\hat{I}),\, f(I_{\text{cur}})\Big), \tag{8}$$

$$r_{\text{TFSF}} \;=\; s_{\text{it}} \;+\; \lambda\, s_{\text{ir}}. \tag{9}$$

s_{it} rewards faithfulness to the textual instruction; s_{ir} anchors the output to the current observation as a soft spatial prior, with λ controlling the trade-off, since several future scenes are plausible.

ST-CQA Reward. The generated caption $\hat{y}$ is scored against the reference caption y by an LLM judge (Qwen3-30B-A3B-Instruct-2507) (Yang et al., 2025) that receives the full prompt context together with explicitly parsed metadata. The judge scores the response on five equally weighted dimensions, namely *changes*, *unchanged content*, *time*, *space*, and *environment*, with 20 points each (prompt in Figure 16). The time, space, and environment dimensions are cross-checked against the parsed metadata, covering location consistency, solar geometry, and cloud and atmospheric conditions. The output scalar in $[0, 100]$ is normalized to

$$r_{\text{ST-CQA}} \;=\; \text{clip}\left(\frac{\text{score}(\hat{y},\, y;\, P)}{100},\; 0,\; 1\right). \tag{10}$$

Optimization. GRPO updates the policy π_θ by maximizing the group-relative advantage A_{grp} with KL regularization toward the frozen SIT policy π_{θ_0}:

$$\max_\theta\; \mathbb{E}[A_{\text{grp}}(P, \hat{y})] \;-\; \gamma\, \text{KL}(\pi_\theta(\cdot \mid P) \,\|\, \pi_{\theta_0}(\cdot \mid P)). \tag{11}$$

Table 2: **Spatiotemporal change question-answering (ST-CQA) results on the 5K test subset.** RS-WorldModel is compared against closed-source, open-source, and domain-specific baselines. Gray rows (closed-source models and open-source models of 32B and above) are listed for reference and excluded from the bold and underline ranking; GPT-Score was not computed for GPT-5.4 and Qwen3.5-397B-A17B (–).

			N-Gram			Contextual Similarity			
Method	**Size**	**GPT-S↑**	**B-1↑**	**MTR↑**	**R-L↑**	**S-BERT↑**	**SimCSE↑**	**ST5↑**	**Len**
Closed-Source Model									
GPT-5.1	–	91.17	16.82	20.87	14.59	77.19	78.28	76.70	817
Gemini-3-Flash	–	88.02	31.75	22.49	19.64	84.31	84.27	82.22	350
GPT-5.4	–	–	27.79	21.73	18.79	79.57	80.58	79.00	514
Open-Source Model									
Qwen3-VL-32B	32B	87.79	33.41	25.25	21.67	87.11	84.95	84.10	385
InternVL3.5-38B	38B	83.44	37.80	18.94	19.72	81.74	79.97	79.30	237
Qwen2.5-VL-72B	72B	86.40	37.06	19.78	19.83	84.30	82.11	81.68	310
Qwen3-VL-235B-A22B	235B	87.64	31.25	24.35	20.22	83.10	83.48	81.90	406
Qwen3.5-397B-A17B	397B	–	28.43	22.12	19.25	83.49	81.97	80.52	413
Qwen3-VL	2B	75.14	36.79	19.01	21.71	79.47	78.10	77.46	257
	4B	80.85	34.26	22.44	21.75	80.76	79.70	78.01	334
	8B	76.79	39.07	19.68	20.35	80.08	78.43	78.78	238
LLaVA-OV-1.5	4B	65.85	36.70	15.90	18.79	75.52	76.26	75.00	183
	8B	68.96	39.71	17.09	18.89	77.54	76.91	77.14	202
InternVL3.5	2B	72.41	31.02	16.60	17.13	77.87	75.81	75.56	259
	4B	78.90	34.76	18.57	18.42	79.14	77.43	77.26	255
	8B	77.05	35.26	18.21	18.02	79.19	77.61	77.30	245
	14B	80.67	34.87	19.42	19.18	80.76	79.33	78.67	263
EarthDial-RGB	4B	17.51	0.00	0.97	3.12	29.34	31.15	38.76	10
TEOChat	7B	36.85	0.03	2.99	7.39	52.38	55.85	50.10	24
RS-WorldModel	2B	**86.20**	**50.59**	**22.50**	**26.35**	**90.45**	**86.75**	**88.32**	207

4 EXPERIMENTS

4.1 MAIN RESULTS

Quantitative Results. Tables 2 and 3 compare RS-WorldModel with closed-source, open-source, and domain-specific baselines.

(1) Understanding. Table 2 reports ST-CQA results. With only 2B parameters, RS-WorldModel ranks first among all open-source baselines on BLEU-1, ROUGE-L, and all three contextual similarity metrics. The gain over the same-scale Qwen3-VL-2B is substantial: ROUGE-L improves by 21% and S-BERT by 14%. RS-WorldModel also surpasses models 16–120× larger on most metrics, *e.g.,* Qwen3-VL-32B scores 84.10 on ST5-SCS, while RS-WorldModel reaches 88.32. On GPT-Score, which judges semantic coverage, RS-WorldModel trails the Qwen models of 32B–235B by at most 1.6 points (86.20 vs. 87.79 for Qwen3-VL-32B). We attribute this to domain training with explicit metadata, which off-the-shelf VLMs lack. As the references are drafted by Qwen3-VL-32B in our pipeline, these scores may partly reward in-domain style; on the independently captioned RSCC and LEVIR-CC, RS-WorldModel improves over its backbone on 9 of 10 metrics, so the gains transfer beyond our annotation style (Appendix E; expert audit in Appendix D).

Two domain-specific baselines, EarthDial-RGB and TEOChat, score below 40 on GPT-Score, indicating that existing remote sensing models are not designed for open-ended temporal captioning. GPT-5.1 (OpenAI, 2025) achieves the highest GPT-Score but answers in 817 tokens on average, nearly 4× RS-WorldModel, with lower n-gram and contextual scores; GPT-5.4 and Qwen3.5-397B-A17B (514 and 413 tokens) trail RS-WorldModel on every metric reported for them.

(2) Forecasting. Table 3 reports TFSF results. RS-WorldModel ranks first among all open-source models on every metric, reducing FID by 48% and 47% relative to CRS-Diff and FLUX.1-Kontext while attaining the highest CosSim and GPT scores. It is also closest to the realized future, with the best PSNR, SSIM, LPIPS, and DISTS among all methods, including closed-source ones; its PSNR of

Table 3: **Text-guided future scene forecasting (TFSF) results on the 1.6K test subset.** RS-WorldModel is compared against closed-source, open-source, and domain-specific baselines. Gray rows (closed-source models) are listed for reference and excluded from the bold and underline ranking.

Method	Size	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	CosSim↑	GPT Scores↑			
								Sim.	Qual.	OA	AA
Closed-Source Model											
Gemini-2.5-Flash Image	–	11.63	0.1871	0.4493	0.2964	46.14	69.21	46.63	46.95	93.58	46.79
GPT-Image-1.5	–	11.14	0.1211	0.5148	0.3367	83.51	66.05	46.94	47.06	94.00	47.00
GPT-Image-1-mini	–	13.23	0.2056	0.5327	0.3344	92.27	65.95	44.76	45.96	90.72	45.36
Open-Source Model											
BAGEL	7B	9.54	0.1025	0.5810	0.3847	78.47	62.82	44.25	42.13	86.38	43.19
SD3.5-Large-IPA	8B	9.95	0.1071	0.6847	0.3903	97.88	66.69	33.15	40.63	73.78	36.89
FLUX.1-Kontext	12B	12.07	0.1979	0.5790	0.3568	81.92	64.67	39.00	42.41	81.41	40.70
CRS-Diff	0.9B	9.89	0.1269	0.6214	0.3673	82.76	63.09	27.04	30.97	58.01	29.01
RS-WorldModel	2B	**17.20**	**0.3124**	**0.3864**	**0.2429**	**43.13**	**68.34**	**44.59**	**44.84**	**89.43**	**44.71**

17.20 exceeds the strongest external value of 13.23. CRS-Diff scores lowest on Similarity; BAGEL, a unified model like ours, is competitive on Similarity (44.25) but has a much higher FID (78.47); GPT-Image-1.5 leads on Similarity and OA with nearly double our FID. RS-WorldModel balances text faithfulness (s_{it}) with consistency to the current observation (s_{ir}) and surpasses Gemini-2.5-Flash Image on FID (43.13 vs. 46.14). It does not simply copy its input, since a persistence control scores only 9.82 PSNR (Appendix E). Without the target text it still reaches 16.09 PSNR, and with only the current image and target time 12.62 (Table 13).

