# THE CHANGE YOU WANT TO TALK ABOUT SPATIAL AND NARRATIVE CHANGE DETECTION IN EARTH OBSERVATION WITH A SINGLE VLM ARCHITECTURE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Bi-temporal change detection (CD) of Earth Observation images is fragmented into disconnected spatial and textual prediction tasks. While recent vision language models (VLM) explore open-vocabulary and query-driven challenges, their semantic flexibility is often constrained by decoupled, post-hoc matching pipelines. Besides, they often struggle to capture fine-grained temporal dynamics and remain restricted to isolated tasks, such as dense prediction, captioning, or grounding. We propose `LOSCAR`, a multi-task VLM capable of addressing diverse CD tasks under a unique paradigm, through specific prompting. By integrating a spatio-temporal architecture with a spatial decoder that translates multimodal features into dense image masks, our model jointly generates natural language descriptions and explicit spatial outputs under a continuous open-vocabulary regime. After multi-dataset training, our experiments show that `LOSCAR` achieves state-of-the-art results in change captioning while delivering accurate spatial localization and robust reasoning on complex semantic change trajectories.

## 1 INTRODUCTION

Earth observation (EO) images are key for the continuous monitoring of our planet's surface. A pivotal task is semantic change detection (SCD), identifying and characterizing surface evolutions between remote sensing images (Benidir et al., 2025). The increased availability of very high resolution (VHR) images calls for solutions capable of interpreting complex transitions (Zheng et al., 2023; Madani et al., 2026; Gonthier, 2026) and neglecting pseudo-changes (Zang et al., 2025). Traditionally, SCD has been formulated as a dense prediction task (*where*, *what*). Convolutional neural networks and transformers efficiently generate multi-class change masks (Wu et al., 2023; Corley et al., 2024; Pang et al., 2025) but (i) remain limited to closed-set scenarios, and (ii) cannot naturally address user-specific queries (*how much*, *describe*, *what replaced the forest?* (Bernhard et al., 2023) or (iii) adapt to the multiple facets of changes (Chang & Ghamisi, 2023; Zhu et al., 2022).

The recent explosion of vision language models (VLMs, (Liu et al., 2023; Bai et al., 2023; Zhang et al., 2026)) has introduced natural language as a flexible interface for querying EO images, with both visual question answering (VQA) and change captioning (Liu et al., 2022; Wang et al., 2025). These approaches are often limited by their purely textual outputs: they provide semantic context but fail to spatially ground these descriptions with precise, pixel-level masks (mainly bounding box coordinates as text). Conversely, *open-vocabulary* CD models (OVCD, (Li et al., 2026a; Zhu & Yang, 2025)) and text-guided referring tools (Liu et al., 2025; Korkmaz et al., 2026) successfully generate masks based on text prompts. Yet, OVCD frameworks generally rely on decoupled, multi-stage pipelines or diffusion models to simulate textures for a small, manually curated set of prototypes (often a few dozen classes). They lack the conversational interface, reasoning depth, and genuine semantic scalability of a VLM. The intrinsic complexity of EO images requires to handle interpretation, abstraction, and explanation simultaneously. Captioning (Karaca et al., 2025), VQA (Yuan et al., 2022), visual grounding (Li et al., 2024), referring (Korkmaz et al., 2026), and OV semantic prediction (Guo et al., 2026) are closely intertwined tasks. Rather than addressing them through disjoint pipelines, we propose an end-to-end multi-task architecture that enables both spatial and narrative change detection within a single VLM, with direct application to localization tasks.



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.

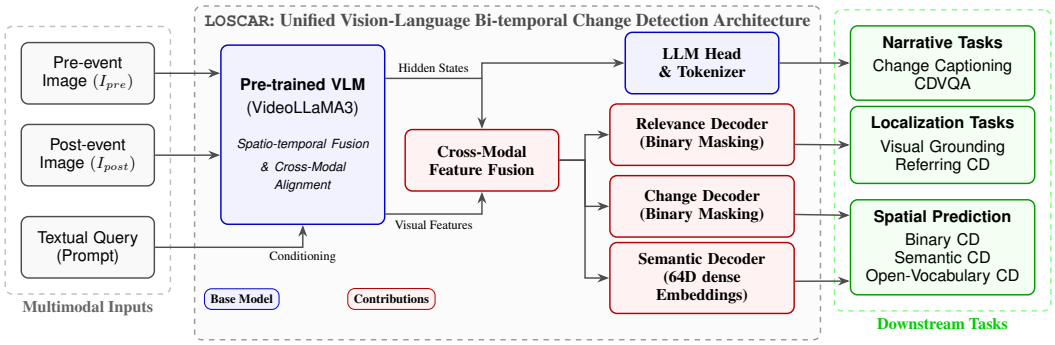



Figure 1: **Conceptual framework of LOSCAR.** The pre-trained VLM handles initial multimodal alignment. Our novel decoding branches and feature fusion modules enable to both process text and images to natively solve narrative, localizing, and open-vocabulary tasks.

Our framework, termed LOSCAR (*Language-guided Open-vocabulary Semantic ChAnge Reasoning*), seamlessly couples conversational reasoning with explicit, dense spatial maps. Grounded on a Video-Large Language Model (LLM) backbone (Zhang et al., 2023b) handling bi-temporal EO sequences, LOSCAR processes text and images and simultaneously outputs both natural language and structured spatial masks (Fig. 1). Beyond off-the-shelf VLM integration, we propose the following core mechanisms:

- A novel spatial decoding mechanism: while standard VLMs output 1D textual tokens, we design an image decoder that produces dense global change masks and semantic maps guided by the textual reasoning, in order to tackle real EO applications. It applies cross-attention directly between the linearized visual patches and the LLM semantic hidden states, before explicitly folding them back into the spatial domain.
- A truly open-vocabulary semantic space: we transition from closed-set class logits to a continuous semantic space. Instead of relying on a-priori classes or texture generation, we leverage large language models to extract an expansive, unconstrained vocabulary directly from our dataset.
- Implicit referring: we isolate specific changes without annihilating the model's global scene understanding through a dynamic *relevance mask*. Language acts as a filter, allowing us to localize a region targeted by a query, providing precise spatial grounding without additional annotation cost.

**Contributions.**

1. A novel, end-to-end multi-task vision-language architecture for change detection. By jointly training on narrative and spatial objectives, LOSCAR efficiently handles multiple CD tasks on Very High Resolution Earth Observation images.
2. A novel image decoder coupled with a genuine open-vocabulary strategy. Our model bridges the gap between conversational AI and pixel-perfect remote sensing predictions.
3. A robust quantitative and qualitative evaluation across three VHR datasets. We demonstrate the effectiveness of our architecture on both text generation and visual grounding.

## 2 Related Work

CD methods fall into three main paradigms, each of which divided into two tasks (Tab. 1).

**Dense spatial predictions.** **Semantic change detection** (SCD) (Chen et al., 2021; Zheng et al., 2021b) consists in predicting pixel-wise transitions. Most recent methods jointly infer semantic maps for each date and a binary change map (Ding et al., 2024; Chen et al., 2024; Long et al., 2025). They remain constrained by fixed taxonomies and closed-set supervision (Benidir et al., 2025) and fail to handle other transition labels (Chen et al., 2025; Zan et al., 2025). Alternatively, **open-vocabulary change detection** (OVCD) targets arbitrary changes defined by natural language or a given semantic label at inference (Zhu et al., 2026), often using multiple decoupled stages (Zheng et al., 2024): first, class-agnostic change proposals, followed by category assignment via visual prototypes or zero-shot matching against a set of target classes (Li et al., 2026a; Dou et al., 2026). Guo

| Task | SCanNet (Ding et al., 2024) | MambaSCD (Chen et al., 2024) | BGSNet (Long et al., 2025) | RSICCformer (Liu et al., 2022) | SAT-Cap (Wang et al., 2025) | MModalCC (Karaca et al., 2025) | CDVQA (Yuan et al., 2022) | VisTA (Li et al., 2024) | TEOChat (Irvin et al., 2025) | EarthDial (Soni et al., 2025) | LG-CD (Liu et al., 2025) | RCDNet (Korkmaz et al., 2026) | UniVCD (Zhu & Yang, 2025) | M-C-I (Li et al., 2026a) | I-M-C (Li et al., 2026a) | AdaptOVCD (Dou et al., 2026) | OpenDPR (Guo et al., 2026) | Semantic-CC (Zhu et al., 2024) | MCI (Liu et al., 2024) | Change3D (Zhu et al., 2025) | RSUniVLM (Liu & Lian, 2026) | DVLChat (Xuan et al., 2025) | **LOSCAR (Ours)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SCD | ✓ | ✓ | ✓ | | | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| OVCD | | | | | | | | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | ✓ |
| RCD | | | | | | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | ✓ | ✓ |
| CVG | | | | | | | | ✓ | ✓ | ✓ | | | | | | | | | | | | ✓ | ✓ |
| CC | | | | ✓ | ✓ | ✓ | | | ✓ | ✓ | | | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| CDVQA | | | | | | | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | ✓ | ✓ | ✓ |

Table 1: **Existing models and supported tasks** (only with multitemporal inputs). Only LOSCAR supports the three CD paradigms. Spatial prediction: SCD–Semantic CD; OVCD–Open-vocabulary CD. Localization: RCD–Referring CD; CVG–Change-based visual grounding; Narrative: CC–Change captioning; CDVQA–CD-based visual question answering.

et al. (2026) focus on category identification and replace image-text matching with prototypes. Zhu & Yang (2025) introduce a lightweight feature alignment module to bridge SAM2 spatial representations and CLIP semantic priors, enabling unsupervised OVCD without mask proposal generation. Alternatively, Seg2Change (Su et al., 2026) limits the influence of pseudo-changes, removing intermediate mask generators (Li et al., 2025) and inherent thresholds.

