

UniChange: Unifying Change Detection with Multimodal Large Language Model

Xu Zhang^{1*}

Danyang Li^{2*}

Xiaohang Dong¹

Tianhao Wu⁴

Hualong Yu¹

Jianye Wang¹

Qicheng Li^{1†}

Xiang Li^{2,3}

¹TMCC, Computer Science, Nankai University

²VCIP, Computer Science, Nankai University

³NKIARI, Futian, Shenzhen

⁴CMEE, Sichuan Agricultural University

{xu_zhang, danyang.li}@mail.nankai.edu.cn, liqicheng@nankai.edu.cn

Abstract

Change detection (CD) is a fundamental task for monitoring and analysing land cover dynamics. While recent high performance models and high quality datasets have significantly advanced the field, a critical limitation persists. Current models typically acquire limited knowledge from single-type annotated data and cannot concurrently leverage diverse binary change detection (BCD) and semantic change detection (SCD) datasets. This constraint leads to poor generalisation and limited versatility. The recent advancements in Multimodal Large Language Models (MLLMs) introduce new possibilities for a unified CD framework. We leverage the language priors and unification capabilities of MLLMs to develop UniChange, the first MLLM-based unified change detection model. UniChange integrates generative language abilities with specialised CD functionalities. We introduce three special tokens: [T1], [T2], and [CHANGE], utilising their embeddings as the key to query variations. This approach successfully accommodates both BCD and SCD tasks. Furthermore, UniChange utilises text prompts to guide the identification of change categories, eliminating the reliance on predefined classification heads. This design allows UniChange to effectively acquire knowledge from multi-source datasets, even when their class definitions conflict. Experiments on four public benchmarks (WHU-CD, S2Looking, LEVIR-CD+, and SECOND) demonstrate SOTA performance, achieving IoU scores of 90.41, 53.04, 78.87, and 57.62, respectively, surpassing all previous methods. The code is available at <https://github.com/Erxucomeon/UniChange>.

*Equal Contribution

†Corresponding Author

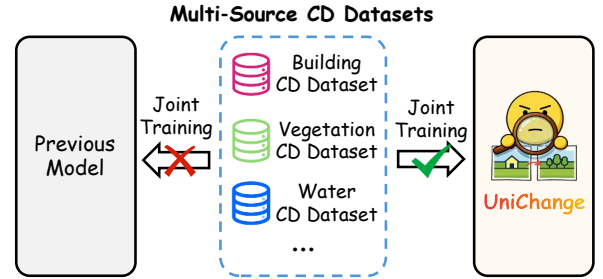


Figure 1. Inconsistency in multi-source change detection datasets.

1. Introduction

Change detection (CD) is the process of observing and analysing multi-temporal remote sensing images to identify changes in land surface cover. It serves as a fundamental task in the remote sensing field and a cornerstone of modern geospatial analysis. Change detection plays a vital role in numerous applications, ranging from sustainable urban planning [48] and natural disaster assessment [58] to ecological monitoring [44] and land resource management [1]. Depending on the required level of detail, CD tasks are typically categorised into two primary sub-domains: binary change detection (BCD) and semantic change detection (SCD). BCD, the simpler task, primarily identifies the locations where change has occurred, but may focus on a single object of interest or not specify the nature of these transformations. While useful for automated monitoring, it provides relatively coarse-grained information. In contrast, SCD is a more complex and refined task. It not only determines the location of change but also identifies the “from-to” semantic transition [37] (e.g., from “forest” to “urban land”). This granular understanding of semantic shifts offers richer insights, providing more robust decision support for the aforementioned applications.

In the deep learning era, methods such as FC-Siam-diff [10] and IFN [53] establish the siamese network architecture as the dominant paradigm for BCD. Subsequent research shifts focus toward enhancing the fusion and interaction of dual-temporal features. BiT [4] fuses dual-temporal features using token-space context modeling and feature refinement. Concurrently, Transformer-based architectures [2, 50, 54] are proposed to fuse and model long-range dependencies between these features at various stages within the encoder. Changer [18] pioneers a dual-temporal feature exchange paradigm by inserting a series of feature interaction layers directly within the feature extractor. Furthermore, recent works like SAM-CD [13] have begun to adapt Vision Foundation Models (VFM) [29, 41, 57] for the BCD task. However, a fundamental limitation of these methods is the necessity of training a specialised model for each individual dataset. RSBuilding [46] partially mitigates this issue by leveraging a VFM, enabling joint training on multiple building datasets for both building extraction and change detection. However, its framework is essentially tailored to building-related tasks and cannot be extended to other categories of land cover change.

Compared to BCD, SCD is considered a more challenging task. Its core complexity lies in the fact that SCD must not only localise changes but also identify specific “from-to” semantic transitions between land cover categories. To address this challenge, the early HRSCD [11] adopts a decoupling strategy, jointly training semantic segmentation and binary change detection, which is subsequently refined by MTSCD-Net [8]. Subsequent studies have turned to deeper feature interaction and temporal modelling. BiSRNet [12] improves semantic change detection via dual-temporal reasoning and consistency modelling. Furthermore, SCD-SAM [37] leverages the representational capabilities of VFMs and designs a refined dual-encoder and dual-decoder architecture. The SCD task’s inherent specificity and complexity necessitate a specialised architecture, thereby hindering architectural unification between BCD and SCD tasks.

By analysing the previous BCD and SCD methods, we have identified two problems that have persisted in the field. The first is dataset incompatibility. As shown in Fig. 1, a positive sample in one dataset (e.g., “building”) may be explicitly defined as a negative background sample in another dataset (e.g., “vegetation”). This semantic conflict prevents traditional models from being jointly trained on such multi-source datasets. Secondly, the different requirements of BCD and SCD lead to incompatible model architectures, as shown in Fig. 2(a) and Fig. 2(b). This has led to a proliferation of highly specialised models, each with limited knowledge from a single annotation type of dataset (BCD or SCD). This mutually independent design approach results in poor model generalisation and severely limits the

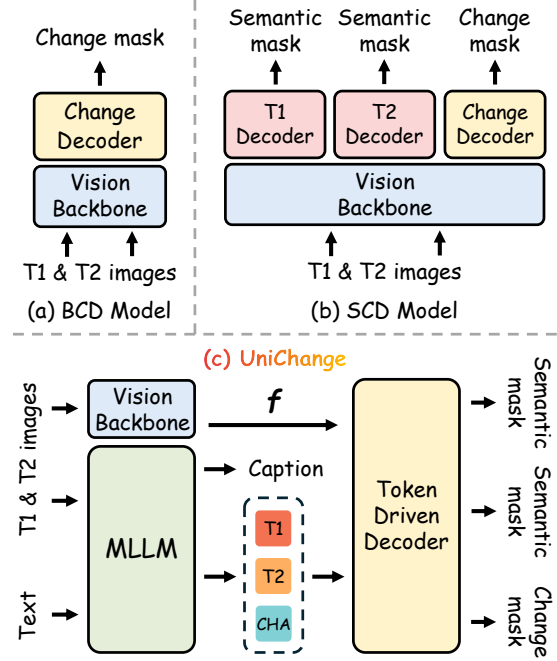


Figure 2. Inconsistency in model architecture.

versatility of the CD system in practical applications.