Human and Spatial Evaluation. Experts chose the forecast of RS-WorldModel in 60.0% of 200 blind forced-choice judgements, three times the rate expected by chance (Table 9). It also attains the best boundary F1 and change-region IoU while altering the fewest stable pixels (Tables 11 and 12).

Qualitative Results. Figures 4 and 5 show representative examples from both tasks.

(1) Understanding. In Figure 4, two images of the same urban area are taken about three years apart. RS-WorldModel preserves the layout, identifies subtle surface-texture changes near the fire station, and attributes shadow differences to sun elevation and acquisition time, whereas EarthDial-RGB and TEOChat either describe a single image or report no change.

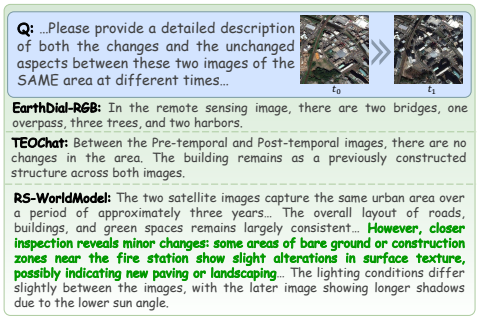


Figure 4: Qualitative comparison on temporal change understanding.

(2) Forecasting. In Figure 5, the prompts specify a snow-covered rural settlement, a stadium in early summer, and a school with denser parking and vegetation changes. RS-WorldModel preserves the input structure while realizing the requested changes, whereas the baselines often distort the layout or miss part of the change.

4.2 Ablation Study

Effect of λ in the TFSF Reward. λ in Equation (9) weighs s_{ir} against s_{it}. We sweep $\lambda \in \{0.0, 0.1, 0.2\}$ and report results on both tasks (Table 4). On TFSF, λ=0.2 is marginally best on every metric, consistent with the current observation supplying spatial priors, such as building layouts and road networks, that text alone cannot convey. On ST-CQA, λ=0.2 is also best on every metric except METEOR, which marginally favors λ=0.1. The gains are modest, so the reference term acts as a regularizer. Although a large λ could favor copying the input, at the adopted λ=0.2 the forecasts exceed a persistence control by 7.38 dB (Appendix E).

Ablation on Training Stages. We compare successive configurations of the pipeline (Table 5). GAGP alone reaches an FID of 50.28 but cannot follow instructions. Adding SIT enables both tasks,



Figure 5: Qualitative comparison on text-guided forecasting. RS-WorldModel better preserves scene structure while realizing the requested seasonal, illumination, and land-cover changes.

Table 4: **Ablation on the reference-adherence weight λ.** TFSF on 1.6K, ST-CQA on 5K.

	Forecasting (TFSF)						Understanding (ST-CQA)							
			GPT Scores↑					N-Gram			Contextual Sim.			
λ	FID↓	CosSim↑	Sim.	Qual.	OA	AA	GPT-S↑	B-1↑	MTR↑	R-L↑	S-BERT↑	SimCSE↑	ST5↑	Len
0.0	44.19	67.22	44.40	44.34	88.75	44.37	86.03	49.78	22.62	26.05	90.44	86.32	88.07	211
0.1	43.64	67.24	44.23	44.41	88.63	44.32	86.05	49.61	**22.79**	26.04	90.44	86.47	88.07	214
0.2	**43.13**	**68.34**	**44.59**	**44.84**	**89.43**	**44.71**	**86.20**	**50.59**	22.50	**26.35**	**90.45**	**86.75**	**88.32**	207

and VRO improves every reported metric, most on the GPT-based OA (83.20 to 89.43) and Quality (40.82 to 44.84) and least on FID (44.23 to 43.13), i.e., mainly faithfulness and quality. Removing geographic and acquisition metadata from GAGP raises its FID from 50.28 to 53.72 (Table 6).

Table 5: **Ablation on the three-stage training paradigm.** Forecasting on the 1.6K subset; understanding on the 5K subset. *GAGP-only uses metadata conditioning without instruction tuning.

			Forecasting (TFSF)					ST-CQA	
GAGP	SIT	VRO	FID↓	Sim.↑	Qual.↑	OA↑	AA↑	GPT-S↑	Len
✓	×	×	50.28*	–	–	–	–	–	–
✓	✓	×	44.23	42.38	40.82	83.20	41.56	85.24	201
✓	✓	✓	**43.13**	**44.59**	**44.84**	**89.43**	**44.71**	**86.20**	207

Table 6: **Geo-metadata ablation in GAGP.** FID on the 1.6K subset. Geo metadata lowers FID by 3.44.

Pre-training	FID↓
w/o Geo Metadata	53.72
w/ Geo Metadata	**50.28**

5 Conclusion

This work presents RSWBench-1.1M, a 1.1M-sample benchmark that pairs spatiotemporal change question answering with text-guided future scene forecasting and attaches geographic and acquisition metadata to every sample, together with RS-WorldModel, a 2B unified autoregressive baseline for both tasks. On ST-CQA, RS-WorldModel surpasses open-source models up to 120× larger on most metrics; on TFSF, it obtains the best PSNR, SSIM, LPIPS, DISTS, and FID among all compared methods. The study has limits: forecasts are single-step and evaluated at 256×256 on RGB imagery, the annotations are LLM-drafted with an expert audit of training records only, and forecasting without the target conditions (12.62 PSNR) remains far from solved; SAR imagery, higher output resolutions, and multi-step forecasting are natural next steps.

Reproducibility Statement

The main text describes the data construction and training. Appendix C gives details on RSWBench-1.1M, and Appendix G gives the experimental setups and implementation details.

AI Use Statement

Generative AI tools were used to polish the grammar and wording of the manuscript. The authors reviewed the final text and take responsibility for its content.

References

Rafael Almar, Erwin WJ Bergsma, Grégoire Thoumyre, Alain Giros, Patrick Marchesiello, Solange Lemai-Chenevier, Stéphanie Artigues, Sophie Loyer, and Jean-Marc Delvit. Global 1-km coastal bathymetry from sentinel-2 wave inversion using the satellite-to-shores (s2hores) toolbox. *Scientific Data*, 12(1):1941, 2025.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

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. *arXiv preprint arXiv:2501.17811*, 2025b.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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. *arXiv preprint arXiv:2510.26583*, 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. *arXiv preprint arXiv:2505.14683*, 2025.

Google. Gemini 3 flash: frontier intelligence built for speed, 2025. URL https://blog.google/products/gemini/gemini-3-flash/. Published: 2025-12-17; Accessed: 2026-02-26.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Zhongbin Guo, Yuhao Wang, Ping Jian, Chengzhi Li, Xinyue Chen, Zhen Yang, et al. Tamms: Change understanding and forecasting in satellite image time series with temporal-aware multimodal models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, Yu Liu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 224:272–286, 2025.