**Localization approaches.** **Referring change detection** (RCD) identifies specific types of change guided by natural language prompts, typically via CLIP-based text embeddings. While Korkmaz et al. (2026); Liu et al. (2025) allow language to act as a semantic filter and improve cross-dataset generalization over SCD, they remain limited to category-level prompts over a closed set of trained classes. A related direction extends visual grounding to bi-temporal remote sensing, jointly addressing VQA and change region localization (Li et al., 2024). We instead define **change-based visual grounding** (CVG) as a *standalone* task that extends RCD in two directions: (i) supporting natural language queries that encode spatial and relational context beyond bare category names,and (ii) generalizing to an open vocabulary, enabling the detection of change types not seen during training

**Narrative tasks.** **Change captioning** (CC) generates natural language descriptions, leveraging capabilities of Transformers (Liu et al., 2022; Wang et al., 2025), even with bi-temporal semantic masks as input (Karaca et al., 2025). VLMs now enable diverse conversational downstream tasks (Irvin et al., 2025; Soni et al., 2025). Recent approaches (Zhu et al., 2024; Liu et al., 2024; Zhu et al., 2025; Liu & Lian, 2026) integrate a dedicated image branch alongside the captioning branch to jointly output a change mask. Our model extends this end-to-end through a dedicated image decoder. **Visual question answering** (VQA) in remote sensing (Lobry et al., 2020) initially queries static images using natural language. Subsequent works (Zheng et al., 2021a) better align textual queries with visual EO (multitemporal) features. Yuan et al. (2022) introduce temporal reasoning and answer explicit queries. VQA architectures require adaptations for question-conditioned dense prediction and for localization using multiple EO images (Li et al., 2024; Deng et al., 2026) and for more intuitive, fine-grained visual explanations (Urooj et al., 2021; Bazi et al., 2026). Beyond VQA, some models (Soni et al., 2025; Irvin et al., 2025) encode bounding-box coordinates directly within their textual outputs, *i.e.*, implicit CVG without a dedicated localization head (Tab. 1).

**Multimodal large-language models in remote sensing.** The application of general-domain LLMs (Liu et al., 2023; Zhang et al., 2023b) to EO is challenged by the domain gap arising from the specific characteristics of VHR images. In many cases, CLIP visual encoding is coupled with a LLaMA2 or LLaVA text model (Muhtar et al., 2024; Kuckreja et al., 2024; Noman et al., 2025). Particular effort is devoted to improving instruction-following capabilities (Pang et al., 2024) and predominantly process single-date imagery. Inspired by Change3D (Zhu et al., 2025), we adopt a video-based approach, taking the VideoLLaMA3 (Zhang et al., 2025) architecture as starting point.

Recent VLM-based visual grounding approaches have explored different paradigms for producing spatial outputs: textual coordinates (Zhou et al., 2024), or explicit bounding boxes and pixel-level masks with dedicated spatial modules (Zhang et al., 2023a). We adopt this direction and extend it to the bi-temporal setting by equipping a video-based VLM with an image decoder.

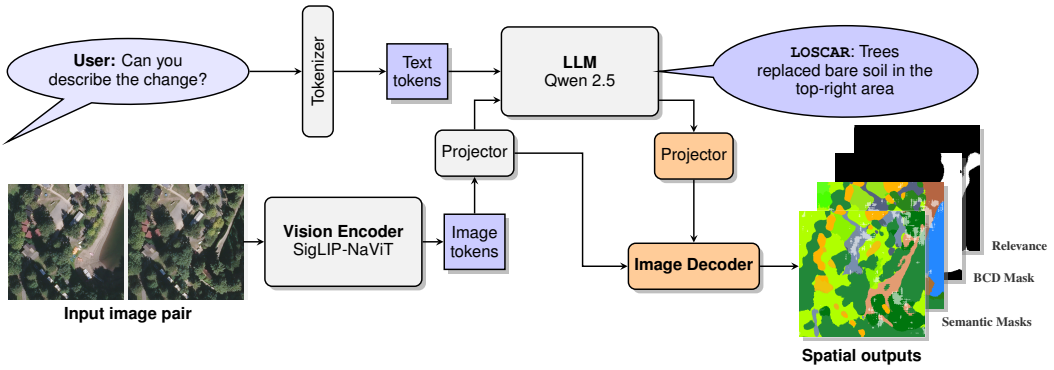



Figure 2: **Overview of LOSCAR.** Images patch embeddings are projected into the LLM token space and fed into the text token sequence. The LLM jointly processes all modalities. In contrast to standard VLMs, we insert an additional **Image Decoder** (Fig. 4) that both leverages the projected image tokens and the LLM hidden states to produce spatial outputs.

# 3 METHOD

## 3.1 OVERVIEW

Change detection is traditionally formulated either as pixel-wise semantic segmentation over a closed set of classes or as natural-language descriptions of change. However, these formulations do not readily extend to jointly handle bi-temporal images and integrate tasks across the three paradigms. LOSCAR introduces a **vision-language model for change detection** that reformulates change detection as a *multimodal reasoning problem*. We propose a unique end-to-end VLM that unifies: (i) **query-driven**, **open-vocabulary** change analysis, (ii) **rich textual explanations** paired with spatial outputs, and (iii) multiple tasks within a **single architecture**. LOSCAR takes as input bi-temporal images and a textual query and generates both natural language answers and structured image-based outputs (semantic embeddings, global change masks, and query-specific relevance masks) conditioned jointly on the images and the query (Fig. 2). Our framework builds upon VideoLLaMA3 (Zhang et al., 2025), chosen for its native ability to process multi-frame inputs and its robust handling of variable spatial resolutions.

**Vision encoder and tokenization.** We use the pretrained SigLIP-NaViT vision encoder. To handle very high resolution EO imagery without sacrificing fine-grained details, the model leverages the any-resolution vision tokenization (AVT) approach. AVT adapts the standard vision transformer architecture by replacing absolute position embeddings with 2D-rotary position embeddings (2D-RoPE). This dynamic tokenization enables that visual features are richly represented across varying spatial resolutions. Following VideoLLaMA3, the vision encoder and the projector are fine-tuned to seamlessly integrate with AVT.

**LLM Backbone.** The reasoning component of our architecture is based on a Qwen2.5-7B language model (Yang et al., 2024). Bi-temporal images ($I_{pre}$ and $I_{post}$) are processed by the vision encoder into patch embeddings, projected, and then concatenated with the textual query tokens before being fed into the LLM. In this linear sequence, each patch and text token is represented as a high-dimensional vector in $\mathbb{R}^{d_{llm}}$ (where $d_{llm} = 3584$). The LLM generates the contextualized hidden states encoding multimodal information (Fig. 1 and 2).

## 3.2 SPATIAL GROUNDING VIA THE IMAGE DECODER

While LLMs excel at generating textual descriptions, their outputs are inherently 1D. To generate pixel-wise change and semantic embeddings, we introduce a spatial Image Decoder. This module re-establishes the 2D spatial arrangement of the image tokens while conditioning them on the LLM's semantic hidden states. The decoding process follows five steps (see Fig. 4 in Appendix A.1):

1. **Feature projection:** The raw visual tokens and the final LLM hidden states are independently projected to a lower-dimensional space $d_{dec} = 256$ using linear layers ($W_v, W_h \in \mathbb{R}^{3584\times256}$).

2. **Cross-attention fusion:** We condition the visual features with the semantic information encoded by the LLM with a multi-head cross-attention mechanism. The projected visual features act as the *query*, while the projected LLM hidden states act as both the *key* and the *value*. A skip-connection is used: $F_{pre/post} = \text{query}_{pre/post} + \text{Attention}(\text{query}_{pre/post}, \text{key}, \text{value})$.
3. **Spatial reshaping:** The fused 1D tokens are folded back into a 2D grid, preserving the original patch locations. For a standard $512 \times 512$ input image with a patch size of $14 \times 14$ pixels, this yields a dense latent feature map of size $37 \times 37 \times 256$.
4. **U-Net decoding:** The 2D latent map is processed by a U-Net-style decoder. An initial bilinear upsampling scales the map to $64 \times 64$. Subsequent blocks, comprising double convolutions and ConvTranspose2D layers, progressively upsample the features back to the original image resolution.
5. **Multi-branch decoders:** We instantiate parallel decoder branches for the different spatial tasks. The *Change Decoder* takes the difference between the bi-temporal latent maps to output a binary $512 \times 512 \times 2$ map (change vs. no-change). The *Embedding Decoder* processes each latent map individually to produce continuous feature representations for open-vocabulary semantics. Finally, a *Relevance Decoder* outputs query-specific spatial masks.