To address the fundamental challenges within the change detection domain, we propose UniChange. UniChange is a unified change detection model based on the multimodal large language model. Not only does UniChange simultaneously support both BCD and SCD tasks, but it also enables joint training using multi-source datasets. The architecture of UniChange is shown in Fig. 2(c). We introduce three special tokens [T1], [T2], and [CHANGE] to enable the model to handle both BCD and SCD tasks concurrently. Furthermore, UniChange leverages textual cues to guide change region classification, thereby eliminating the reliance on pre-defined classification heads found in traditional approaches. This design facilitates the extraction of unified knowledge from diverse, multi-source change detection datasets, even when semantic category conflicts exist between datasets. Overall, our contributions are as follows:

- We propose UniChange, the first unified framework for change detection based on the MLLM. It is designed to simultaneously accommodate both binary change detection (BCD) and semantic change detection (SCD) tasks within a single end-to-end model, whilst enabling training using multi-source change detection datasets.
- We introduce a Token Driven Decoder strategy. The strategy incorporates three special tokens into MLLM: [T1], [T2] and [CHANGE]. It enables the model to be trained simultaneously on change detection tasks and the semantic understanding of dual-temporal images.
- Comprehensive experiments demonstrate that UniChange

achieves state-of-the-art performance, significantly outperforming previous methods on both the BCD and SCD benchmark datasets.

2. Related Work

2.1. Change Detection

The field of change detection (CD) is broadly divided into two sub-fields: binary change detection (BCD) and the more complex semantic change detection (SCD). The latter is more challenging as it requires identifying the “from-to” semantic information, not just the presence of change.

Early deep learning methods for BCD are dominated by CNN-based architectures. FC-Siam-diff [10] is one of the early methods to utilise the paradigm of comparing features from two weight-sharing encoders. IFN [53] extends this by adding deep supervision and attention mechanisms. To better model global context, Transformer-based methods are introduced, including BiT [4], ChangeFormer [2], Swin-SUNet [54] and Changer [18]. Other notable works, such as DMINet [19], and HATNet [50], further explore hybrid architectures and deep feature interaction. More recently, VFM-based methods have emerged. SAM-CD [13] focuses on adapting VFMs [29, 41, 57] to the BCD task, while RS-Building [46] leverages VFMs to unify building-specific extraction and change detection tasks. However, these methods are difficult to jointly train on diverse and multi-class change detection datasets, limiting their generalisation.

Compared with BCD task, SCD task is more complex. HRSCD [11] and MTSCD-Net [8] decouple the task into binary change detection and semantic segmentation sub-tasks, integrating them within a multi-task framework for refined feature association. SCDNet [38] employs siamese networks and further utilises attention mechanisms to refine multi-scale difference features, while Bi-SRNet [12] proposes cross-temporal semantic reasoning blocks to enhance feature consistency. Most recently, VFM-based adaptation emerges. SCD-SAM [37] adapts VFMs by designing a specialised dual-encoder and dual-decoder system specifically for the SCD task. The differences between BCD and SCD tasks lead to inconsistency between their model architectures, resulting in poor versatility.

2.2. Multimodal Large Language Model

Initial MLLMs, such as LLaVA [34], MiniGPT-4 [60], and InstructBLIP [9], focus on tasks such as image captioning and visual question answering (VQA). Subsequent research has further developed this capability, enabling it to support refined visual understanding, perception and generation at both regional and pixel levels [31, 33, 39, 49, 52, 56]. While transformative, their application to remote sensing (RS) has been challenging. RS-specific MLLMs such as RSGPT [26], GeoChat [30], and GeoPixel [42] have emerged, but

these methods largely focus on single-image interpretation. This architecture, which focuses on single images, makes it ill-suited for comparative, dual-temporal analysis. It leaves a significant gap in the ability to perform pixel-level grounding for change detection.

3. Method

In this section, we first define the unified change detection task. As detailed in Sec. 3.1, this task involves dual-temporal remote sensing images and textual queries. Subsequently, we present the UniChange architecture in detail in Sec. 3.2. This architecture comprises an MLLM for understanding changes, a Vision Backbone for extracting features from dual-temporal remote sensing images, and a specialised Token Driven Decoder. Finally, we detail the training strategy in Sec. 3.3, including the unified loss functions designed to manage both binary and semantic constraints.

3.1. Task Definition

We define the task of unified change detection as follows. Let the input consist of a pair of dual-temporal remote sensing images, $x_{img1}, x_{img2} \in R^{H \times W \times C}$, captured at times T_1 and T_2 , where H , W , and C represent the height, width, and channel dimensions, respectively. An accompanying text input, x_{txt} , provides the change query instruction. This instruction may be a binary change query (e.g., “Please segment all areas that have undergone change.”) or a more complex semantic change query (e.g., “Please segment the semantic masks of the changed areas.”). The core goal is to generate a set of masks ($\hat{M}$) that accurately adhere to the semantic instructions provided by the query x_{txt} . This overall process can be formulated as

$$\hat{M} = \Phi_{UniChange}(x_{img1}, x_{img2}, x_{txt}), \quad (1)$$

where $\Phi_{UniChange}$ represents the entire unified change detection model.

3.2. Architecture

Embedding as Change. Current change detection (CD) models suffer from two fundamental problems: dataset incompatibility due to semantic conflicts across different sources, and architectural divergence between specialised binary change detection (BCD) and semantic change detection (SCD) tasks. This fragmentation prevents joint training, leading to poor generalisation and limited versatility. Therefore, we propose UniChange. This framework implements our proposed “embedding as change” paradigm. This paradigm integrates generative language abilities with specialised CD functionalities, thereby achieving a true unification of multi-source datasets and multi-tasks. The framework of our method is illustrated in Fig. 3(a). MLLM takes the dual-temporal remote sensing images x_{img1}, x_{img2}