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

Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. *arXiv preprint arXiv:2312.03606*, 2023.

Aybora Köksal and A Aydın Alatan. Samchat: Introducing chain-of-thought reasoning and grpo to a multimodal small language model for small-scale remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 19:795–804, 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*, pp. 27831–27840, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *Advances in Neural Information Processing Systems*, 37:3229–3242, 2024.

Chenyang Liu, Keyan Chen, Rui Zhao, Zhengxia Zou, and Zhenwei Shi. Text2earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model. *IEEE Geoscience and Remote Sensing Magazine*, 2025.

Yuxi Lu, Biao Wu, Zhidong Li, Kunqi Li, Chenya Huang, Huacan Wang, Qizhen Lan, Ronghao Chen, Ling Chen, and Bin Liang. Remote sensing-oriented world model. *arXiv preprint arXiv:2509.17808*, 2025.

Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024.

OpenAI. Gpt-5.1: A smarter, more conversational chatgpt, 2025. URL https://openai.com/index/gpt-5-1/. Published: 2025-11-12; Accessed: 2026-02-26.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, et al. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14303–14313, 2025.

Datao Tang, Xiangyong Cao, Xingsong Hou, Zhongyuan Jiang, Junmin Liu, and Deyu Meng. Crs-diff: Controllable remote sensing image generation with diffusion model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

InstantX Team. Instantx sd3.5-large ip-adapter page, 2024.

Guoqing Wang, Pin Tang, Xiangxuan Ren, Guodongfang Zhao, Bailan Feng, and Chao Ma. Learning vision-language-action world models for autonomous driving. *arXiv preprint arXiv:2604.09059*, 2026.

Jinping Wang, Weiwei Song, Hao Chen, Jinchang Ren, and Huimin Zhao. Fusdreamer: Label-efficient remote sensing world model for multimodal data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2025a.

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

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. *arXiv preprint arXiv:2508.18265*, 2025c.

Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5805–5813, 2024.

Xingxing Weng, Chao Pang, and Gui-Song Xia. Vision-language modeling meets remote sensing: Models, datasets, and perspectives. *IEEE Geoscience and Remote Sensing Magazine*, 2025.

Nicholas Wright, John MA Duncan, J Nik Callow, Sally E Thompson, and Richard J George. Training sensor-agnostic deep learning models for remote sensing: Achieving state-of-the-art cloud and cloud shadow identification with omnicloudmask. *Remote Sensing of Environment*, 322:114694, 2025.

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 Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 2025.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. *arXiv preprint arXiv:2506.07218*, 2025.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

Linrui Xu, Ling Zhao, Wang Guo, Qiujun Li, Kewang Long, Kaiqi Zou, Yuhan Wang, and Haifeng Li. Rs-gpt4v: A unified multimodal instruction-following dataset for remote sensing image understanding. *arXiv preprint arXiv:2406.12479*, 2024.

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. *arXiv preprint arXiv:2505.21076*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Junyan Ye, Jun He, Xiang Zhang, Yi Lin, Honglin Lin, Conghui He, and Weijia Li. Satellite image synthesis from street view with fine-grained spatial textual guidance: A novel framework. *IEEE Geoscience and Remote Sensing Magazine*, 2025.

Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, Xing Wei, and Ning Guo. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*, 2025.

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.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024.

Zhenghui Zhao, Chen Wu, Xiangyong Cao, Di Wang, Hongruixuan Chen, Datao Tang, Liangpei Zhang, and Zhuo Zheng. Changebridge: Spatiotemporal image generation with multimodal controls for remote sensing. *arXiv preprint arXiv:2507.04678*, 2025.

Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425, 2022.

Qi Zhu, Jiangwei Lao, Deyi Ji, Junwei Luo, Kang Wu, Yingying Zhang, Lixiang Ru, Jian Wang, Jingdong Chen, Ming Yang, et al. Skysense-o: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14733–14744, 2025.

APPENDIX

The appendix is organized as follows. Appendix A discusses broader impacts, Appendix B reviews related work, Appendix C gives additional details on RSWBench-1.1M, Appendix D reports the human evaluations of the training annotations and of the test-set forecasts, Appendix E reports additional evaluations of both tasks, Appendix F presents case studies, Appendix G describes the experimental setups, and the last appendix lists the prompts.

A BROADER IMPACTS

Positive Impacts. Remote sensing world models may support socially beneficial applications when used as assistive tools for Earth-observation analysis. By linking observed Earth-surface changes with plausible future scenarios, RS-WorldModel can help analysts summarize land-cover change, inspect urban expansion, assess post-disaster damage, and explore future scenes under specified geographic and acquisition conditions. When paired with expert validation, such capabilities may make Earth-observation analysis more accessible and reduce the manual effort required to screen large satellite archives.

Risks and Limitations. The same capabilities require careful use. Generated satellite scenes are synthetic forecasts rather than real observations, and should not be used as direct evidence without verification. Forecasts are inherently uncertain, so incorrect predictions may affect downstream analysis if they are used without expert review. Model performance may also vary across regions, land-cover types, and acquisition conditions because public remote sensing archives are not uniformly distributed. Finally, training and evaluating large multimodal models incurs compute and energy costs.

Mitigation. We view RS-WorldModel as a research model rather than an autonomous decision system. Generated outputs should be clearly labeled as synthetic or forecasted, accompanied by the conditioning metadata when possible, and verified against real observations before operational use. For high-stakes settings such as disaster response, environmental enforcement, or public-sector planning, outputs should remain subject to human expert review and uncertainty-aware validation. Future releases should consider provenance metadata, output auditing, geographic bias evaluation, and more efficient training and adaptation strategies.

B RELATED WORK

Unified Multimodal Understanding and Generation. Unified multimodal models seek to handle visual understanding and controllable generation within a shared model interface. One line of work discretizes visual content and trains language-model backbones to perform mixed text-visual next-token prediction (Wu et al., 2025; Chen et al., 2025b;a; Xie et al., 2025); another develops native multimodal models with continuous or hybrid representations (Cui et al., 2025). Together, these methods show that a shared generative interface can transfer information between perception and synthesis. However, they are designed for generic visual domains and do not encode the geographic metadata, acquisition conditions, and temporal surface changes that govern remote sensing observations. Inspired by the broader idea of coupling perception and future prediction (Zeng et al., 2025), RS-WorldModel applies unified text-visual sequence modeling to two remote-sensing-specific tasks: spatiotemporal change question answering and text-guided future scene forecasting.

Remote Sensing Understanding, Generation, and World Modeling. Remote sensing vision-language models have produced strong understanding-oriented systems (Weng et al., 2025). This literature has expanded from image-level VQA and captioning to grounded spatial reasoning, open-world interpretation, multisensor comprehension, and temporal Earth-observation reasoning (Kuckreja et al., 2024; Hu et al., 2025; Zhan et al., 2025; Zhu et al., 2025; Zhang et al., 2024; Irvin et al., 2024; Köksal & Alatan, 2025). These models improve perception and language grounding, but they do not natively generate controllable future satellite observations. Remote sensing generation studies address the complementary problem of synthesizing satellite imagery. Generative foundation models, spatiotemporal image generation methods, and text-driven satellite synthesis resources have improved visual realism, condition adherence, and controllability (Khanna et al., 2023; Zhao et al., 2025; Liu et al., 2025; Ye et al., 2025). However, these methods are usually evaluated as

image generators rather than as models that jointly explain observed changes and forecast future states under the same geospatial context. Several recent works share partial overlap with our formulation: FusDreamer (Wang et al., 2025a) adopts the world-model framing for label-efficient multimodal classification and fusion, Remote Sensing-Oriented World Model (Lu et al., 2025) studies direction-conditioned spatial extrapolation under the same framing, and TAMMs (Guo et al., 2026) combines temporal change description with future image forecasting from observed time series. RS-WorldModel differs from the closest of these, TAMMs, by targeting fine-grained open-ended change question answering rather than coarse captioning, conditioning both tasks on explicit geographic and acquisition metadata, and conditioning future scene forecasting on an explicit change description and target metadata.