### 3.3 Open-Vocabulary Semantics and Relevance Prediction

We move beyond from closed-set classification with an Embedding Decoder: it outputs a dense embedding map $M_{emb} \in \mathbb{R}^{512\times512\times64}$, where 64 is our open-vocabulary embedding dimension. Simultaneously, we project pretrained CLIP embeddings of a broad land-cover vocabulary (from $\mathbb{R}^{768}$ to $\mathbb{R}^{64}$). Pixel-level alignment is achieved directly in the embedding space by computing the cosine similarity between the predicted dense pixel embeddings and the projected text embeddings, without relying on a traditional cross-entropy loss over a fixed number of classes (Append. B.1–B.2).

Because not all changes are relevant to a user's specific query, we strictly decouple semantic understanding from query-targeting. We introduce a Relevance Decoder that projects the multimodal embeddings into a binary spatial mask. This mask acts as an attention filter, highlighting only the specific instances, trajectories, or damage regions explicitly requested by the user, while the Embedding Decoder retains a holistic semantic understanding of the entire scene.

### 3.4 Dataset Construction and Automated Query Generation

Training such a multi-task open-vocabulary model requires highly diverse, instruction-following Q/A data with perfectly aligned spatial supervision. We build a comprehensive dataset by aggregating three datasets that captures the heterogeneity of changes (Wang et al., 2026): FSC-180k (0.2 m GSD, (Benidir et al., 2025)), HiUCD (0.1 m GSD (Tian et al., 2020)), SECOND-CC ($\approx$ 1m GSD (Karaca et al., 2025)). To bridge the gap between pixel-level annotations and rich conversational capabilities, we develop a hybrid data-generation pipeline combining automated spatial analysis with LLM-driven semantic enrichment (details in Appendix B.4):

- **Generation of spatialized questions and answers:** A comprehensive set of diverse and complex questions is generated using the labels of the pre- and post-images within the reference datasets. The spatial and semantic difference between the corresponding bi-temporal masks over a predefined spatial grid (*e.g.*, $3\times3$) allows us to extract change trajectories, calculate affected areas, count specific instances, and identify unchanged regions. This ensures strict, deterministic accuracy for quantitative visual question answering.
- **LLM-Driven vocabulary expansion:** To break free from generic class names, we prompt the Gemini API with mono-temporal images to extract highly descriptive open-vocabulary terms. These terms are clustered and mapped to our base nomenclature using CLIP embeddings. This LLM-based enrichment is twofold: it injects precise, natural language descriptions into the generated Q/A texts, and it populates our open-vocabulary masks with rich semantic targets.
- **Spatial supervision alignment:** Each generated Q/A pair is associated with four spatial tensors: a binary change mask, a closed-set semantic mask, an open-vocabulary mask (global level), and a query-specific, instance level, relevance mask, used to supervise the Relevance Decoder.

### 3.5 Anti-Forgetting Regularization and End-to-End Training

To mitigate the catastrophic forgetting of general VQA capabilities that occurs when fine-tuning our VLM on domain-specific data, we adopt an aggressive data mixing strategy. We construct a distinct regularization subset by specifically prompting Gemini to act as a general conversational agent. Gemini generates generic visual Q/A pairs based on images from our dataset, as well as from the street-view images of the ViewDelta dataset (Varghese et al., 2025). This ensures our model retains its ground-level, non-EO perspective, and its general image-based descriptive capabilities.

The entire architecture is trained end-to-end with module-specific learning rates: lower rates for the pretrained LLM backbone (this ensures full sentence generation instead of few uninformative words), and higher rates for the randomly initialized Image Decoder. The overall training objective is a weighted sum of four losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{txt}}\mathcal{L}_{\text{txt}} + \lambda_{\text{chg}}\mathcal{L}_{\text{chg}} + \lambda_{\text{OV}}\mathcal{L}_{\text{OV}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}}, \quad \text{where} \tag{1}$$

- $\mathcal{L}_{\text{txt}}$ is the standard autoregressive cross-entropy loss for text generation.
- $\mathcal{L}_{\text{chg}}$ is computed against the ground-truth global binary change mask using a combination of binary cross-entropy (BCE) and Dice loss to mitigate spatial class imbalance.
- $\mathcal{L}_{\text{OV}}$ is the open-vocabulary loss, which maximizes the cosine similarity between the predicted 64D pixel embeddings and the projected CLIP text embeddings of our Gemini-based expanded vocabulary.
- $\mathcal{L}_{\text{rel}}$ supervises the Relevance Decoder using the query-specific instance masks, leveraging focal BCE and Dice losses to ensure the model accurately spatially grounds the user's specific request. Both losses help address the extreme class imbalance caused by the small number of relevance pixels (Appendix A.2).

## 4 Experiments and Results

### 4.1 Experimental Setup

**Evaluation metrics.** For textual generation, we report standard NLP metrics, with a particular focus on CIDEr and BERTScore-F1 (Anderson et al., 2016; Zhang et al., 2020). Since our LLM generates natural, conversational responses, these semantic metrics are better suited to capturing semantic similarity than rigid n-gram overlap metrics. Spatial predictions are evaluated using Binary IoU (IoU), mean IoU (mIoU), and Separated Kappa (SeK, (Yang et al., 2022)).
**Implementation details.** Our architecture is initialized from the VideoLLaMA3-7B checkpoint and trained end-to-end. Detailed training hyperparameters, hardware setup, and computational costs are provided in Appendix A.3. We will release the code, data, and model weights.

### 4.2 Spatial and Localization Tasks

Our quantitative analysis of pixel-level predictions is twofold: spatial tasks (full-image segmentation) and localization tasks (anchoring attention on specific regions related to a query).

| | SECOND-CC | | | | LEVIR-MCI | | |
|---|---|---|---|---|---|---|---|
| | BCD | | SCD | | BCD | | SCD |
| **Method** | IoU | F1 | mIoU | SeK | IoU | F1 | mIoU |
| SCanNet (Ding et al., 2024) | 61.1 | 75.8 | 73.4 | 22.3 | - | - | - |
| Delta-LLaVA (Li et al., 2026b) | - | - | 69.7 | 19.2 | - | - | - |
| ChangeFormer (Bandara & Patel, 2022) | - | - | - | - | 83.4 | 90.3 | - |
| CTMTNet (Shi et al., 2024) | - | - | - | - | 85.4 | 91.7 | - |
| MCI (Liu et al., 2024) | - | - | - | - | - | - | 86.4 |
| **LOSCAR (ours)** | **64.6** | **78.8** | **78.5** | **21.4** | **80.8** | **89.7** | **89.4** |

Table 2: **Dense spatial prediction.** Metrics are grouped into binary CD (assessing the global presence of change) and semantic CD (evaluating multi-class land-cover transitions). "-": the method has not been evaluated on such dataset and task in the original paper.

**Spatial tasks: binary and semantic change detection.** We evaluate the ability of the model to generate dense representations over the entire scene. LOSCAR demonstrates strong performances

for both Binary Change Detection (BCD) and Semantic Change Detection (SCD). On LEVIR-MCI, the high IoU and mIoU confirm that integrating a multimodal spatial decoding branch enhances spatial boundary delineation (Tab. 2). The model also maintains spatial consistency on on complex transitions (Fig. 3), yielding strong SepKappa scores on SECOND-CC.

**Open Vocabulary CD (OVCD).** Only qualitative assessment was performed (Fig. 3 and 8), showing satisfactory spatial and semantic discrimination. The model can distinguish closely related concepts, although this occasionally leads to slight redundancy (Fig. 3). Standard protocols typically compute a class-wise IoU on a restricted set of novel classes. Such protocols are not well suited for our ∼4,000 concept vocabulary. At this scale, the spatial frequency of any specific concept is extremely low: minor pixel mismatches cause large IoU drops. Additionally, exact-match metrics fail to account for semantic proximity.

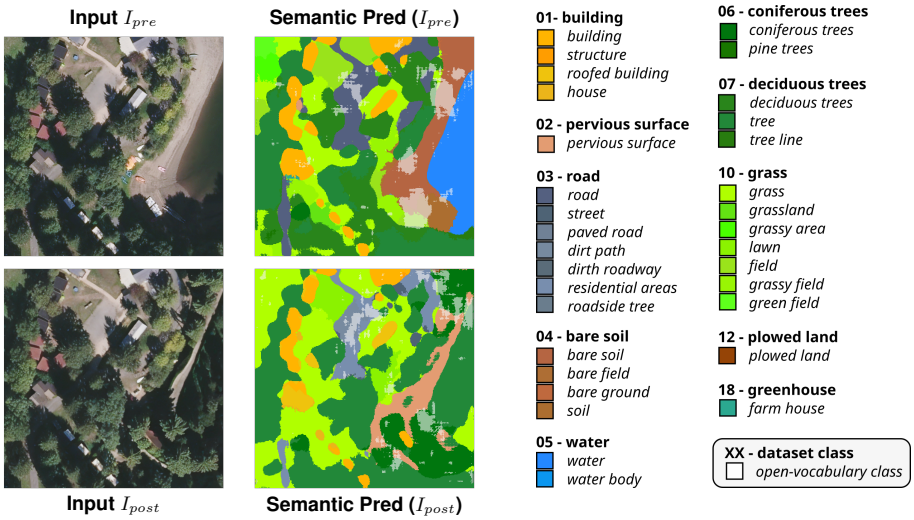



Figure 3: LOSCAR jointly provides spatial and narrative information for the Visual Question Answering task (FSC-180k dataset). The answer comes with pre- and post- dense semantic masks, using the open vocabulary labels discovered from the labels available in the initial dataset.