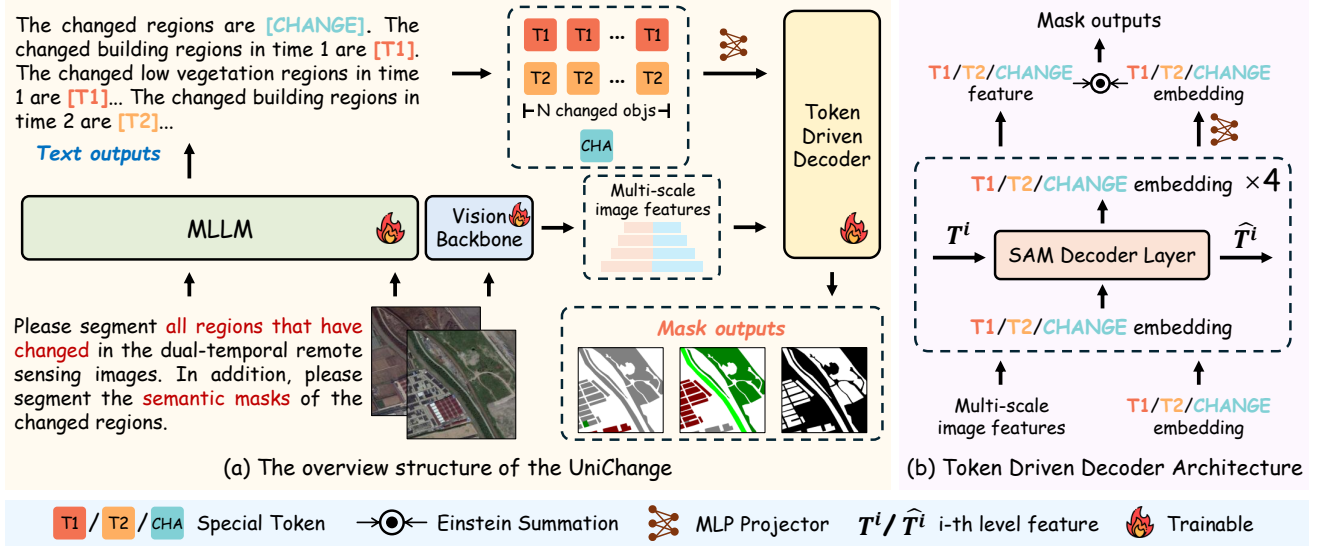


Figure 3. The overview structure of UniChange is shown in (a). UniChange generates text outputs based on text instructions and dual-temporal remote sensing images. The text outputs incorporate a series of special tokens ([T1], [T2], and [CHANGE]). Concurrently, the dual-temporal remote sensing images undergo feature extraction via the Vision Backbone. These features, alongside the embeddings corresponding to the special tokens, are fed into the Token Driven Decoder. Ultimately, UniChange generates the masks corresponding to the text instructions. The overall structure of the Token Driven Decoder is shown in (b). It receives dual-temporal remote sensing image features and special token embeddings, generating corresponding masks.

and a given textual instruction x_{txt} as input. To unify both binary and semantic change detection within a single framework, we introduce three specialised tokens into the MLLM’s vocabulary: [T1], [T2] and [CHANGE]. Conditioned on the textual instruction x_{txt} , the MLLM autoregressively generates the response sequence y_{txt} . These tokens are strategically placed within this sequence in accordance with the specific task query. This process can be formulated as

$$y_{txt} = \mathcal{F}_{MLLM}(x_{img1}, x_{img2}, x_{txt}). \quad (2)$$

When the MLLM intends to generate masks for a specific change (e.g., the binary change mask or a semantic mask for T_1 and T_2), its output y_{txt} includes the corresponding special token (e.g., [T1], [T2] or [CHANGE]). Subsequently, we extract the MLLM last-layer embedding h_{task} corresponding to the position of a specific token ([T1], [T2] or [CHANGE]) within y_{txt} . This raw embedding is then projected by a dedicated $MLP(\cdot)$ layer to make it compatible with the downstream vision module’s feature space:

$$\hat{h}_{task} = MLP(h_{task}). \quad (3)$$

These projected sparse embeddings (where each is individually generated as $\hat{h}_{task}$ via Eq. (3)) serve as the dynamic, instruction-guided queries.

Concurrently, the vision backbone $\mathcal{F}_{enc}$ first extracts base dense visual features from the dual-temporal remote

sensing images. A feature pyramid is then constructed using max-pooling and transposed convolutional layers. It generates four levels of multi-scale image features for each temporal image. This results in two multi-scale image feature sets, denoted as $\{F_1^i\}$ and $\{F_2^i\}$, where $i \in \{1, 2, 3, 4\}$. Finally, these dual-temporal visual features, along with the projected task embeddings $\hat{h}_{task}$, are fed into the token driven decoder $\mathcal{F}_{dec}$ to generate the final mask $\hat{M}_{task}$. The entire process can be described as

$$\begin{aligned} \{F_1^i\}, \{F_2^i\} &= \mathcal{F}_{enc}(x_{img1}, x_{img2}), \\ \hat{M}_{task} &= \mathcal{F}_{dec}(\{F_1^i\}, \{F_2^i\}, \hat{h}_{task}). \end{aligned} \quad (4)$$

We employ a hybrid fine-tuning strategy. Only the language decoder of the MLLM undergoes LoRA fine-tuning [25]. All other components of the model are fully parameterised for fine-tuning.

Token Driven Decoder. Inspired by RSBuilding, the token driven decoder, $\mathcal{F}_{dec}$, is specifically architected to translate the MLLM’s high-level instructions into pixel-level segmentation masks for the change detection task. Its architecture is illustrated in Fig. 3(b). This decoder is employed to process the dual-temporal remote sensing image features and the projected special token embeddings. Specifically, the projected embeddings $\{\hat{h}_{t1}, \hat{h}_{t2}, \hat{h}_{change}\}$ (obtained from Eq. (3)) serve as the initial task queries. Let E^0 represent the concatenation (Φ_{Cat}) of these initial queries:

$$E^0 = \Phi_{Cat}(\hat{h}_{t1}, \hat{h}_{t2}, \hat{h}_{change}), \quad (5)$$

the concatenated task query E^0 is then hierarchically refined through four decoder layers. Each decoder layer processes its corresponding multi-scale image features F_1^i and F_2^i by flattening (Φ_{Flat}) them into token sequence, and concatenating (Φ_{Cat}) them into a unified visual sequence T^i :

$$T^i = \Phi_{Cat}(\Phi_{Flat}(F_1^i), \Phi_{Flat}(F_2^i)). \quad (6)$$

The SAM Decoder Layer then refines both the task query and the visual sequence. The task query E^{i-1} (from the previous level) first attends to itself ($\mathcal{A}_{self}$), then attends to the visual sequence T^i ($\mathcal{A}_{cross}$), and is subsequently processed by an FFN ($\mathcal{F}_{FFN}$). Next, the visual sequence T^i is also updated by attending to this newly refined task query ($\mathcal{A}_{cross}$). The positional encodings are added to queries and keys in all attention operations. This process is defined as

$$\begin{aligned} E^i &= \mathcal{F}_{FFN}(\mathcal{A}_{cross}(\mathcal{A}_{self}(E^{i-1}), T^i)), \\ \hat{T}^i &= \mathcal{A}_{cross}(T^i, E^i). \end{aligned} \quad (7)$$

The refined task query E^4 is taken from the final level. The query and the refined visual sequences $\{\hat{T}^i\}$ from all levels are then used for the final mask generation.