Remote Sensing Datasets for Understanding and Generation. Large-scale remote sensing vision-language datasets mainly support understanding tasks. VRSBench (Li et al., 2024) provides a versatile benchmark for image understanding, SkySenseGPT/FIT-RS (Luo et al., 2024) offers fine-grained instruction tuning data, RS-GPT4V (Xu et al., 2024) builds a multimodal instruction-following corpus, and SkyScript (Wang et al., 2024) contributes semantically diverse image-text annotations. EarthDial (Soni et al., 2025) and TEOChatlas (Irvin et al., 2024) increase the scale and temporal coverage of remote sensing understanding data, but they do not provide paired future-scene generation targets. Generation-oriented resources such as Git-10M (Liu et al., 2025), Street2Sat-Text, and CVACT-Text (Ye et al., 2025) support text-driven satellite image synthesis, yet they are not designed for spatiotemporal change question answering. In contrast, RSWBench-1.1M jointly supports understanding and forecasting: it pairs rich geographic metadata and fine-grained language annotations with tasks for both observed-change reasoning and text-guided future scene generation.

C Additional Details on RSWBench-1.1M

The 1.1M training corpus is constructed exclusively from the fMoW training split and allocated across the three training stages of RS-WorldModel: 371K samples for Geo-Aware Generative Pre-training (GAGP), 742K mixed samples for Shared Instruction Tuning (SIT), and 16K samples reserved for Verifiable Reinforcement Optimization (VRO). The 1.1M figure counts training samples across the stages, not unique image pairs. The 6.6K evaluation benchmark is drawn exclusively from the fMoW *test* split, balanced into 5K ST-CQA and 1.6K TFSF instances, preserving the global geographic diversity of the original test set.

Linguistic Characteristics. To complement the dataset overview in the main paper (Section 2), we report token-length distributions and word-cloud visualizations for the two subsets of RSWBench-1.1M in Figures 6 and 7. The forecasting subset (TFSF) has an average token length of 110.5, dominated by visual-attribute terms such as *vegetation*, *shadows*, *buildings*, and *unchanged*, reflecting prompts that specify both target appearance and stable structure. The understanding subset (ST-CQA) has a longer average length of 245.8 tokens, with a vocabulary centered on temporal and comparative terms such as *changes*, *remains*, *later*, and *consistent*, consistent with bi-temporal change descriptions. Together, the two distributions show that the forecasting prompts are compact and appearance-oriented, whereas the change descriptions are longer and comparative.

D Human Evaluation

Two parts of this work cannot be verified by automatic metrics alone: whether the LLM-drafted annotations of RSWBench-1.1M describe what the images actually show, and whether the forecasts of RS-WorldModel look plausible to people who interpret satellite imagery. We ran one expert study for each.

Expert Audit of Training Annotations. The annotations of RSWBench-1.1M are drafted and refined by LLMs (Section 2), so any systematic error would enter both training and evaluation. Two experts with backgrounds in remote sensing image interpretation audited 20,000 records from the training corpus, stratified by fMoW category: 10,000 temporal captions for ST-CQA and 10,000 future-scene generation prompts for TFSF. Each expert judged every record against the real image pair and its acquisition metadata, without seeing the other expert's scores. The binary rubric in Table 7 targets the errors that LLM-written change descriptions are prone to: hallucinated objects, statements that contradict the metadata, generic wording that would fit any scene, wrong or missed changes, and

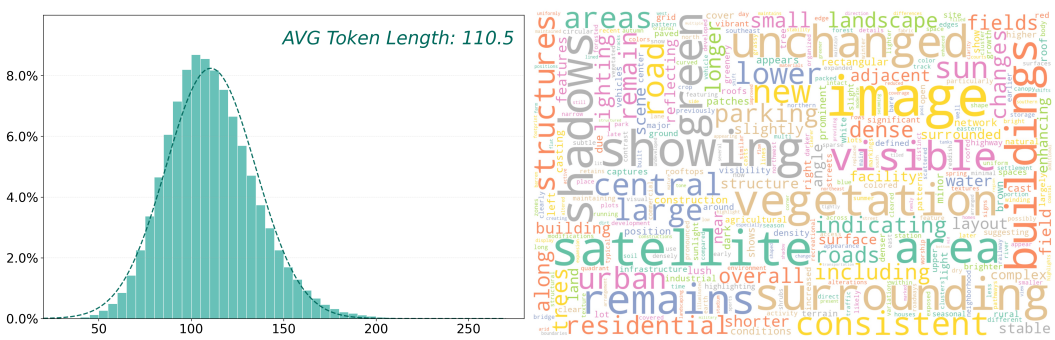


Figure 6: **Forecasting subset (TFSF).** Token-length distribution (left) and word cloud of frequent terms (right) for the text-guided future scene forecasting subset.

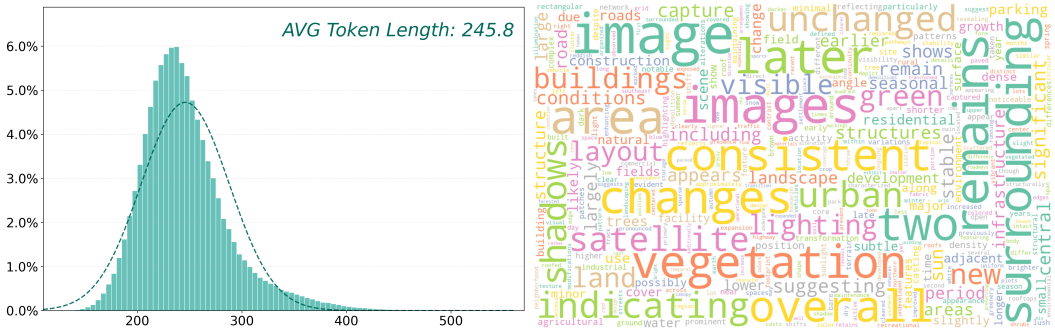


Figure 7: **Understanding subset (ST-CQA).** Token-length distribution (left) and word cloud of frequent terms (right) for the spatiotemporal change question-answering subset.

Table 7: **Audit rubric.** Three criteria apply to both annotation types, and each type adds its own.

Criterion	What it checks	ST-CQA caption	TFSF prompt
Object existence	Named objects and land-cover types are findable at the resolvable scale	✓	✓
Metadata fidelity	Statements about imaging conditions do not contradict the metadata	✓	✓
Descriptive precision	At least two location-bearing anchors identify this specific scene	✓	✓
Change correctness	Asserted changes are verified in both directions against the image pair	✓	
Causal attribution	Acquisition effects are not described as real surface change	✓	
Future novelty	The prompt adds scene content beyond the current image and metadata		✓

Table 8: **The experts pass about 90% of the criteria and agree on 88% of them.** Pass rate is the fraction of satisfied criteria per record; the overall pass rate weights records equally. Agreement is the raw agreement between the two experts over individual criteria, so its overall value is weighted by the number of criteria per annotation type.

Annotation	Criteria	Expert 1↑	Expert 2↑	Mean↑	Agreement↑
ST-CQA temporal caption	5	92.32	89.68	91.00	87.44
TFSF generation prompt	4	88.60	89.50	89.05	89.30
Overall	–	90.46	89.59	**90.03**	**88.27**

acquisition effects mistaken for surface change. Table 8 reports a mean pass rate of 91.00% for captions and 89.05% for prompts, 90.03% overall, and the two experts agree on 88.27% of individual criteria. The annotations are therefore largely faithful to the imagery, and the agreement indicates that the rubric can be applied consistently.