**Localization tasks: grounding and referring.** We assess the model's capacity to anchor its visual attention based on a textual prompt. For visual grounding (evaluated on QAG-360K (Li et al., 2024), Tab. 4) and referring change detection (evaluated on DVL-Suite (Xuan et al., 2025), Tab. 5), we rely on our Relevance Decoder. Unlike global change maps, the generated relevance mask isolates the exact area corresponding to the user's natural language query. On our pretraining set, LOSCAR achieves a mask IoU of 86.1% and an F1-score of 92.5%. Since queried targets usually cover a small fraction of the scene, we evaluate across all decision thresholds to compute the ROC Area Under Curve (AUC). The relevance head achieves an AUC of 71.1%, illustrating its capacity to isolate targeted queries from complex backgrounds under highly skewed spatial distributions.

### 4.3 Narrative Tasks

**Change captioning (CC).** We evaluate captioning on LEVIR-MCI (Liu et al., 2024) and SECOND-CC (Karaca et al., 2025). For LEVIR-MCI, we employ a prompt-free setting (empty strings) during fine-tuning. For SECOND-CC, we evaluate both the standard setting and a prompt-free fine-tuning setting. As shown in Tab. 3, LOSCAR consistently outperforms MCI (Liu et al., 2024). We achieve significantly higher METEOR and CIDEr scores, confirming that our continuous OV training leverages a wider vocabulary. To quantify this, we measure lexical diversity (Guo et al., 2025): LOSCAR achieves 0.26 on SECOND-CC (vs. 0.14 for MModalCC), breaking free from rigid

| Dataset | Method | B-1 | B-4 | METEOR | ROUGE-L | CIDEr | SPICE | BERT-F1 |
|---|---|---|---|---|---|---|---|---|
| SECOND-CC | MModalCC (Karaca et al., 2025) (w/o mask) | 56.6 | 7.9 | 38.4 | 40.9 | 43.0 | 17.3 | 70.5 |
| | MModalCC (Karaca et al., 2025) (w/ SCanNet) | 69.1 | 20.5 | 53.9 | 55.5 | 85.7 | 23.4 | 78.4 |
| | MModalCC (Karaca et al., 2025) (w/ GT masks) | 72.7 | 23.7 | 57.1 | 58.3 | 93.3 | 24.9 | 80.1 |
| | **LOSCAR (ours)** | 58.1 | 11.0 | 49.2 | 49.9 | 76.4 | 18.9 | 77.2 |
| | **LOSCAR (ours)** – prompt-free | **63.0** | **13.6** | **53.6** | **52.2** | **83.5** | **19.8** | **81.1** |
| LEVIR-MCI | MCI (Liu et al., 2024) | 75.7 | 29.5 | 66.3 | 66.6 | 118.0 | 25.0 | 84.0 |
| | **LOSCAR (ours)** – prompt-free | **81.2** | **33.4** | **73.8** | **73.5** | **136.0** | **28.8** | **88.1** |

Table 3: **Change captioning performance.** Higher values always indicate better performance. The best results (among methods relying solely on VHR images) are highlighted in **bold**. Grayed-out rows represent models that explicitly require reference semantic segmentation masks as side inputs.

captioning templates. The narrower gap on LEVIR-MCI (0.21 vs 0.19) may be explained by its more restricted thematic scope and its absence from the pretraining data.

| Method | QAG-360K | | | | CDVQA | | | |
|---|---|---|---|---|---|---|---|---|
| | CVG | | Textual answer | | Test set 1 (Text) | | Test set 2 (Text) | |
| | mIoU | oIoU | AA | OA | AA | OA | AA | OA |
| CDVQA (Yuan et al., 2022) | 17.7 | 29.3 | 62.5 | 68.6 | 55.3 | 65.9 | 55.4 | 61.1 |
| SOBA (Wang et al., 2024) | 32.7 | 41.8 | 66.3 | 72.0 | 60.3 | 69.2 | 60.3 | 64.8 |
| VisTA (Li et al., 2024) | 44.9 | 62.1 | 73.0 | 77.4 | 65.9 | 73.1 | 65.9 | 68.5 |
| **LOSCAR (ours)** | **54.9** | **67.2** | **67.2** | **74.4** | **38.3** | **74.8** | **67.0** | **69.7** |

Table 4: **Visual Question Answering performance.** We report narrative text generation metrics (Average Accuracy – AA, and Overall Accuracy – OA) alongside localization visual grounding metrics (mIoU and oIoU) on QAG-360K. For CDVQA, evaluation focuses on textual answers.

| Method | Localization | Narrative | | |
|---|---|---|---|---|
| | RCD (IoU) | BCA-QA Single Choice (Acc) | BCA-QA Multi Choice (Acc) | BCA-Report (GPT-Eval) |
| EarthDial (Soni et al., 2025) | - | 62.2 | 20.3 | 1.10 |
| ChangeMamba (Chen et al., 2024) | 32.4 | - | - | - |
| DVLChat 7B(Xuan et al., 2025) | 29.1 | 64.9 | 21.3 | 3.47 |
| Earth-OneVision (Cai et al., 2026) | - | 68.6 | 50.6 | - |
| **LOSCAR (ours)** | **34.7** | **83.6** | **62.9** | **3.81** |

Table 5: **Quantitative results on DVL-Suite** (Xuan et al., 2025). Performance is evaluated across spatial localization (Referring Change Detection), requiring dense mask prediction, and text-only narrative tasks (Basic Change Analysis: BCA-QA and BCA-Report).

**Visual question answering (VQA).** Beyond captioning, we evaluate our method on QAG-360K, CDVQA (Yuan et al., 2022), and DVL-Suite (Tab. 4). LOSCAR demonstrates competitive results on CDVQA and QAG-360K. On QAG-360K, it performs well in both textual reasoning and spatial grounding, validating our dual-decoder architecture. Interestingly, short-answer VQA presents a challenge for pretrained LLMs, which exhibit a verbosity bias. Adapting these conversational models to output short answers requires careful optimization, such as gradient accumulation, to suppress extra tokens that heavily penalize exact-match metrics. On DVL-Suite (Tab. 5), LOSCAR shows solid results across diverse textual formats. It accurately discriminates options in multiple-choice questions (DVL-Choice) and successfully anchors complex descriptions into precise relevance masks (RCD). However, the DVL-Report task highlights the inherent complexity of open-ended narrative generation. Synthesizing a comprehensive report requires the model to autonomously determine the hierarchy and level of detail of concurrent transitions. This highly unconstrained generation remains difficult to evaluate: numerous structural variations can validly describe the same scene.

### 4.4 Ablation Study

**Pretraining impact.** For **Spatial prediction** and **Localization** tasks, pretraining our model consistently improves its performances (Tab. 6): we observe slight improvement on BCD and RCD and a large gain in CVG (+6.6 points). Regarding **Narrative** tasks, the results are mixed, depending on the task complexity. On complex scenarios, such as the DVL-Suite Multi-Choice evaluation, the pretrained model exhibits a substantial improvement (53.3% to 62.9%). Conversely, we observe a small

performance drop on simpler, highly constrained tasks, requiring short answers (Single-Choice accuracy: 84.2% to 83.6%) as well as QAG-360K text answers. This behavior aligns with our design: our pretraining regime encourages the generation of descriptive, full-sentence narratives which may make the model becomes less specialized to outputting terse, rigid outputs. These results suggest that, rather than being specialized for straightforward, dataset-specific formats, the pretrained model retains greater flexibility when responding to ambiguous and complex queries.

| | Spatial | Localization | | Narrative | | |
|---|---|---|---|---|---|---|
| **LOSCAR setup** | LEVIR-MCI **BCD** (IoU) | QAG-360K **CVG** (mIoU) | DVL-Suite **RCD** (IoU) | QAG-360K **Text. Ans.** (OA) | CDVQA **Test Set 1** (OA) | DVL-Suite **BCA-QA** Single / Multi Choice (Acc) |
| w/o pretraining | 80.8 | 48.3 | 33.9 | **74.4** | 74.2 | **84.2** / 53.3 |
| w/ pretraining | **81.9** | **54.9** | **34.7** | 72.2 | **74.8** | 83.6 / **62.9** |

Table 6: **Impact of the pretraining step.** We compare our model trained from scratch on the downstream tasks (w/o pretraining) with the full pipeline utilizing our pretraining scheme (see Sec. 3.4).

**Architecture design.** Removing any of the task-specific decoders (Relevance, Change, Semantic) disables their corresponding downstream output (Tab. 7). Removing one branch leads to a noticeable a noticeable performance drop across the remaining tasks. For instance, removing the Relevance Decoder decreases SCD mIoU from 78.5 to 72.7. This confirms that the relevance branch helps bridge textual prompts and spatial layout, helping the model to link natural language land-cover concepts to their dense semantic embeddings. Omitting the Open-Vocabulary expansion yields only marginal changes in spatial metrics, but leads to a slight drop in narrative quality. The impact of the anti-forgetting strategy is inherently difficult to quantify using standard, task-specific benchmarks. Discarding this strategy does not necessarily decrease specialized metrics. It can even artificially inflate them, as the model becomes more specialized to narrow domain distributions at the potential cost of retaining broader knowledge. Its contribution lies in the preservation of the model's ability to provide structured paragraphs and to accurately respond to out-of-domain or unconstrained visual questions, which are relevant to broader practical use cases.

| | Spatial prediction | | Localization | Narrative |
|---|---|---|---|---|
| **Ablated component** | SECOND-CC **BCD** (IoU) | SECOND-CC **SCD** (mIoU) | SECOND-CC **RCD** (IoU) | SECOND-CC **CC** (CIDEr) |
| **Full Model (LOSCAR)** | **64.6** | **78.5** | 72.3 | **83.5** |
| w/o LLM head | 63.2 | 74.1 | 51.2 | ✗ |
| w/o Relevance Decoder | 60.3 | 72.7 | ✗ | 80.4 |
| w/o Change Decoder | ✗ | 71.2 | 65.7 | 81.8 |
| w/o Semantic Decoder | 61.4 | ✗ | 69.6 | 81.5 |
| w/o Open-vocabulary expansion | 64.4 | 77.6 | **72.4** | 83.1 |

Table 7: **Ablation study of key architectural components.** We report the intersection over union (IoU or mIoU) for spatial/localization tasks and CIDEr for narrative generation. The ✗ symbol indicates that the specific task cannot be performed when the corresponding component is discarded.