The visual sequences $\{\hat{T}^i\}$ from each level are split and reshaped ($\Phi_{Split/Reshape}$) back into their corresponding 2D dual-temporal feature maps, $\{\hat{F}_1^i\}$ and $\{\hat{F}_2^i\}$:

$$\{\hat{F}_1^i, \hat{F}_2^i\} = \Phi_{Split/Reshape}(\{\hat{T}^i\}). \quad (8)$$

Next, three feature sets are prepared: $\{\hat{F}_1^i\}$, $\{\hat{F}_2^i\}$, and their element-wise difference $\{\hat{F}_1^i - \hat{F}_2^i\}$. All three sets are then uniformly processed by a three-step procedure: up-sampling (Φ_{Up}), concatenation (Φ_{Cat}), and fusion (Φ_{fuse}). This derives F_{t1} , F_{t2} , and F_{change} , respectively:

$$\begin{aligned} F_{t1} &= \Phi_{fuse}(\Phi_{Cat}(\{\Phi_{Up}(\hat{F}_1^i)\})), \\ F_{t2} &= \Phi_{fuse}(\Phi_{Cat}(\{\Phi_{Up}(\hat{F}_2^i)\})), \\ F_{change} &= \Phi_{fuse}(\Phi_{Cat}(\{\Phi_{Up}(\hat{F}_1^i - \hat{F}_2^i)\})). \end{aligned} \quad (9)$$

Concurrently, the refined task query E^4 from the final decoder level is split and projected ($\Phi_{Split/Proj}$) into the individual task embeddings $\hat{e}_{t1}$, $\hat{e}_{t2}$, and $\hat{e}_{change}$:

$$\hat{e}_{t1}, \hat{e}_{t2}, \hat{e}_{change} = \Phi_{Split/Proj}(E^4). \quad (10)$$

Finally, these projected embeddings are used to filter their corresponding features (F_{t1} , F_{t2} , F_{change}) to generate the final masks:

$$\begin{aligned} \hat{M}_{task} &= \mathcal{M}_{gen}(F_{task}, \hat{e}_{task}), \\ task &\in \{t1, t2, change\} \end{aligned} \quad (11)$$

where $\mathcal{M}_{gen}$ represents the mask generation function. This function performs Einstein Summation.

This unified structure allows UniChange to flexibly generate the corresponding masks based on user instructions.

3.3. Training Strategy

Our method is trained end-to-end to jointly optimise UniChange. The overall objective function, $\mathcal{L}_{total}$, which guides the entire optimisation process, is defined as a summation of all contributing loss terms:

$$\mathcal{L}_{total} = \mathcal{L}_{txt} + \mathcal{L}_{mask}. \quad (12)$$

In Eq. (12), $\mathcal{L}_{txt}$ is the standard autoregressive Cross-Entropy loss applied to the MLMM’s generated token sequence, optimising its next-token prediction capability.

The comprehensive mask loss, $\mathcal{L}_{mask}$, is designed to enforce pixel-level accuracy across the unified change detection task. This loss is defined as the summation of four distinct components:

$$\mathcal{L}_{mask} = \lambda_{BCE}\mathcal{L}_{BCE} + \lambda_{Dice}\mathcal{L}_{Dice} + \lambda_{SS}\mathcal{L}_{SS} + \lambda_{SC}\mathcal{L}_{SC}. \quad (13)$$

In Eq. (13), λ_{BCE} , λ_{Dice} , λ_{SS} , and λ_{SC} serve as the weighting coefficients for the Binary Cross-Entropy loss, Dice loss, Semantic Segmentation loss, and Semantic Change loss, respectively. $\mathcal{L}_{BCE}$ and $\mathcal{L}_{Dice}$ are the Binary Cross-Entropy loss and the Dice loss, respectively. Both of them optimise the model’s prediction by applying supervision between the predicted change masks and the ground truth change masks. $\mathcal{L}_{SS}$ is the Cross-Entropy loss for Semantic Segmentation, which computes pixel-wise classification errors across all spatial locations between the dual-temporal predictions and their respective ground truth labels, thereby penalizing misclassification across all semantic categories. $\mathcal{L}_{SC}$ is the Semantic Change loss, designed to enhance the discriminability of dual-temporal feature representations. This loss calculates a cosine embedding distance between the semantic feature maps based on the binary ground truth masks. Specifically, it enforces similarity for unchanged regions and encourages divergence for changed regions.

Crucially, during training on BCD datasets, the semantic constraint losses, $\mathcal{L}_{SS}$ and $\mathcal{L}_{SC}$, are set to zero. They are only computed when training on SCD datasets. By optimising this comprehensive objective function, UniChange ensures both robust and effective alignment between textual instructions and pixel-level perception, simultaneously achieving high accuracy in both binary and multi-class semantic change detection tasks.

4. Experiments

4.1. Experimental Setting

We employ LLaVA-7B-v1-1 [34] as the base MLMM. For Vision Backbone, we adopt the RSBUILDING-ViT-L [46] backbone, which is a SAM backbone structure pre-trained on remote sensing imagery.

UniChange is trained for 10 epochs (with 400 steps per epoch) using 4 NVIDIA 80G H100 GPUs. We employ the AdamW [36] optimizer with a base learning rate of 5×10^{-5} . The training is performed using a per-device batch size of 1 and a gradient accumulation step of 8, leveraging the deepspeed [40] engine for efficiency.

The comprehensive loss function, which is introduced in Sec. 3.3, utilises several weighting coefficients to balance its multiple objective components. Specifically, the mask loss components are weighted with λ_{BCE} set to 2.0, λ_{Dice} set to 0.5, λ_{SS} set to 0.5, and λ_{SC} set to 1.0.

4.2. Datasets

BCD Datasets. We evaluate the binary change detection (BCD) capability of UniChange on three widely adopted remote sensing datasets: WHU-CD [28], S2Looking [43], and LEVIR-CD+ [3]. The WHU-CD dataset consists of a pair of large dual-temporal remote sensing images, each with a size of 32507×15354 . For standardised evaluation, we crop these large images into 1024×1024 patches. We then split these patches into training, validation, and testing sets following an 8:1:1 ratio. The S2Looking dataset provides a total of 5000 image pairs captured from side-looking rural areas globally. These image pairs are divided into 3500 for training, 1000 for testing and 500 for validation. Finally, LEVIR-CD+ is an extended version of LEVIR-CD. It provides 985 pairs of 1024×1024 images, primarily focusing on building changes. We utilise the original 1024×1024 image size for training and testing, which provides 637 training patches and 348 testing patches. These datasets collectively allow for robust assessment across different resolutions and geographic regions.

SCD Datasets. We evaluate the semantic change detection (SCD) capability of UniChange on the SECOND dataset [51]. The dataset is specifically designed for fine-grained change analysis, requiring models to simultaneously localise changes and classify semantic transitions. The dataset comprises 4662 pairs of dual-temporal remote sensing images with a 512×512 resolution. It covers six detailed land-cover categories: building, low vegetation, tree, water, playground, and bare ground. The predominance of unchanged pixels in the dataset creates a significant class imbalance, which places higher demands on the model’s generalisation capability.