Table 9: **Experts choose the forecast of RS-WorldModel in 60% of the judgements.** Four annotators answered 50 forced-choice questions from the TFSF test set, giving 200 judgements; the methods were anonymised throughout the study.

Method	Selections	Selection rate (%)↑	Rank
CRS-Diff	1	0.5	5
BAGEL	3	1.5	4
Gemini-2.5-Flash Image	30	15.0	3
GPT-Image-1.5	46	23.0	2
RS-WorldModel	**120**	**60.0**	**1**

Expert Preference Study on Forecasts. Reference-based metrics compare a forecast with the one observation that was actually acquired, whereas an analyst asks whether a forecast is a plausible future of the scene. Four remote sensing annotators answered 50 forced-choice questions drawn from the TFSF test set. Each question shows one current observation and the futures predicted by five methods, namely RS-WorldModel, Gemini-2.5-Flash Image, GPT-Image-1.5, BAGEL, and CRS-Diff, anonymised as A to E and shuffled per question. Each option is played as an 8-second clip: the current observation, a smooth transition, the generated future, and finally the two side by side. The annotator picks the single most plausible future. The realized target image is not shown, so the study measures plausibility rather than agreement with the actual future, which the reference-based metrics cover. RS-WorldModel is chosen in 60.0% of the 200 judgements (Table 9), three times the 20.0% expected from an undiscriminating five-way choice. It is the strict plurality choice on 30 of the 50 questions, and all 10 questions on which the four annotators agreed were decided in its favour. BAGEL and CRS-Diff fall far below chance, at 1.5% and 0.5%, so the annotators actively rejected them rather than choosing at random.

E Additional Evaluation

Table 3 scores each forecast by its pixel and perceptual fidelity to the realized future. This appendix examines what those scores leave open: whether the model is simply copying its input, whether structures and changes land in the right places, how much of the result depends on the conditioning, and whether change understanding carries over to data built without our pipeline. Apart from the encoder similarity in Table 10 and the FID in Table 13, none of the forecasting measures below relies on a learned evaluator or an LLM judge.

Table 10: **RS-WorldModel is second only to Gemini-2.5-Flash Image in encoder similarity to the real target.** Cosine similarity ($\times 100$) between the embeddings of generated and target images on the 1.6K TFSF test subset.

Method	DINOv2 ViT-L/14↑	CLIP ViT-L/14↑
Gemini-2.5-Flash Image	**85.25**	**87.72**
GPT-Image-1.5	76.02	81.13
GPT-Image-1-mini	72.56	79.85
BAGEL	77.50	78.51
SD3.5-Large-IPA	67.00	74.35
FLUX.1-Kontext	72.58	77.46
CRS-Diff	73.90	75.95
RS-WorldModel	84.83	85.66

Persistence Control and Encoder Similarity. A forecast that returns its input unchanged already preserves every stable structure, so fidelity scores alone cannot rule this out. We therefore score a persistence control that outputs the current observation as its prediction. It reaches 9.82 PSNR, 0.1579 SSIM, 0.4686 LPIPS, and 0.3110 DISTS, against 17.20 PSNR for RS-WorldModel in Table 3, a gap of 7.38 dB. The control is not a weak one: on LPIPS and DISTS it is beaten only by RS-WorldModel and Gemini-2.5-Flash Image. Semantic agreement with the target points the same way. Measured as DINOv2 ViT-L/14 and CLIP ViT-L/14 similarity (Table 10), RS-WorldModel trails only Gemini-2.5-Flash Image and is well ahead of the remaining baselines.

Boundary Consistency. Fine-grained forecasting requires building footprints, roads, and field edges to stay where they are in the realized future. We binarize each forecast and its target with

Table 11: **RS-WorldModel preserves the boundaries of the real target best at every tolerance.** Mean boundary F1 per image (×100) on the 1.6K TFSF test subset, averaged over six deterministic edge extractors.

Method	BF1@1px↑	BF1@3px↑	BF1@5px↑
Gemini-2.5-Flash Image	51.01	76.96	86.75
GPT-Image-1.5	48.60	74.86	85.29
GPT-Image-1-mini	42.16	67.18	78.68
BAGEL	47.97	73.16	83.46
SD3.5-Large-IPA	36.42	59.34	70.85
FLUX.1-Kontext	40.90	62.82	72.75
CRS-Diff	39.33	64.20	76.13
RS-WorldModel	**54.23**	**79.13**	**87.85**

six deterministic edge extractors (automatic Canny, Otsu-Canny, Otsu-Sobel, Otsu-Scharr, Otsu-Laplacian, and Otsu morphological gradient) and compute the boundary F1 at tolerances of 1, 3, and 5 pixels, averaged over the extractors. RS-WorldModel obtains the highest BF1 at every tolerance (Table 11) and the best value under every extractor at the primary tolerance of 3 pixels. Its edge density stays within 0.99 to 1.02 times that of the reference, so the gain does not come from drawing more edges.

Table 12: **RS-WorldModel localizes changes best and alters the fewest stable pixels.** Mean scores per site (×100) on the 358 sites with exactly two acquisitions, averaged over five deterministic change-mask operators. Stable FCR is the fraction of pixels that are stable in the reference but marked as changed; Tol. F1@3 is mask F1 with a tolerance of 3 pixels.

Method	Mask IoU↑	Mask F1↑	Stable FCR↓	Tol. F1@3↑
Gemini-2.5-Flash Image	21.52	32.39	20.40	53.62
GPT-Image-1.5	27.46	40.51	33.14	65.97
GPT-Image-1-mini	28.28	41.43	35.18	66.61
BAGEL	23.49	35.00	31.95	58.06
SD3.5-Large-IPA	27.21	40.00	45.97	64.34
FLUX.1-Kontext	18.57	27.14	28.02	42.71
CRS-Diff	25.84	38.52	41.48	63.20
RS-WorldModel	**30.57**	**44.19**	**19.42**	**67.78**

Change-Region Consistency. Boundary agreement does not show whether the model changed the right regions. We compare a reference change mask, computed between the source image and the real target, with a predicted mask, computed between the same source image and the forecast. The source is unambiguous only at the 358 sites with exactly two acquisitions (22.4% of the test subset); all of them are used, a rule fixed before any model output was examined. Masks come from five deterministic operators whose Otsu thresholds are estimated once from the source and target responses and then held fixed for every method. RS-WorldModel has the highest mask IoU, mask F1, and tolerant F1, and the lowest stable false-change rate (Table 12): it changes more of what did change and less of what did not. These masks are a reproducible proxy for change regions, not human-annotated ground truth.

Table 13: **Removing the target imaging conditions hurts more than removing the target text.** TFSF on the 1.6K test subset; the model and sampling parameters are fixed, and only the conditioning part of the prompt varies.

Setting	Target text	Target conditions	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓
Minimal	None	Removed	12.62	0.1747	0.5695	0.3217	53.55
No target text	None	Given	16.09	0.2450	0.4741	0.2794	52.86
Coarse	One line	Given	17.07	0.3060	0.3914	0.2468	44.35
Detailed	Full	Given	**17.20**	**0.3124**	**0.3864**	**0.2429**	**43.13**

Conditioning Ablation. If TFSF reduced to text-guided editing, the target description would carry most of the result. We fix the model and the sampling parameters and weaken the conditioning in four steps (Table 13): the full target-scene description; a coarse one-line instruction filled from a fixed template that states only the elapsed time, the target date, the season, and the cloud cover; no target text, with the target acquisition metadata kept; and a minimal setting with only the current image

and the target time. The coarse instruction costs 1.22 FID. Removing the text entirely costs 1.11 dB of PSNR, yet the remaining 16.09 is still above every external method in Table 3, each of which receives the full description. Removing the target sun geometry, off-nadir angle, and cloud cover costs a further 3.47 dB, so the acquisition conditions weigh more than the description, consistent with their role in fixing illumination and viewing geometry.