## 5 Conclusion

We propose a framework that encompasses the main change detection paradigms for change detection in bi-temporal remote sensing imagery. LOSCAR leverages the capabilities of recent LLM and VLM backbones and combines them with a new image decoder to generate both dense spatial predictions and textual outputs. Experiments show the versatility and performance of our framework across six tasks. Although LOSCAR demonstrates broad versatility, it exhibits certain limitations in terms of spatial precision. Compared with dedicated, task-specific segmentation networks, the generated semantic masks can be somewhat coarse. Similarly, the relevance masks and the localization of unseen classes currently lack fine-grained boundary sharpness. Nevertheless, LOSCAR exhibits broader multi-task and interactive capabilities than traditional task-specific models. By seamlessly bridging dense spatial prediction and complex multimodal understanding, this framework paves the way for a fully unified approach to remote sensing change detection.

### AI USE STATEMENT

We used generative AI tools (Gemini) to extract open-vocabulary concepts from remote sensing images (Sec. 3.4) and to construct the anti-forgetting question-answer dataset (Sec. 3.5).

We used generative AI tools to polish writing. We did not use generative AI tools as an help for developing theoretical models or conceptual frameworks, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, supporting qualitative and thematic data analysis, interpreting results.

### ETHICS STATEMENT

This work uses only publicly available Earth observation datasets (FSC-180k, LEVIR-MCI, SECOND-CC, ViewDelta). It does not involve human subjects, personal data, or crowdsourced annotation. Very high resolution aerial images were acquired at a spatial resolution ($> 0.2\,\mathrm{m}$) that does not enable identification of individuals. Similarly to other change detection workflows, `LOSCAR`'s learned representations could in principle be applied to dual-use tasks such as surveillance or military reconnaissance purposes. Our dataset, model weights, and code will be made publicly available in the forthcoming weeks.

### REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our work, we have thoroughly documented our methodology throughout the main text and supplementary material. The complete architectural details of our spatial Image Decoder, including the cross-attention fusion and the multi-branch U-Net topology, are illustrated and described in Section 3.2 and Appendix A.1. Comprehensive training details (encompassing differential learning rates, hardware setup, and computational costs) are provided in Appendix A.3. Furthermore, the exact procedure used to construct our open-vocabulary dataset, including concepts extraction via the Gemini API, prompt designs, NLP normalization, and semantic clustering, is fully detailed in Section 3.4, Section 3.5, and Appendix B. To facilitate future research and verify our claims, the complete source code, pre-trained model checkpoints, and our newly generated dataset will be made publicly available upon publication.

## REFERENCES

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and S. Süsstrunk. SLIC Superpixels Compared to State-of-the-art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:2274–2282, 2012. 18

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *ECCV*, 2016. 6

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv:2308.12966*, 2023. 1

Wele Gedara Chaminda Bandara and Vishal M. Patel. A transformer-based siamese network for change detection. In *IGARSS*, 2022. 6

Yakoub Bazi, Mohamad M. Al Rahhal, Mansour Zuair, and Faroun Mohamed. Revisiting change VQA in remote sensing with structured and native multimodal qwen models. *arXiv:2604.18429*, 2026. 3

Yanis Benidir, Nicolas Gonthier, and Clement Mallet. The change you want to detect: Semantic change detection in Earth observation with hybrid data generation. In *CVPR*, 2025. 1, 2, 5

Maximilian Bernhard, Niklas Strauß, and Matthias Schubert. MapFormer: Boosting Change Detection by Using Pre-change Information. In *ICCV*, 2023. 1

Miaoxin Cai, Guanqun Wang, Wei Zhang, Guangyao Zhou, Yin Zhuang, Tong Zhang, Hao Wang, He Chen, and Jun Li. Earth-OneVision: Extending Remote Sensing Multimodal Large Language Models to More Sensor Modalities and Tasks. *arXiv:2606.10819*, 2026. 8

Shizhen Chang and Pedram Ghamisi. Changes to Captions: An Attentive Network for Remote Sensing Change Captioning. *IEEE Transactions on Image Processing*, 32:6047–6060, 2023. 1

Hao Chen, Zipeng Qi, and Zhenwei Shi. Efficient transformer for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021. 2

Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. ChangeMamba: Remote sensing change detection with spatiotemporal state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024. 2, 3, 8

Yongqi Chen, Shou Feng, Chunhui Zhao, Yingjie Tang, Junjie Tang, Di Dong, and Nan Su. Deep learning for high-resolution remote sensing change detection: Task taxonomy, datasets, and perspectives. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:27886–27904, 2025. 2

Isaac Corley, Caleb Robinson, and Anthony Ortiz. A change detection reality check. In *ICLR*, 2024. 1

Pei Deng, Wenqian Zhou, and Hanlin Wu. DeltaVLM: Interactive remote sensing image change analysis via instruction-guided difference perception. *Remote Sensing*, 18:541, 2026. 3

Lei Ding, Jing Zhang, Haitao Guo, Kai Zhang, Bing Liu, and Lorenzo Bruzzone. Joint spatio-temporal modeling for semantic change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. 2, 3, 6

Mingyu Dou, Shi Qiu, Ming Hu, Yifan Chen, Huping Ye, Xiaohan Liao, and Zhe Sun. AdaptOVCD: Training-free open-vocabulary remote sensing change detection via adaptive information fusion. *arXiv:2602.06529*, 2026. 2, 3

Nicolas Gonthier. Operational change detection for geographical information: Overview and challenges. *International Journal of Applied Earth Observation and Geoinformation*, 149, 2026. 1

Qi Guo, Jue Wang, Yinhe Liu, and Yanfei Zhong. OpenDPR: Open-vocabulary change detection via vision-centric diffusion-guided prototype retrieval for remote sensing imagery. In *CVPR*, 2026. 1, 2, 3

Yanzhu Guo, Guokan Shang, and Chloé Clavel. Benchmarking linguistic diversity of large language models. *arXiv:2412.10271*, 2025. 7

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. In *ICLR*, 2025. 3

Ali Can Karaca, Enes Ozelbas, Saadettin Berber, Orkhan Karimli, Turabi Yildirim, and Mehmet Fatih Amasyali. Robust Change Captioning in Remote Sensing: SECOND-CC Dataset and MModalCC Framework. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:21494–21513, 2025. 1, 3, 5, 7, 8

Yilmaz Korkmaz, Jay N. Paranjape, Celso M. de Melo, and Vishal M. Patel. Referring change detection in remote sensing imagery. In *WACV*, 2026. 1, 3

Kartikeya Kuckreja et al. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. In *CVPR*, 2024. 3

Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. SegEarth-OV: Towards training-free open-vocabulary segmentation for remote sensing images. In *CVPR*, 2025. 3

Kaiyu Li, Xiangyong Cao, Yupeng Deng, Chao Pang, Zepeng Xin, Hui Qiao, Tieliang Gong, Deyu Meng, and Zhi Wang. DynamicEarth: How far are we from open-vocabulary change detection? In *AAAI*, 2026a. 1, 2, 3

Ke Li, Fuyu Dong, Di Wang, Shaofeng Li, Quan Wang, Xinbo Gao, and Tat-Seng Chua. Show me what and where has changed? Question answering and grounding for remote sensing change detection. *arxiv:2410.23828*, 2024. 1, 3, 7, 8

Xiaohe Li, Jiahao Li, Kaixin Zhang, Yuqiang Fang, Leilei Lin, Hong Wang, Haohua Wu, and Zide Fan. Decoding the Delta: Unifying Remote Sensing Change Detection and Understanding with Multimodal Large Language Models. *arXiv:2604.14044*, 2026b. 6

Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022. 1, 3

Chenyang Liu, Keyan Chen, Haotian Zhang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. Change-Agent: Toward interactive comprehensive remote sensing change interpretation and analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 3, 6, 7, 8

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 3

Xu Liu and Zhouhui Lian. RSUniVLM: A unified vision-language model for remote sensing via Granularity-oriented MoE. *Pattern Recognition*, 179:113717, 2026. 3

Yixiao Liu, Yizhou Yang, Jinwen Li, Jun Tao, Ruoyu Li, Xiangkun Wang, Min Zhu, and Junlong Cheng. LG-CD: Enhancing language-guided change detection through SAM2 adaptation. *arXiv:2509.21894*, 2025. 1, 3

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. RSVQA: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58:8555–8566, 2020. 3