4.3. Evaluation Metrics

BCD Metrics. To evaluate BCD performance, we employ four standard pixel-level metrics: Precision (P), Recall (R), F1 score (F1), and Intersection over Union (IoU). Precision measures the ratio of correctly predicted change pixels to all predicted change pixels. Recall measures the ratio of true positive pixels to all ground-truth positive pixels. The F1 score is the harmonic mean of P and R. IoU quantifies the

overlap between predicted and ground-truth change regions.

SCD Metrics. To evaluate SCD performance, we employ three metrics: mIoU, F_{scd} , and SeK. mIoU evaluates the overall segmentation quality. F_{scd} quantifies semantic transition accuracy using the harmonic mean of precision and recall. SeK measures semantic discrimination, mitigating the impact of the unchanged class. UniChange’s SCD performance is comprehensively evaluated using BCD metrics (IoU, F1 (F_{bcd})) and SCD metrics (mIoU, F_{scd} , SeK). Further details on the BCD and SCD metrics are provided in Sec. A of the supplementary materials.

4.4. Results

BCD Results. As shown in Tab. 1, UniChange achieves the best F1 score and IoU score across all three evaluated benchmarks: WHU-CD, S2Looking, and LEVIR-CD+. On the WHU-CD dataset, our method achieves top F1 and IoU scores of 94.96 and 90.41. On the S2Looking dataset, our method obtains an F1 score of 69.32 and an IoU score of 53.04, exceeding the LSKNet result (67.52 F1 score, 50.96 IoU score). Furthermore, on the LEVIR-CD+ dataset, our approach demonstrates superior performance, setting a new SOTA F1 score of 88.19 and IoU score of 78.87, outperforming the second-best model, SFCD-Net (86.44 F1 score, 76.12 IoU score). This consistent superiority in F1 and IoU scores across diverse benchmarks highlights UniChange’s effectiveness and broad applicability.

SCD Results. As shown in Tab. 2, UniChange achieves the highest IoU score, mIoU score and F_{bcd} score, confirming its superior capability in accurately delineating the overall regions of change and non-change. Furthermore, the model’s SOTA performance on F_{scd} and SeK metrics verifies that UniChange’s semantic change perception capability is equally outstanding. This dual excellence across both binary and semantic accuracy capabilities exemplifies the unique, robust unification achieved by our MLLM-based framework, setting it apart from prior specialised models. For instance, MambaSCD, while achieving a strong binary detection result (mIoU 72.73, F_{bcd} 72.81), exhibits a substantial drop in semantic accuracy (F_{scd} 54.17). Similarly, HGINet, despite attaining a competitive semantic score (F_{scd} 62.88, the second highest), shows weaker binary detection performance (mIoU 71.73, F_{bcd} 71.90). This comprehensive superiority across all metrics confirms that UniChange’s architecture is highly effective for handling the challenging task of multi-class change detection.

4.5. Ablation Study

In this section, we aim to validate the efficacy of the proposed components and strategies by conducting a series of ablation experiments. Unless otherwise specified, we employ LLaVA-7B-v1-1 as the base MLMM and RSBuilding-ViT-L as our vision backbone. Furthermore, semantic su-

Table 1. Comparison with existing SOTA methods on WHU-CD, S2Looking and LEVIR-CD+.

Methods	WHU-CD				S2Looking				LEVIR-CD+			
	P	R	F1	IoU	P	R	F1	IoU	P	R	F1	IoU
FC-Siam-diff [10]	48.84	88.96	63.06	46.05	83.49	32.32	46.60	30.38	80.88	77.65	79.23	65.60
DASNet [6]	83.77	91.02	87.24	77.37	45.06	48.71	47.29	30.97	77.51	78.03	77.77	63.63
SNUNet [17]	91.28	87.25	89.22	80.54	45.25	50.60	47.78	31.39	78.90	78.23	78.56	64.69
BIT [4]	91.56	87.84	89.66	81.26	70.26	56.53	62.65	45.61	82.37	79.73	81.03	68.11
ChangeFormer [2]	91.76	84.85	88.17	78.85	70.73	49.25	58.07	40.91	84.29	80.81	82.51	70.23
RFL-CDNet [20]	91.33	91.46	91.39	84.15	65.72	60.82	63.17	46.17	79.95	84.04	81.94	69.41
SAM-CD [13]	96.87	85.67	90.92	83.35	72.80	58.92	65.13	48.29	79.71	81.35	81.96	69.43
Meta-CD [22]	89.00	90.35	89.67	81.27	74.08	54.00	62.47	45.42	80.17	84.13	82.10	69.93
SA-CDNet [21]	95.29	93.67	94.47	89.52	81.28	56.24	66.48	49.79	85.55	83.44	84.43	73.06
ChangeCLIP [16]	96.02	93.58	94.78	90.08	-	-	-	-	88.46	83.90	86.12	75.63
Changer [18]	-	-	-	-	73.01	62.04	67.08	50.47	-	-	-	-
LSKNet [32]	-	-	-	-	71.90	63.64	67.52	50.96	-	-	-	-
SFCD-Net [55]	95.42	93.14	94.27	89.16	72.60	61.70	66.71	50.05	87.25	85.65	86.44	76.12
TTP [7]	96.05	92.76	94.37	89.34	73.51	62.19	67.38	50.80	85.81	84.36	85.08	74.03
UniChange	95.83	94.11	94.96	90.41	73.48	65.60	69.32	53.04	87.88	88.49	88.19	78.87

Table 2. Comparison with existing SOTA methods on SECOND. * denotes re-implementation under identical settings.

Methods	IoU	mIoU	F _{scd}	F _{bcd}	SeK
HRSCD.Str4 [11]	53.34	69.44	-	69.57	15.97
ChangeMask [59]	54.23	-	-	70.32	17.89
SAAN [23]	53.49	-	-	-	18.03
DESNet [47]	-	70.73	58.75	-	17.97
STSP-Net [24]	-	72.03	60.77	72.05	20.91
DFINet [45]	-	72.61	-	-	20.12
MLFA-Net [14]	56.33	72.45	-	72.06	20.11
TextSCD [27]	-	72.38	61.90	-	21.66
GLAI-Net [15]	56.21	-	-	71.97	20.63
Bi-SRNet* [12]	56.89	72.25	62.12	72.52	21.69
SCDNet* [38]	55.51	71.63	62.40	71.39	21.43
MTSCD-Net* [8]	54.10	68.99	58.31	70.21	18.83
SCD-SAM* [37]	56.82	71.92	60.71	72.46	20.60
HGINet* [35]	56.13	71.73	62.88	71.90	21.83
MambaSCD* [5]	57.24	72.73	54.17	72.81	21.31
UniChange	57.62	72.85	63.50	73.12	23.02

pervision is applied to the dual-temporal remote sensing images within the WHU-CD dataset. The lora rank is set to 8, and lora alpha is consistently set to twice the lora rank.