Table 14: **RS-WorldModel improves over its base backbone on every metric except ROUGE-L on LEVIR-CC.** ST-CQA on RSCC and LEVIR-CC; values in parentheses are the relative changes over Qwen3-VL-2B, and bold marks the better value within each dataset.

Dataset	Method	ST5↑	S-BERT↑	SimCSE↑	MTR↑	R-L↑
RSCC	Qwen3-VL-2B (base)	49.48	43.11	39.12	9.35	6.40
	RS-WorldModel	**50.31**	**45.87**	**44.73**	**11.26**	**10.53**
	vs. base	*(+1.7%)*	*(+6.4%)*	*(+14.3%)*	*(+20.4%)*	*(+64.5%)*
LEVIR-CC	Qwen3-VL-2B (base)	37.39	26.76	24.74	3.75	**6.35**
	RS-WorldModel	**45.75**	**33.02**	**41.08**	**5.57**	5.41
	vs. base	*(+22.4%)*	*(+23.4%)*	*(+66.0%)*	*(+48.5%)*	*(−14.8%)*

Transfer to External Datasets. RSCC, with pre- and post-disaster pairs over 31 global events (988 examples), and LEVIR-CC, with bi-temporal pairs over 20 regions of Texas (1,929 examples), were captioned independently of our pipeline and record none of the acquisition metadata that RSWBench-1.1M provides. Across the five metrics on each dataset, RS-WorldModel improves over its base backbone Qwen3-VL-2B in 9 of the 10 cases (Table 14), by up to 64.5% on RSCC and 66.0% on LEVIR-CC; the exception is ROUGE-L on LEVIR-CC. Absolute scores stay below the in-domain ones partly because the reference captions are much shorter, averaging 11.9 and 59.4 words against 206.2 in RSWBench-1.1M.

F Case Studies

We present four representative case studies that qualitatively illustrate the behavior of RS-WorldModel on its two core tasks: one case for spatiotemporal change understanding and three cases for text-guided future scene forecasting.

Spatiotemporal Change Understanding. For this task (Figure 8), we provide a bi-temporal pair together with the instruction *"...Please provide a detailed description of both the changes and the unchanged aspects between these two images of the SAME area at different times..."* Most baselines fall into one of two failure modes: under-reporting, where subtle but real changes are missed, and over-reporting, where major modifications are hallucinated. RS-WorldModel instead correctly identifies the consistent overall layout, captures fine-grained vegetation growth near the buildings, and explains the difference in shadows in terms of acquisition time and sun angle, achieving the highest GPT-Score among all evaluated models.

Text-Guided Future Scene Forecasting. For this task (Figure 9), we evaluate three diverse scenarios spanning a coastal property, a rural settlement, and a military facility, each conditioned on a distinct textual instruction together with the corresponding geographic metadata. RS-WorldModel produces forecasts with markedly higher structural fidelity (preserved building footprints, road networks, and shoreline geometry), shadow consistency (illumination matching the prescribed sun angle), and text adherence (faithful realization of described seasonal and weather conditions). In these three examples, RS-WorldModel attains the highest GPT-based Similarity and Quality scores among strong open-source baselines.

G Experimental Setups

Evaluation Benchmarks. We evaluate RS-WorldModel on two tasks. *Spatiotemporal Change Question-Answering* (ST-CQA) measures how well a model describes bi-temporal changes; we report GPT-Score, BLEU-1, METEOR, ROUGE-L, S-BERT, SimCSE, ST5-SCS, and response length on a 5K subset (Table 2). *Text-Guided Future Scene Forecasting* (TFSF) measures whether a model synthesizes a plausible post-temporal image from a text instruction and geographic context; we report PSNR, SSIM, LPIPS, and DISTS against the real target image, FID, CosSim, and four GPT-based

Q: …Please provide a detailed description of both the changes and the unchanged aspects between these two images of the SAME area at different times…

EarthDial-RGB: There is no change to mention between the two images. Score:40

TEOChat: Between the Pre-temporal and Post-temporal images, there have been no changes to the area. The building remains as a previously constructed structure in both images. Score:45

LLaVA-OV-1.5-4B: … The pre-temporal image shows a clear day … the sun was lower in the sky, casting longer shadows … The post-temporal image also shows a clear day, but the sun is higher, resulting in shorter shadows … A large building that was previously empty has now been constructed, and there are new roads and parking lots that were not present before … The surrounding trees have grown taller … Overall, the area has undergone significant development and growth since the pre-temporal image was taken. Score:35



t_0

LLaVA-OV-1.5-8B: … The buildings within the area have undergone some modifications … Some structures appear to have been expanded or altered in shape, indicating possible construction or renovation activities … The layout of the roads seems to have changed slightly, with some sections appearing wider or having additional lanes, which could imply road expansion or traffic management adjustments … The overall shape and size of the central building complex remain consistent … The general pattern of the surrounding roads and their connections … have not changed significantly. Score:73



t_1

InternVL3.5-4B: … The arrangement and presence of most buildings remain consistent … Major roads and open clearings … maintain their layout … The large industrial complex at the bottom right … is unchanged in position and appearance … The landscaped areas … are now fully covered with lush green trees … there has been a noticeable change in vegetation … The sunlight angle and elevation have changed significantly … Post-temporal sunlight is higher and more direct compared to the Pre-temporal image. Score:72

InternVL3.5-8B: … The area has seen increased construction and urban sprawl … New buildings and structures are more visible in the post-temporal image, indicating expansion and development in the urban landscape … There is noticeable growth in greenery, particularly along roads and in previously open areas … Roads and pathways may have been widened or altered … The overall layout of major buildings and infrastructure seems to retain its core structure … the fundamental structure of the area remains recognizable across the time frames. Score:75

Qwen3-VL-4B: … the most noticeable change is the significant increase in vegetation cover … in the post-temporal image, dense green foliage has taken over, suggesting seasonal growth or the passage of time … The road network, including the highway and intersecting streets, remains unchanged in layout and structure … Buildings and industrial structures … remain visually identical in shape, size, and placement … However, this does not affect the underlying features … the changes observed are purely in the vegetation and seasonal growth patterns, not in the built environment or infrastructure. Score:80

Qwen3-VL-8B: … A new road segment appears in the lower right portion of the second image, connecting to the existing highway … the layout of roads near the central buildings has been adjusted … there seems to be an addition or modification to the structure itself, possibly involving a new roof or expansion … The major highway running horizontally through the center of the scene remains largely unchanged … The large, prominent building complex located centrally … retain their fundamental form and location relative to each other. Score:58

RS-WorldModel: … The overall layout of roads, parking lots, and buildings remains consistent … The central ground transportation station … retains its position and general footprint … The surrounding **parking areas** show minor variations in vehicle density and arrangement, likely due to daily or seasonal fluctuations rather than permanent changes … Vegetation near buildings appears **more lush and mature** in the later image … **the later image showing a higher sun angle and shorter shadows.** Score:93

Figure 8: **Case study on spatiotemporal change understanding (ST-CQA).** Given a bi-temporal pair of the same area, RS-WorldModel accurately reports the consistent layout, fine-grained vegetation growth, and illumination differences attributable to acquisition time, while baselines either miss the changes entirely or hallucinate large structural modifications. GPT-Scores are shown next to each response.

scores (Similarity, Quality, Overall, Avg.) on a 1.6K subset (Table 3). The ST-CQA GPT-Score is computed by GPT-5-Nano (Figure 12) and the TFSF GPT-based scores by GPT-4o (Figure 15).