Jiang Long, Sicong Liu, Mengmeng Li, Hang Zhao, and Yanmin Jin. BGSNet: A boundary-guided Siamese multitask network for semantic change detection from high-resolution remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 225:221–237, 2025. 2, 3

Seyedehanita Madani, Rama Chellappa, and Vishal M. Patel. DiffRegCD: Integrated registration and change detection with diffusion features. In *WACV*, 2026. 1

Dilxat Muhtar, Zhenshi Li, and P. Xiao. LHRS-Bot: Empowering remote sensing with VGI-enhanced large multimodal language model. In *ECCV*, 2024. 3

Mubashir Noman, Noor Ahsan, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. CDCHAT: A large multimodal model for remote sensing change description. In *IGARSS*, 2025. 3

Chao Pang, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Xingxing Weng, Shuai Wang, Litong Feng, Gui-Song Xia, and Conghui He. H2RSVLM: Towards helpful and honest remote sensing large vision language model. *arXiv:2403.20213*, 2024. 3

Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. ComPtr: Toward Diverse Bi-Source Dense Prediction Tasks via a Simple Yet General Complementary Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:8613–8629, 2025. 1

Jingye Shi, Mengge Zhang, Yuewu Hou, Ruicong Zhi, and Jiqiang Liu. A multitask network and two large-scale datasets for change detection and captioning in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024. 6

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, and Salman Khan. EarthDial: Turning multi-sensory Earth Observations to interactive dialogues. In *CVPR*, 2025. 3, 8

You Su, Yonghong Song, Jingqi Chen, and Zehan Wen. Seg2change: Adapting open-vocabulary semantic segmentation model for remote sensing change detection. *arXiv:2604.11231*, 2026. 3

Shiqi Tian, Ailong Ma, Zhuo Zheng, and Yanfei Zhong. Hi-UCD: A large-scale dataset for urban semantic change detection in remote sensing imagery. In *NeurIPS Workshop on Machine Learning for the Developing World*, 2020. 5

Aisha Urooj, Hilde Kuehne, Kevin Duarte, Chuang Gan, Niels Lobo, and Mubarak Shah. Found a Reason for me? Weakly-supervised Grounded Visual Question Answering using Capsules. In *CVPR*, 2021. 3

Subin Varghese, Joshua Gao, and Vedhus Hoskere. Viewdelta: Scaling scene change detection through text-conditioning. In *ICCV*, 2025. 6

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. EarthVQA: Towards Queryable Earth via Relational Reasoning-Based Remote Sensing Visual Question Answering. In *AAAI*, 2024. 8

Yelin Wang, Zijia Song, Shuo Ye, Chuanguang Yang, Miaoyu Wang, Yong Xu, Zhulin An, Yongjun Xu, and Zitong Yu. RSICCLLM: A Multimodal Large Language Model for Remote Sensing Image Change Captioning. In *ECCV*, 2026. 5

Yuduo Wang, Weikang Yu, and Pedram Ghamisi. Change Captioning in Remote Sensing: Evolution to SAT-Cap – A Single-Stage Transformer Approach. *arXiv:2501.08114*, 2025. 1, 3

Chen Wu, Bo Du, and Liangpei Zhang. Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:9774–9788, 2023. 1

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 *NeurIPS*, 2025. 3, 7, 8

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv:2407.10671*, 2024. 4

Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022. 6

Zhenghang Yuan, Lichao Mou, Zhitong Xiong, and Xiao Xiang Zhu. Change detection meets visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 1, 3, 8

Yujie Zan, Shunping Ji, Songtao Chao, and Muying Luo. Open-vocabulary generative vision-language models for creating a large-scale remote sensing change detection dataset. *ISPRS Journal of Photogrammetry and Remote Sensing*, 225:275–290, 2025. 2

Qi Zang, Dong Zhao, Shuang Wang, Dou Quan, and Zhun Zhong. Feature spectrum learning for remote sensing change detection. In *CVPR*, 2025. 1

Ao Zhang, Wei Ji, and Tat-Seng Chua. NExT-Chat: An LMM for Chat, Detection and Segmentation. *arXiv:2311.04498*, 2023a. 3

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv:2501.13106*, 2025. 3, 4

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv:2306.02858*, 2023b. 2, 3

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *ICLR*, 2020. 6

Xu Zhang, Danyang Li, Xiaohang Dong, Tianhao Wu, Hualong Yu, Jianye Wang, Qicheng Li, and Xiang Li. UniChange: Unifying Change Detection with Multimodal Large Language Model. In *CVPR*, 2026. 1

Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu. Mutual Attention Inception Network for Remote Sensing Visual Question Answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021a. 3

Zhuo Zheng, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Change is Everywhere: Single-temporal supervised object change detection in remote sensing imagery. In *ICCV*, 2021b. 2

Zhuo Zheng, Shiqi Tian, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Scalable multi-temporal remote sensing change data generation via simulating stochastic change process. In *ICCV*, 2023. 1

Zhuo Zheng, Yanfei Zhong, Liangpei Zhang, and Stefano Ermon. Segment Any Change. In *NeurIPS*, 2024. 2

Yue Zhou, Mengcheng Lan, Xiang Li, Litong Feng, Yiping Ke, Xue Jiang, Qingyun Li, Xue Yang, and Wayne Zhang. GeoGround: A unified large vision-language model for remote sensing visual grounding. *arXiv:2411.11904*, 2024. 3

Duowang Zhu, Xiaohu Huang, Haiyan Huang, Hao Zhou, and Zhenfeng Shao. Change3D: Revisiting change detection and captioning from a video modeling perspective. In *CVPR*, 2025. 3

Yongshuo Zhu, Lu Li, Keyan Chen, Chenyang Liu, Fugen Zhou, and Zhenwei Shi. Semantic-CC: Boosting remote sensing image change captioning via foundational knowledge and semantic guidance. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 3

Yongshuo Zhu, Lu Li, Keyan Chen, Zhenwei Shi, and Fugen Zhou. Free-CD: Probabilistically Decoupled Training-Free Open-Vocabulary Change Detection with Resolution-Invariant Feature Inversion. In *ECCV*, 2026. 2

Zhe Zhu, Shi Qiu, and Su Ye. Remote sensing of land change: A multifaceted perspective. *Remote Sensing of Environment*, 282:113266, 2022. 1

Ziqiang Zhu and Bowei Yang. UniVCD: A new method for unsupervised change detection in the open-vocabulary era. *arXiv:2512.13089*, 2025. 1, 3

# A Architecture and Implementation Details

## A.1 Image Decoder Architecture

As mentioned in Section 3.2, our framework introduces a dedicated spatial Image Decoder to transform the LLM's token-level representations into dense 2D pixel-wise predictions. Figure 4 provides an overview of this module.

The figure illustrates the complete forward pass, starting with the linear projection of the visual and textual tokens, followed by their multimodal fusion via cross-attention and their subsequent reshaping into 2D latent feature maps. We further detail the parallel multi-branch U-Net topology. It highlights how the *Relevance* and *Change* decoders operate on the temporal difference between the latent maps to produce binary masks. In parallel, the *Embedding* decoder independently processes the pre- and post-event states to generate the 64-dimensional open-vocabulary semantic embeddings. The selection of a 64-dimensional latent space rather than a more compact 32-dimensional latent space was selected based on the results of an empirical ablation study.. As detailed in Tab. 8, projecting the features into a 64D latent space retains more of the variance present in the original representations and exhibits better alignment with the embeddings produced by the CLIP text encoder, resulting in improved semantic prediction performance.

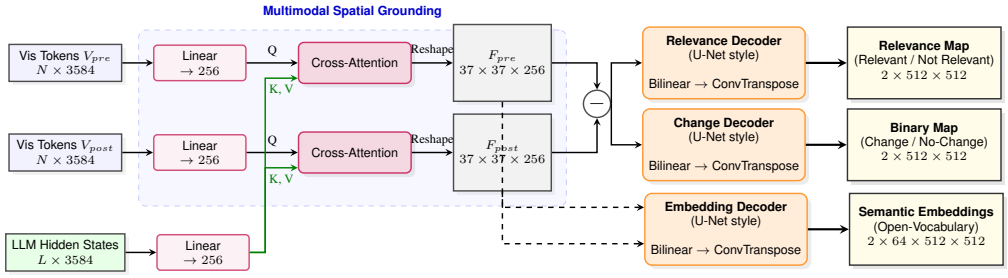



Figure 4: **Detailed Architecture of the Spatial Decoders.** Visual tokens from the *pre-* and *post-* images and textual tokens ($d_{llm} = 3584$) are projected and fused via Cross-Attention, then reshaped into 2D latent maps. The difference map feeds both the *Relevance Decoder* (query-targeted) and the *Change Decoder* (global change). Concurrently, the *Embedding Decoder* processes individual temporal states to produce high-resolution open-vocabulary embeddings.

| Latent Dimension ($D$) | mIoU (FSC) | Cosine Similarity ($\rho$) | Explained Variance (PCA) |
|---|---|---|---|
| 32 | 0.79 | 0.23 | 56% |
| **64** | **0.83** | **0.26** | **70%** |

Table 8: **Embedding space dimensionality ablation.** Comparison of spatial semantic grounding (mIoU) and embedding alignment quality (Cosine Similarity) when projecting visual features to a 32-dimensional versus a 64-dimensional latent space. The 64D space captures significantly more variance and aligns better with the fixed CLIP text embeddings.