Effects of Dual-Temporal Images Semantic Supervision. The WHU-CD dataset provides comprehensive annotations, including a change mask and individual semantic segmentation masks for both dual-temporal images (T_1 and T_2). To validate the effects of this dual-temporal semantic supervision, we conduct an ablation study categorizing experiments into four groups. As shown in Tab. 3, the

Table 3. Effects of dual-temporal images semantic supervision.

Semantic Supervision		WHU-CD			
T_1	T_2	P	R	F1	IoU
-	-	95.38	93.51	94.42	89.46
✓	-	95.77	93.45	94.59	89.74
-	✓	95.16	94.39	94.77	90.06
✓	✓	95.83	94.11	94.96	90.41

optimal performance is realized by the final group, which applies semantic supervision to both temporal images (T_1 and T_2). The intermediate configurations (supervising only T_1 or T_2) achieve progressive performance improvements over the baseline (lacking semantic supervision), but are outperformed by the fully supervised setting. This monotonic trend confirms the synergistic effect of dual-temporal semantic supervision in enhancing feature discrimination, thereby affirming its efficacy within our unified framework.

Effects of Different Vision Backbones. To validate the effects of different vision backbones, we conduct experiments using three backbones (SAM, SAM2, and RSBuilding-ViT-L) under frozen and fine-tuned (ft) configurations. Note that SAM and RSBuilding backbones share the same architecture, differing only in pre-trained weights. Results in Tab. 4 show two trends: First, the fine-tuned group consistently outperforms the frozen group. Second, RSBuilding-ViT-L delivers best results in both settings.

Effects of LoRA Ranks. To validate the effects of different LoRA ranks on model performance, we conduct experiments with LoRA ranks of 4, 8, 16, and 32, with the re-

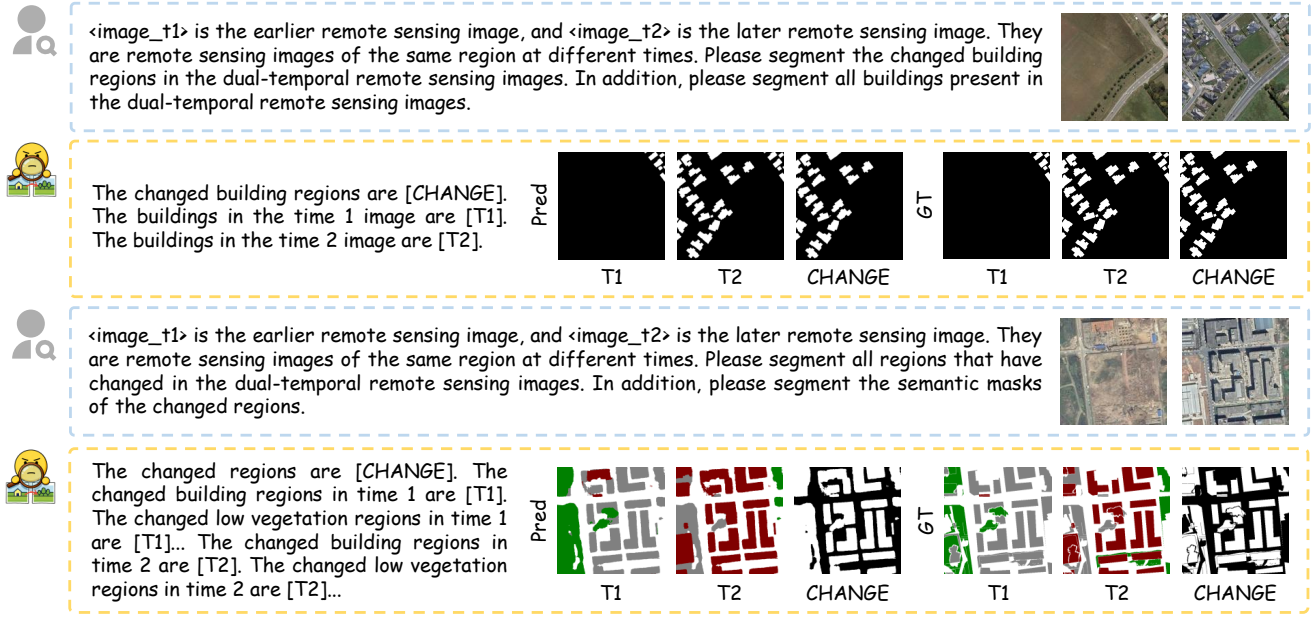


Figure 4. Visualisation results from UniChange. The images for the first question are sourced from the binary change detection dataset WHU-CD. The images for the second question are sourced from the semantic change detection dataset SECOND.

Table 4. Effects of different vision backbones.

Vision Backbone	SECOND		S2Looking	
	SeK	mIoU	F1	IoU
SAM [29]	14.63	67.78	59.51	42.36
SAM2 [41]	13.84	67.48	59.13	41.97
RSBuilding-ViT-L [46]	19.30	71.11	66.76	50.10
SAM(ft)	22.50	72.53	66.87	50.23
SAM2(ft)	22.82	72.63	67.42	50.85
RSBuilding-ViT-L(ft)	23.02	72.85	69.32	53.04

Table 5. Effects of different LoRA ranks.

LoRA Rank	WHU-CD		S2Looking	
	F1	IoU	F1	IoU
4	94.68	89.89	68.53	52.13
8	94.96	90.41	69.32	53.04
16	94.82	90.15	68.42	52.00
32	94.81	90.13	68.16	51.70

sults detailed in Tab. 5. As shown in the table, the model’s performance varies with different LoRA ranks, achieving optimal results when the LoRA rank is set to 8.

Effects of Joint Training Datasets. We design a progressive experimental setup to validate the effects of joint training with multiple datasets, with results presented in

Table 6. Effects of joint training datasets.

Training Dataset				A	B	C	D	
A	B	C	D	IoU	IoU	IoU	mIoU	SeK
✓	-	-	-	89.68	-	-	-	-
✓	✓	-	-	89.77	52.57	-	-	-
✓	✓	✓	-	90.16	52.51	78.43	-	-
✓	✓	✓	✓	90.41	53.04	78.87	72.85	23.02

Tab. 6. In this study, A, B, C, and D represent the WHU-CD, S2Looking, LEVIR-CD+, and SECOND datasets, respectively. We progressively increase the number of datasets in the training mixture, starting with A, then A+B, A+B+C, and finally A+B+C+D. Results show a clear upward trend in overall performance as more datasets are added. This indicates that joint training allows the model to learn more general and robust representations. The best results for all datasets are achieved when the model is trained jointly on all four datasets.

4.6. Visualisation Results

As illustrated in Fig. 4, we present the visualisation results for UniChange. It can be observed that UniChange demonstrates commendable performance across both BCD and SCD tasks. Further detailed visualisation results are provided in Sec. B of the supplementary materials.