Baselines. For ST-CQA, we compare against closed-source models (GPT-5.1 (OpenAI, 2025), Gemini-3-Flash (Google, 2025), GPT-5.4), generic open-source VLMs spanning 2B–397B parameters (Qwen-VL series (Bai et al., 2025), Qwen3.5-397B-A17B, LLaVA-OV (An et al., 2025), InternVL3.5 (Wang et al., 2025c)), and two remote sensing domain-specific VLMs (EarthDial-RGB (Soni et al., 2025), TEOChat (Irvin et al., 2024)). For TFSF, baselines include closed-source generators (Gemini-2.5-Flash Image (Comanici et al., 2025), GPT-Image-1.5, GPT-Image-1-mini), generic open-source generators covering complementary paradigms (adapter-based SD3.5-Large-IPA (Team, 2024) and FLUX.1-Kontext (Labs et al., 2025), and the unified model BAGEL (Deng



Figure 9: **Case study on text-guided future scene forecasting (TFSF).** Across three scenarios (coastal property, rural settlement, military facility), RS-WorldModel produces forecasts with stronger structural fidelity, shadow consistency, and text adherence than competing baselines. GPT-based Similarity and Quality scores are reported below each generation.

et al., 2025)), and the remote sensing domain-specific generator CRS-Diff (Tang et al., 2024). All baselines are evaluated as released, without training on RSWBench-1.1M. Each generation baseline receives the current image, the full target-scene description, and the target acquisition metadata written into the prompt as text, and all forecasts are evaluated at 256×256.

Implementation Details. RS-WorldModel is built on Qwen3-VL-2B-Instruct and trained on 8 NVIDIA A800 GPUs (80 GB each) in bf16 precision, using DeepSpeed ZeRO-3 and FlashAttention-2, with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.05) as the optimizer throughout. Input images are encoded by the Qwen3-VL vision encoder, and target images are tokenized by MoVQGAN. Acquisition metadata, including geographic coordinates, ground sampling distance, timestamps, sun azimuth and elevation, off-nadir angle, and cloud-cover ratio, is serialized as text in the prompt

alongside the visual context, so that physical conditions are processed in the same autoregressive sequence as image and text tokens. For Stages 1 and 2, the image resolution is capped at 256×256 pixels, with a context length of 32,768 tokens and a maximum generation length of 2,048 tokens; both stages use a per-device batch size of 16 with gradient accumulation of 2, and hold out 10% of the training data for validation every 1,000 steps. *Stage 1 (GAGP)* performs text-free generative pre-training on 371K forecasting samples for 8 epochs with a cosine schedule, peak learning rate 5×10^{-4}, and warmup ratio 0.10. *Stage 2 (SIT)* continues from the Stage-1 checkpoint and trains for another 4 epochs on 742K mixed understanding and forecasting samples, with the peak learning rate reduced to 1×10^{-4} and the warmup ratio to 0.02. *Stage 3 (VRO)* continues from the Stage-2 checkpoint and applies GRPO (Shao et al., 2024; Guo et al., 2025) on 16K mixed ST-CQA and TFSF samples; for each prompt we sample $G = 5$ rollouts and update the policy by maximizing the group-relative advantage in Eq. (11), with KL regularization retained against the frozen Stage-2 policy to preserve instruction-following behavior. The TFSF reward follows Eq. (9) with $\lambda = 0.2$.

H PROMPTS

To ensure reproducibility, we present all prompt templates used in our data construction, evaluation, and training pipeline. These prompts are carefully engineered for their respective roles: the *Qwen3-VL-32B Draft Generation Prompt* (Figure 10) and *Qwen2.5-72B Text Refinement Prompt* (Figure 11) enable scalable, high-quality annotation of RSWBench-1.1M; the *GPT-5-Nano ST-CQA Scoring Prompt* (Figure 12) and *GPT-4o TFSF Scoring Prompt* (Figure 15) provide reliable automatic scoring for understanding and generation tasks; the *Qwen3 LLM-as-a-Judge Prompt* (Figure 16) drives verifiable reinforcement optimization (VRO); and the *Stage-1 System Prompt* (Figure 13) together with the *Stage-2/3 System Prompt* (Figure 14) define RS-WorldModel's behavior across training stages. In particular, we adopt a pure LLM as the judge in VRO, inspired by Perception-R1 (Xiao et al., 2025), rather than a VLM. This design delivers more stable, semantically rich, and metadata-grounded reward signals for geographic and physical plausibility.

Draft Generation Prompt

System Prompt

You are an advanced AI model capable of combining geospatial metadata to generate temporal captions describing both changes and unchanged aspects between two temporal image phases, as well as constructing text prompts for generating post-temporal images.

You will be provided with two satellite images of the same geographic location, centered at coordinates (latitude, longitude) (`{center_lat_lon}`), but captured at different times. Geospatial metadata includes:

- **Acquisition times:** Pre-temporal image at `{time[0]}`, Post-temporal image at `{time[1]}`.
- **Bounding boxes for key objects** (normalized coordinates in the range [0, 1000)): Pre-temporal image `{"bbox_2d": {new_xyxy_box[0]}, "label": "{category[0]}"}`, Post-temporal image `{"bbox_2d": {new_xyxy_box[1]}, "label": "{category[1]}"}`.
- **Sun azimuth:** Pre-temporal image `{sun_azimuth[0]}` degrees, Post-temporal image `{sun_azimuth[1]}` degrees.
- **Sun elevation:** Pre-temporal image `{sun_elevation[0]}` degrees, Post-temporal image `{sun_elevation[1]}` degrees.
- **Off-nadir angle:** Pre-temporal image `{off_nadir_angle[0]}` degrees, Post-temporal image `{off_nadir_angle[1]}` degrees.
- **Cloud cover:** Pre-temporal image `{cloud_cover[0]}`%, Post-temporal image `{cloud_cover[1]}`%.

`<metadata-interpretation-guide>`

- **Hemisphere:** Positive latitude = Northern; negative = Southern.

- **Seasons (month-based):**
  - Northern: Mar-May (spring), Jun-Aug (summer), Sep-Nov (autumn), Dec-Feb (winter).
  - Southern: Sep-Nov (spring), Dec-Feb (summer), Mar-May (autumn), Jun-Aug (winter).
- **Sun azimuth** (sunlight direction, clockwise from north):
  - 0°: north.
  - 45°: northeast.
  - 90°: east.
  - 135°: southeast.
  - 180°: south.
  - 225°: southwest.
  - 270°: west.
  - 315°: northwest.
- **Sun elevation** (above horizon):
  - Low (∼0°): long shadows.
  - Medium (∼45°): moderate shadows.
  - High (∼90°): minimal shadows.

</metadata-interpretation-guide>
<objective>
Analyze the two satellite images to describe both changes and unchanged aspects between them. Provide a comprehensive analysis that covers visual differences and similarities across all observable elements, including but not limited to key objects, land cover, structures, vegetation, water bodies, urban development, natural features, and any other relevant aspects. Use the provided geospatial metadata only as a reference to inform your analysis, not make it the main focus or describe it directly. Instead, convert relevant metadata into natural, descriptive language without including raw numerical values or technical terms like degrees, percentages, or coordinates. Explain visual influences on changes and unchanged aspects, ensuring descriptions are reasonable, accurate, and not dominated by metadata details or speculation.
</objective>
<response-format>
Output as a structured JSON object:

```
{
  "temporal_caption": "A comprehensive caption describing both the
    changes and unchanged aspects between the two images.",
  "Post-temporal_image_generation": "A textual prompt for
    generating the Post-temporal image, phrased as 'A satellite
    image showing [detailed predicted changes based on observed
    differences].'"
}
```

</response-format>

Figure 10: Prompt template for draft generation using Qwen3-VL-32B-Instruct in the scalable data construction pipeline.