## A.2 Spatial Loss Functions Formulation

Training our spatial decoders requires handling severe class imbalance. This issue is particularly pronounced for the *Relevance* Decoder, where the queried target object often occupies only a small fraction of the overall image. To prevent the network from collapsing into trivial background predictions, we employ a combination of Dice Loss and Focal Binary Cross-Entropy (BCE).

Let $y \in \{0, 1\}$ denote the ground-truth binary mask and $p \in [0, 1]$ the predicted probability for a given pixel.

**Dice Loss.** The Dice Loss is applied to both the global Change Decoder ($\mathcal{L}_{\text{chg}}$) and the Relevance Decoder ($\mathcal{L}_{\text{rel}}$) to measure the spatial overlap between the prediction and the ground truth, which

helps mitigate spatial class imbalance. It is defined as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum yp}{\sum y + \sum p + \epsilon}, \tag{2}$$

where the sum is over the pixels and $\epsilon$ is a small smoothing constant to prevent division by zero.

**Focal Loss for Relevance Prediction.** To strictly supervise the Relevance Decoder ($\mathcal{L}_{\text{rel}}$) against query-specific instance masks, the standard BCE loss is insufficient due to the overwhelming number of negative (background) pixels. We therefore replace it with a class-weighted Focal BCE Loss:

$$\mathcal{L}_{\text{Focal}} = -\alpha y(1-p)^{\gamma}\log(p) - (1-\alpha)(1-y)p^{\gamma}\log(1-p). \tag{3}$$

We set the focusing parameter to $\gamma = 2.0$ to down-weight the contribution of easy-to-classify background pixels. Crucially, we assign a high weight to the positive class by setting $\alpha = 0.9$. This configuration heavily penalizes any missed targeted pixels (false negatives), helping the model ground the user's specific request in the 2D space despite the highly skewed spatial distribution.

**Combined Objectives.** The final spatial loss components are defined as follows (see Equation 1 for the full loss formulation).

$$\mathcal{L}_{\text{chg}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}, \tag{4}$$

$$\mathcal{L}_{\text{rel}} = \mathcal{L}_{\text{Focal}} + \mathcal{L}_{\text{Dice}}. \tag{5}$$

### A.3 Implementation Details and Computational Costs

**Training Setup.** Our unified architecture is trained end-to-end. The Large Language Model, the Vision Encoder, and the multimodal projector (MLP2x_GELU) are initialized from the official VideoLLaMA3-7B checkpoint[1]. To ensure stable convergence, we apply differential learning rates: a base learning rate of $1 \times 10^{-5}$ is used to fine-tune the pretrained modules, while a higher learning rate of $1 \times 10^{-4}$ is applied to the newly initialized spatial Image Decoder. The full model is trained on a single computing node equipped with 4 NVIDIA H100 GPUs for approximately 40 hours, using a global batch size of 40 without gradient accumulation.

**Computational Costs.** We benchmark the inference efficiency of `LOSCAR` on a single NVIDIA H100 GPU. Operating with a batch size of 1, the model requires approximately 40 TFLOPs per image pair for a complete spatial and textual forward pass, executing in 247.1 ms. When accounting for the autoregressive generation loop, a full text sequence generation combined with spatial decoding takes an average of 875.7 ms per image pair.

## B Open-Vocabulary Dataset Construction and Mask Generation

To bridge the gap between traditional closed-set remote sensing nomenclature (*e.g.*, generic labels such as "*building*" or "*vegetation*") and the rich, descriptive capabilities required for open-ended Visual Question Answering, we redesign our dataset construction pipeline using a Large Multimodal Model (LMM). Specifically, we design a highly scalable, automated pipeline leveraging the Gemini API to extract, curate, and spatially ground an open vocabulary of highly descriptive concepts.

### B.1 LMM-Driven Concept Extraction via Gemini API

Unlike previous approaches that rely on unconstrained image captioning, we formulate the concept extraction as a constrained JSON generation task. For each mono-temporal image in our training corpus, we use the Gemini API, with images processed asynchronously via the Batch API to efficiently scale to hundreds of thousands of images. To ensure that the newly discovered open-vocabulary terms remain spatially grounded, the prompt provided to Gemini includes the list of the generic classes known to be present in the image (derived from the three original datasets). The model is instructed to analyze the visual content and to output a structured JSON file, mapping each

[1] VideoLLaMA3-7B: https://huggingface.co/DAMO-NLP-SG/VideoLLaMA3-7B

of these generic classes to a list of highly specific, context-aware descriptive terms. For example, for the generic class *building*, the API is prompted to return specific descriptors observed in the image, yielding outputs such as {`"building": ["residential house with red roof", "industrial warehouse"]`}. This explicit mapping allows each generated open-vocabulary term to be associated with a corresponding generic class and its spatial footprint.

Implementing this large-scale constrained extraction required careful logistical planning to remain cost-free and respect API rate limits. Specifically, we relied on the free tier of the `gemini-3.5-flash-lite` model. To process more than 100,000 mono-temporal images within the free tier's strict requests-per-minute constraints, the extraction script was executed continuously in the background over several days. We optimized our asynchronous calls by grouping 30 images per request, prompting the model to generate the structured JSON outputs for the entire image grid simultaneously.

While one might question why such a capable commercial Vision-Language Model (VLM) was not employed directly for the change detection task, several practical limitations make their use as the primary predictive model less suitable for our setting. Fundamentally, closed-source APIs lack the architectural flexibility required for our pipeline and rely on black-box mechanics that hinder scientific reproducibility and may limit reproducibility and access to domain-specific fine-tuning. Furthermore, these large commercial models would introduce substantial computational and financial costs for our use case.

Therefore, the Gemini API was strictly leveraged as an offline "data engine" to overcome the limitations of generic class names and enrich our training set. To support open science and foster further research, the resulting LLM-enriched open-vocabulary masks and the generated Q/A datasets will be made publicly available alongside our code.

### B.2 Vocabulary Refinement and Semantic Clustering

The raw outputs from the Gemini API form a large corpus of diverse linguistic expressions. We apply a strict curation process to transform them into a robust training target. First, we apply Natural Language Processing (NLP) normalization techniques (lowercasing, punctuation removal, and lemmatization) to eliminate syntactic noise and deduplicate identical concepts. Next, we embed all unique expressions using a pretrained CLIP text encoder. We perform semantic clustering based on cosine similarity to group synonymous or overly granular descriptions (*e.g.*, merging "small dirt path" and "unpaved dirt trail"). This refinement process successfully reduces the initial large pool of raw extractions down to a highly curated, semantically coherent vocabulary of approximately 4,000 distinct open-vocabulary concepts, while preserving their hierarchical link to the original classes.

### B.3 CLIP Embedding Alignment and Open-Vocabulary Mask Generation

The final step consists of converting the traditional semantic supervision into an open-vocabulary format compatible with our embedding-based loss ($\mathcal{L}_{\text{emb}}$, see Equation 1). First, the vocabulary of $\sim$4,000 concepts is processed through the CLIP text encoder and stored as a static dictionary of high-dimensional text embeddings. Then, for each image in the dataset, we dynamically construct *open-vocabulary semantic masks*. Rather than assigning a standard integer class ID (*e.g.*, 0 to 30) for each pixel, we replace the generic labels with the specific open-vocabulary concepts identified by Gemini for the corresponding image. Consequently, the ground-truth spatial masks are transformed into dense tensors referencing the CLIP embeddings of these specific terms. During training, the Image Decoder is supervised to project visual features into this shared text-vision latent space, optimizing the cosine similarity between the pixel-level visual embeddings and the CLIP embeddings of the target concepts. Simultaneously, these specific terms are injected into the automated question-answer templates (see Section 3.4), promoting alignment between the rich textual descriptions generated by the language model and the precise spatial targets predicted by the vision decoder.

### B.4 QUESTION-ANSWERING GENERATION PIPELINE

Training LOSCAR requires converting raw semantic maps into structured visual question-answering (VQA) pairs. As illustrated in Fig. 5, our automated QA generation pipeline systematically extracts spatial and semantic attributes from pre- and post-change labels provided by the dataset.

1. Intersecting the two maps allows us to retrieve both the spatial and semantic information of changed areas (*e.g.*, grass → warehouse) as well as stable areas (*e.g.*, unchanged ponds).
2. We decompose these areas into individual objects (instances) using a SLIC-based segmentation (Achanta et al., 2012).
3. The spatial and semantic information is then injected into a "Template Bank". This bank contains various query structures, including standard change descriptions, spatial extent inquiries, and hard negatives to mitigate model hallucinations. The six categories are detailed in Fig. 5.

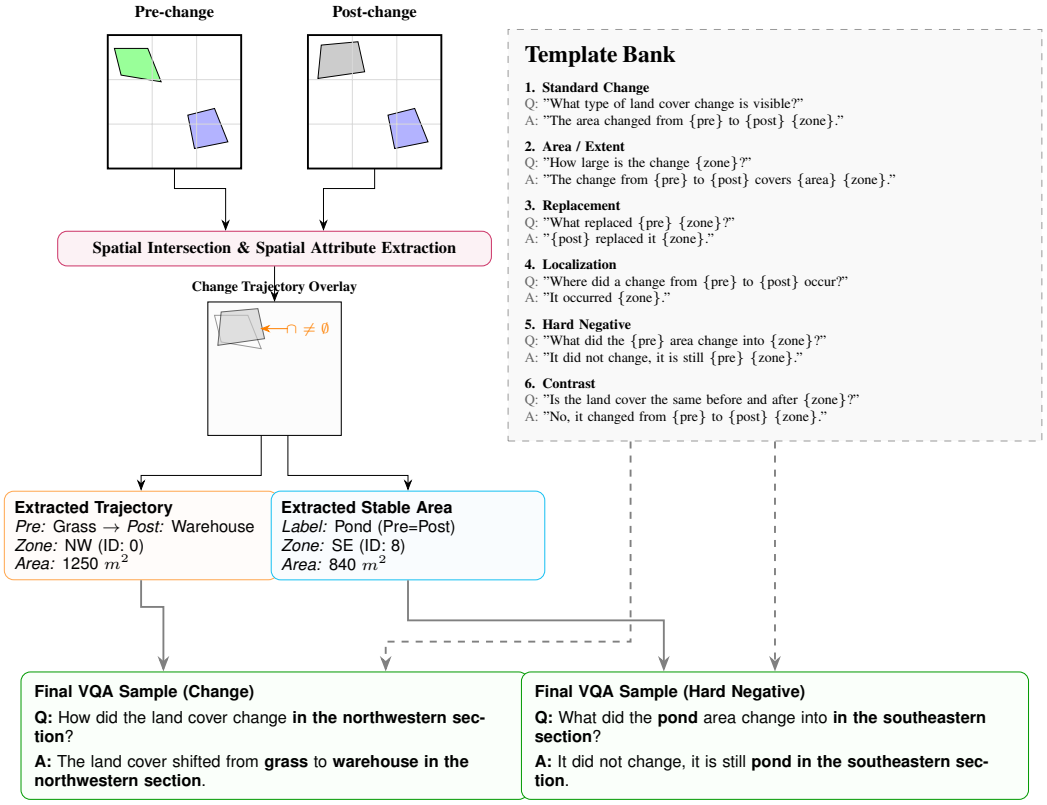
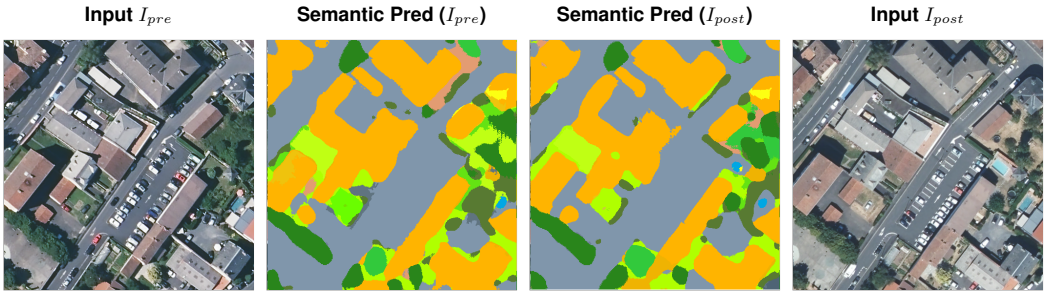



Figure 5: Overview of our automated Question-Answering (QA) generation pipeline.

## C QUALITATIVE ASSESSMENT

Fig. 3 illustrates the performance of LOSCAR on both dense semantic segmentation and localized CDVQA. The predicted semantic masks show detailed spatial delineation across a wide variety of classes and intra-class variations. Fine-grained segmentation is obtained by refining the raw dense predictions using SLIC superpixels, majority voting, and frequency-based weighting, which favors more frequent vocabulary terms. The model also grounds localized text prompts, enabling localized change captioning for specific areas.

Fig. 6 highlights our model's versatility and robustness in complex environments. Notably, the model accurately answers standard, non-change VQA queries, such as counting small objects. This qualitative example suggests that our anti-forgetting strategy successfully preserves general VQA capabilities, reducing the tendency to rely on rigid, formatted change-captioning templates. Simul-




**User**: Can you count the swimming pools on each image?
**LOSCAR**: In the first image, there is 1 swimming pool. In the second picture, we can see 2 pools: one rectangular and another round-shaped.

**User**: What was present before it became swimming pool in the middle-right part of the image?
**LOSCAR**: Grass was present before.

Figure 6: Multimodal understanding and spatial reasoning in complex urban scenes. The model performs standard VQA (counting) without collapsing into rigid change-captioning templates, while also accurately grounding and describing prior states in trajectory-based queries (color code in Fig. 3).

taneously, the model demonstrates strong RCD capabilities by identifying the prior state of a specific transitioned area based on a trajectory query.

Beyond image-level CD, LOSCAR isolates specific transitions based on the user's query (Fig. 7). While the standard binary change mask captures all modified areas in the scene (*e.g.*, both new buildings and new parking lots), the predicted relevance mask aligns with the user's targeted query. By precisely masking only the requested transition and ignoring unrelated changes, the model is able to perform query-conditioned spatial grounding.




**User**: What replaced the tree in the upper central area?
**LOSCAR**: It has been replaced by a building.

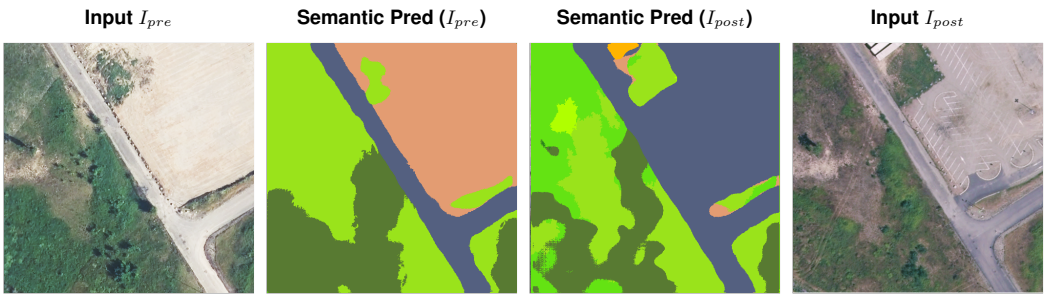
Figure 7: Spatial visual grounding using the relevance mask. Unlike a global binary change mask that highlights all scene modifications, the predicted relevance mask isolates the specific transition (tree $\rightarrow$ building), targeted by the user's query.

Finally, Fig. 8 illustrates the extension of RCD into an open-vocabulary setting. Moving beyond a predefined closed set of classes, the model successfully localizes open-vocabulary concepts by computing similarity scores between dense spatial embeddings and text queries, using a proportional thresholding mechanism. The model successfully localizes small instances like swimming pools and performs well on gardens. While parking areas are generally well delineated, we observe occasional leakage into visually and texturally continuous roads, alongside minor boundary deviations near building walls. These artifacts underscore the difficulty of localizing concepts on visually continuous man-made surfaces, while also illustrating the model's strong generalization capabilities.




Figure 8: Open-Vocabulary Change Detection (OVCD) capabilities. The model dynamically localizes open-vocabulary concepts across the input images by thresholding similarity scores between dense spatial embeddings and textual queries (Fig. 4).

To further characterize the behavior of LOSCAR, Fig. 9 and 10 present qualitative results across diverse conversational scenarios and image resolutions. Fig. 9 highlights complex multimodal interactions, including temporal ordering, detailed land-cover change descriptions, and localized stable-state identification (hard negatives). Moreover, Fig. 10 demonstrates the model's ability to summarize large-scale macroscopic transitions. These examples feature significantly different spatial resolutions, ranging from Very High Resolution imagery (Fig. 9) to the lower-resolution inputs typical from the SECOND-CC dataset (Fig. 10). This contrast illustrates the ability of our multi-dataset training strategy, allowing the model to consistently consistent semantics and reason effectively across varying scales.

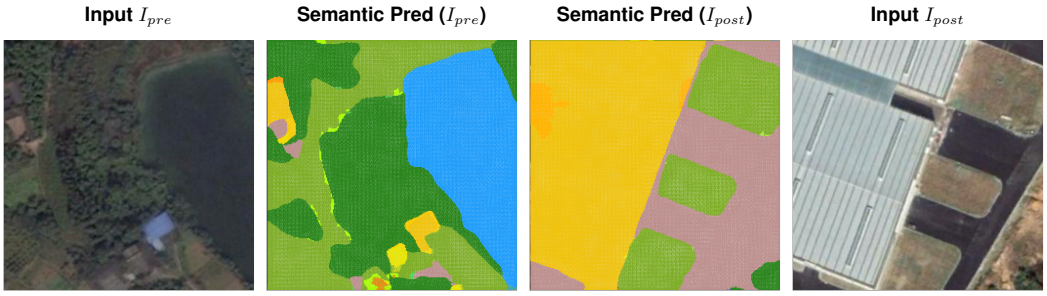



**User**: Which one of the images is the most recent?
**LOSCAR**: Image $I_{post}$ is the most recent.

**User**: Use language to describe the changes between two images.
**LOSCAR**: The large field of dirt has been replaced by parking lot. Grassy area replaced bare soil in the upper central area.

**User**: Has the land cover of the left zone changed between the two images?
**LOSCAR**: No, it is still tree in the left zone.

Figure 9: Complex multimodal tasks and localized question answering on VHR imagery (HiUCD dataset). The color code is provided in Fig. 3.




**User**: How has the area changed?
**LOSCAR**: A huge industrial zone was built in place of the green vegetation area and the pond.

Figure 10: Large-scale change summarization on lower-resolution imagery (SECOND-CC dataset). The color code is provided in Fig. 3.