Text Refinement Prompt

System Prompt

You are an advanced AI model specialized in refining geospatial analyses, correcting and polishing temporal captions, and enhancing image generation prompts based on satellite imagery insights.

You are provided with geospatial metadata and initial analysis outputs for two satellite images of the same geographic location, centered at coordinates (latitude, longitude) (`{center_lat_lon}`), but captured at different times. Geospatial metadata includes:

- **Acquisition times:** Pre-temporal image at `{time[0]}`, Post-temporal image at `{time[1]}`.
- **Bounding boxes for key objects** (normalized coordinates in the range [0, 1000)): Pre-temporal image `{"bbox_2d": {new_xyxy_box[0]}, "label": "{category[0]}"}`, Post-temporal image `{"bbox_2d": {new_xyxy_box[1]}, "label": "{category[1]}"}`.
- **Sun azimuth:** Pre-temporal image `{sun_azimuth[0]}` degrees, Post-temporal image `{sun_azimuth[1]}` degrees.
- **Sun elevation:** Pre-temporal image `{sun_elevation[0]}` degrees, Post-temporal image `{sun_elevation[1]}` degrees.
- **Off-nadir angle:** Pre-temporal image `{off_nadir_angle[0]}` degrees, Post-temporal image `{off_nadir_angle[1]}` degrees.
- **Cloud cover:** Pre-temporal image `{cloud_cover[0]}`%, Post-temporal image `{cloud_cover[1]}`%.

The following are the initial outputs from a previous analysis:

- **temporal_caption:** `"{temporal_caption}"`.
- **Post-temporal_image_generation:** `"{post_temporal_image_generation}"`.

`<metadata-interpretation-guide>`

- **Hemisphere:** Positive latitude = Northern; negative = Southern.
- **Seasons (month-based):**
  - Northern: Mar-May (spring), Jun-Aug (summer), Sep-Nov (autumn), Dec-Feb (winter).
  - Southern: Sep-Nov (spring), Dec-Feb (summer), Mar-May (autumn), Jun-Aug (winter).
- **Sun azimuth** (sunlight direction, clockwise from north):
  - 0°: north.
  - 45°: northeast.
  - 90°: east.
  - 135°: southeast.
  - 180°: south.
  - 225°: southwest.
  - 270°: west.
  - 315°: northwest.
- **Sun elevation** (above horizon):
  - Low ($\sim$0°): long shadows.
  - Medium ($\sim$45°): moderate shadows.
  - High ($\sim$90°): minimal shadows.

</metadata-interpretation-guide>
<objective>
Refine and polish the provided temporal_caption and Post-temporal_image_generation based on the context and metadata guide.
For the temporal_caption:

- Correct any inaccuracies or inconsistencies with the geospatial metadata.
- Improve clarity, conciseness, and flow for better readability.
- Ensure it comprehensively covers both changes (e.g., urban development, vegetation shifts) and unchanged aspects (e.g., stable land features) across all observable elements.
- Naturally incorporate metadata insights (e.g., lighting effects, seasonal influences) into descriptive language without using raw numerical values or technical terms.

For the Post-temporal_image_generation:

- Reference key details from the refined temporal_caption to supplement important information (e.g., add specific predicted changes, visual effects from metadata).
- Enhance it to be a more detailed, vivid prompt for image generation, focusing on observed differences and alignments with the caption.
- Keep the phrasing as `'A satellite image showing [detailed predicted changes based on observed differences].'`
- Avoid speculation; base enhancements on provided analysis.

Use the provided geospatial metadata only as a reference to inform your analysis, not make it the main focus or describe it directly. Instead, convert relevant metadata into natural, descriptive language without including raw numerical values or technical terms like degrees, percentages, or coordinates.
</objective>
<response-format>
Output as a structured JSON object:

```
{
  "temporal_caption": "Refined comprehensive caption describing
    both the changes and unchanged aspects between the two images.",
  "Post-temporal_image_generation": "Enhanced textual prompt for
    generating the Post-temporal image, phrased as 'A satellite
    image showing [detailed predicted changes based on observed
    differences].'"
}
```

</response-format>

Figure 11: Prompt template for text refinement with Qwen2.5-72B-Instruct in the data construction pipeline.

ST-CQA GPT-Score Prompt

Evaluate a model-generated change caption against a human-generated caption (ground truth) for the SAME area in pre-/post-temporal remote sensing images. Identify the aspects mentioned in the human caption (both changes and unchanged aspects) and calculate the percentage of these aspects correctly mentioned or partially matched in the model caption. Score from 0 to 100, where each aspect contributes equally to the score. Consider similar concepts for partial score. Provide your score (0–100) and a short justification (less than 15 words) in the format of `'score#reason'`.

Now score the following:
Human: `{ground_truth}`

Model: `{model_output}`
Output:

Figure 12: Prompt used by GPT-5-Nano to compute GPT-Score for the spatiotemporal change understanding (ST-CQA) task.

RS-WorldModel Stage-1 System Prompt

You are a remote sensing world model. Based on the provided particulars, you can generate future remote sensing imagery at any specified time in the future.

Figure 13: System prompt for RS-WorldModel in Stage 1 (Geo-Aware Generative Pre-training, GAGP).

RS-WorldModel Stage-2/3 System Prompt

You are a remote sensing world model. Based on the provided particulars, you can understand the images and generate future remote sensing imagery at any specified time in the future.

Figure 14: System prompt for RS-WorldModel used in Stage 2 (Shared Instruction Tuning) and Stage 3 (Verifiable Reinforcement Optimization).

TFSF GPT-Score Prompt

You are a strict judge for remote sensing future-image prediction.
`[TEXT_DESCRIPTION]`
`{text_description}`

`[PROMPT_METADATA]`
`{meta_tse}`

Return TWO integer scores (0–50) and one-sentence reason.
(**Similarity_score** (0–50): Text-Image Consistency
Evaluate whether the predicted post image matches `TEXT_DESCRIPTION`.
Use `PROMPT_METADATA` only to penalize obvious contradictions (e.g., target cloud=0 but heavy clouds). Do not require exact physics.

Quality_score (0–50): Image Quality
Evaluate the predicted post image as a remote sensing image: whether the scene looks realistic for the Earth surface and whether edges/boundaries are visually reasonable.
Focus on (1) scene realism/plausibility (no obviously impossible or incoherent land-cover/land-use patterns) and (2) edge/boundary quality (no wobbly/jagged/double edges, halos, boundary bleeding, or melting).

Output STRICTLY one-line JSON:
`{"Similarity_score": <int 0-50>, "Quality_score": <int 0-50>, "reason": "<one sentence>"}`.

Figure 15: Prompt used by GPT-4o to compute GPT-based scores for the text-guided future scene forecasting (TFSF) task.

LLM-as-a-Judge Prompt

Evaluate the Model caption against the Human caption (ground truth) for the SAME area in pre-/post-temporal remote sensing images.

Compare the key information stated in Human and Model for each dimension (changes, unchanged, time, space, environment). Give full/partial credit for semantically consistent information and reduce for contradictions. Score 0–100 using 5 dimensions (20 points each).

Output EXACTLY one line: `score#reason` (reason is one sentence).

PROMPT_METADATA:
`{meta_tse}`

Human Caption:
`{ground_truth}`

Model Caption:
`{model_output}`

Output:

Figure 16: LLM-as-a-Judge prompt template based on Qwen3-30B-A3B-Instruct-2507 for verifiable reinforcement optimization (VRO).