5. Conclusion

In this paper, we introduce UniChange, the first unified change detection framework built upon MLLM. This framework addresses the fundamental challenges of task fragmentation and dataset incompatibility that constrain conventional change detection models. Our approach reframes both BCD and SCD tasks as “embedding as change”. By querying semantics or changes through the embeddings of three special tokens ([T1], [T2], and [CHANGE]), UniChange eliminates the need for predefined classification headers. This design uniquely allows the model to acquire comprehensive knowledge by jointly training on diverse, multi-source datasets, even when their semantic labels conflict. Our approach achieves state-of-the-art performance on four public benchmarks (WHU-CD, S2Looking, LEVIR-CD+, and SECOND), demonstrating its superiority.

References

- [1] Saeid Asadzadeh, Wilson José de Oliveira, and Carlos Roberto de Souza Filho. Uav-based remote sensing for the petroleum industry and environmental monitoring: State-of-the-art and perspectives. *Journal of Petroleum Science and Engineering*, 208:109633, 2022. 1
- [2] Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210. IEEE, 2022. 2, 3, 7
- [3] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote sensing*, 12(10):1662, 2020. 6
- [4] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:3095166, 2022. 2, 3, 7
- [5] 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. 7
- [6] Jie Chen, Ziyang Yuan, Jian Peng, Li Chen, Haozhe Huang, Jiawei Zhu, Yu Liu, and Haifeng Li. Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1194–1206, 2020. 7
- [7] Keyan Chen, Chengyang Liu, Wenyuan Li, Zili Liu, Hao Chen, Haotian Zhang, Zhengxia Zou, and Zhenwei Shi. Time travelling pixels: Bitemporal features integration with foundation model for remote sensing image change detection. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 8581–8584. IEEE, 2024. 7
- [8] Fengzhi Cui and Jie Jiang. Mtsd-net: A network based on multi-task learning for semantic change detection of bitemporal remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 118:103294, 2023. 2, 3, 7
- [9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 3
- [10] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 4063–4067. IEEE, 2018. 2, 3, 7
- [11] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187:102783, 2019. 2, 3, 7
- [12] Lei Ding, Haitao Guo, Sicong Liu, Lichao Mou, Jing Zhang, and Lorenzo Bruzzone. Bi-temporal semantic reasoning for the semantic change detection in hr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 2, 3, 7
- [13] Lei Ding, Kun Zhu, Daifeng Peng, Hao Tang, Kuiwu Yang, and Lorenzo Bruzzone. Adapting segment anything model for change detection in vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–11, 2024. 2, 3, 7
- [14] Qing Ding, Zhenfeng Shao, Xiao Huang, Fengyan Wang, and Mingchang Wang. Mlfa-net: multi-level feature-aggregated network for semantic change detection in remote sensing images. *International Journal of Digital Earth*, 17(1):2398070, 2024. 7
- [15] Qing Ding, Fengyan Wang, Mingchang Wang, Ying Zhang, and Gui Cheng. Glai-net: Global-local awareness integrated network for semantic change detection in remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025. 7
- [16] Sijun Dong, Libo Wang, Bo Du, and Xiaoliang Meng. Changeclip: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208:53–69, 2024. 7
- [17] Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li. Snunet-cd: A densely connected siamese network for change detection of vhr images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 7
- [18] Sheng Fang, Kaiyu Li, and Zhe Li. Changer: Feature interaction is what you need for change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–11, 2023. 2, 3, 7
- [19] Yuchao Feng, Jiawei Jiang, Honghui Xu, and Jianwei Zheng. Change detection on remote sensing images using dual-branch multilevel intertemporal network. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 3
- [20] Yuhang Gan, Wenjie Xuan, Hang Chen, Juhua Liu, and Bo Du. Rfl-cdnet: Towards accurate change detection via richer feature learning. *Pattern Recognition*, 153:110515, 2024. 7

- [21] Yuhang Gan, Wenjie Xuan, Zhiming Luo, Lei Fang, Zeng-mao Wang, Juhua Liu, and Bo Du. Detect changes like humans: Incorporating semantic priors for improved change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 7
- [22] Junyu Gao, Da Zhang, Feiyu Wang, Lichen Ning, Zhiyuan Zhao, and Xuelong Li. Combining sam with limited data for change detection in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 7
- [23] Haonan Guo, Xin Su, Chen Wu, Bo Du, and Liangpei Zhang. Saan: Similarity-aware attention flow network for change detection with vhr remote sensing images. *IEEE Transactions on Image Processing*, 33:2599–2613, 2024. 7
- [24] You He, Hanchao Zhang, Xiaogang Ning, Ruiqian Zhang, Dong Chang, and Minghui Hao. Spatial-temporal semantic perception network for remote sensing image semantic change detection. *Remote Sensing*, 15(16):4095, 2023. 7
- [25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4
- [26] 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. 3
- [27] Haiyan Huang, Qimin Cheng, Duowang Zhu, Xiao Huang, and Qunshan Zhao. Textscd: Leveraging text-based semantic guidance for remote sensing image semantic change detection. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 383–389, 2025. 7
- [28] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2019. 6
- [29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2, 3, 8
- [30] Kartik Kuckreja, Muhammad S Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad S Khan. Geochat: Grounded large vision-language model for remote sensing. the ieee. In *CVF Conference on Computer Vision and Pattern Recognition*, page 4, 2024. 3
- [31] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 3
- [32] Yuxuan Li, Xiang Li, Yimain Dai, Qibin Hou, Li Liu, Yongxiang Liu, Ming-Ming Cheng, and Jian Yang. Lsknet: A foundation lightweight backbone for remote sensing: Y. li et al. *International Journal of Computer Vision*, 133(3):1410–1431, 2025. 7
- [33] Yuxuan Li, Yicheng Zhang, Wenhao Tang, Yimian Dai, Ming-Ming Cheng, Xiang Li, and Jian Yang. Visual instruction pretraining for domain-specific foundation models. *arXiv preprint arXiv:2509.17562*, 2025. 3
- [34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3, 5
- [35] Jiang Long, Mengmeng Li, Xiaoqin Wang, and Alfred Stein. Semantic change detection using a hierarchical semantic graph interaction network from high-resolution remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 211:318–335, 2024. 7
- [36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [37] Liye Mei, Zhaoyi Ye, Chuan Xu, Hongzhu Wang, Ying Wang, Cheng Lei, Wei Yang, and Yansheng Li. Scd-sam: Adapting segment anything model for semantic change detection in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. 1, 2, 3, 7
- [38] Daifeng Peng, Lorenzo Bruzzone, Yongjun Zhang, Haiyan Guan, and Pengfei He. Scdnet: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery. *International Journal of Applied Earth Observation and Geoinformation*, 103:102465, 2021. 3, 7
- [39] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, 2024. 3
- [40] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020. 6
- [41] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 3, 8
- [42] Akashah Shabbir, Mohammed Zumri, Mohammed Benamoun, Fahad S Khan, and Salman Khan. Geopixel: Pixel grounding large multimodal model in remote sensing. *arXiv preprint arXiv:2501.13925*, 2025. 3
- [43] Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. S2looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24):5094, 2021. 6
- [44] Ashbindu Singh. Change detection in the tropical forest environment of northeastern india using landsat. *Remote sensing and tropical land management*, 44:273–254, 1986. 1
- [45] Biao Wang, Zhenghao Jiang, Weichun Ma, Xiao Xu, Peng Zhang, Yanlan Wu, and Hui Yang. Dual-dimension feature interaction for semantic change detection in remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:9595–9605, 2024. 7
- [46] Mingze Wang, Lili Su, Cilin Yan, Sheng Xu, Pengcheng Yuan, Xiaolong Jiang, and Baochang Zhang. Rsbuilding:

- Towards general remote sensing image building extraction and change detection with foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2, 3, 5, 8
- [47] Renfang Wang, Hucheng Wu, Hong Qiu, Feng Wang, Xiufeng Liu, and Xu Cheng. A difference enhanced neural network for semantic change detection of remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. 7
- [48] Thilo Wellmann, Angela Lausch, Erik Andersson, Sonja Knapp, Chiara Cortinovis, Jessica Jache, Sebastian Scheuer, Peleg Kremer, André Mascarenhas, Roland Kraemer, et al. Remote sensing in urban planning: Contributions towards ecologically sound policies? *Landscape and urban planning*, 204:103921, 2020. 1
- [49] Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. *arXiv preprint arXiv:2507.01467*, 2025. 3
- [50] Chuan Xu, Zhaoyi Ye, Liye Mei, Haonan Yu, Jianchen Liu, Yaxiaer Yalikun, Shuangtong Jin, Sheng Liu, Wei Yang, and Cheng Lei. Hybrid attention-aware transformer network collaborative multiscale feature alignment for building change detection. *IEEE Transactions on Instrumentation and Measurement*, 73:1–14, 2024. 2, 3
- [51] 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*, 60:3113912, 2022. 6
- [52] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 3
- [53] Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:183–200, 2020. 2, 3
- [54] Cui Zhang, Liejun Wang, Shuli Cheng, and Yongming Li. Swinsunet: Pure transformer network for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 2, 3
- [55] Da Zhang, Feiyu Wang, Lichen Ning, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Integrating sam with feature interaction for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 7
- [56] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. In *European conference on computer vision*, pages 52–70. Springer, 2024. 3
- [57] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023. 2, 3
- [58] Zhuo Zheng, Yanfei Zhong, Junjie Wang, Ailong Ma, and Liangpei Zhang. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. *Remote Sensing of Environment*, 265:112636, 2021. 1
- [59] Zhuo Zheng, Yanfei Zhong, Shiqi Tian, Ailong Ma, and Liangpei Zhang. Changemask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183:228–239, 2022. 7
- [60] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3

UniChange: Unifying Change Detection with Multimodal Large Language Model

Supplementary Material

Overview

This supplementary material provides additional details to support the main manuscript. The document is structured as follows: Sec. A begins with a detailed breakdown of the evaluation metrics, covering the specific formulas used for both BCD and SCD tasks. Following this, Sec. B offers extensive qualitative visual comparisons, benchmarking our model against other state-of-the-art methods on several key datasets. Finally, Sec. C presents a comprehensive description of the datasets utilised in our experiments, including their technical specifications and key characteristics.

A. Evaluation Metrics

A.1. BCD Metrics

To evaluate the performance of UniChange on the binary change detection (BCD) task, we employ four standard pixel-level metrics: Precision (P), Recall (R), F1 Score (F1), and Intersection over Union (IoU). Precision measures the proportion of correctly predicted change pixels among all pixels classified as change. Recall indicates the proportion of true positive pixels among all truly positive pixels in the ground truth. The F1 Score is the harmonic mean of Precision and Recall, providing a single metric that balances these two measures. Finally, IoU measures the overlap between predicted and ground-truth positive regions, serving as a robust measure of segmentation quality.

The metrics are individually defined as follows, where TP, TN, FP, and FN represent the number of true positive pixels, true negative pixels, false positive pixels, and false negative pixels, respectively:

$$\begin{aligned} P &= TP / (TP + FP), \\ R &= TP / (TP + FN), \\ F1 &= 2 \times P \times R / (P + R), \\ IoU &= TP / (TP + FP + FN). \end{aligned} \quad (14)$$

A.2. SCD Metrics

The assessment of semantic change detection (SCD) model performance is executed using a collection of specialised metrics, all derived from the confusion matrix $Q = \{q_{ij}\}$, where q_{ij} records the count of pixels classified as class i with a ground truth label of j .

The mean Intersection over Union for SCD (mIoU) is utilised to evaluate overall segmentation quality, established

as the arithmetic mean of the Intersection over Union for the regions without change (IoU_{nc}) and all changing regions (IoU_c):

$$mIoU = (IoU_{nc} + IoU_c) / 2. \quad (15)$$

IoU_{nc} measures the overlap between the predicted unchanged regions and the ground-truth unchanged regions:

$$IoU_{nc} = q_{00} / (\sum_{i=0}^N q_{i0} + \sum_{j=0}^N q_{0j} - q_{00}). \quad (16)$$

Conversely, IoU_c measures the overall segmentation quality of all change regions, treating all distinct semantic change categories as a single change class:

$$IoU_c = \sum_{i=1}^N \sum_{j=1}^N q_{ij} / (\sum_{i=0}^N \sum_{j=0}^N q_{ij} - q_{00}). \quad (17)$$

The Separation kappa coefficient (SeK) provides a valuable measure of semantic discrimination amidst class imbalance, particularly designed to diminish the influence of the prevalent unchanged class. SeK is computed from the confusion matrix $\hat{Q} = \{\hat{q}_{ij}\}$, where $\hat{q}_{ij} = q_{ij}$, but $\hat{q}_{00} = 0$, and is calculated as

$$\begin{aligned} \rho &= \sum_{i=0}^N \hat{q}_{ii} / \sum_{i=0}^N \sum_{j=0}^N \hat{q}_{ij}, \\ \eta &= \sum_{i=0}^N \left(\sum_{j=0}^N \hat{q}_{ij} \times \sum_{j=0}^N \hat{q}_{ji} \right) / \left(\sum_{i=0}^N \sum_{j=0}^N \hat{q}_{ij} \right)^2, \\ SeK &= e^{IoU_c - 1} \cdot (\rho - \eta) / (1 - \eta). \end{aligned} \quad (18)$$

The F1-Score for SCD (F_{scd}) offers a focused quantification of semantic transition accuracy within change regions. This metric is derived from the precision (P_{scd}) and recall (R_{scd}) over the pixels determined to have changed.

The SCD precision P_{scd} and recall R_{scd} are defined as

$$\begin{aligned} P_{scd} &= \sum_{i=1}^N q_{ii} / \sum_{i=1}^N \sum_{j=0}^N q_{ij}, \\ R_{scd} &= \sum_{i=1}^N q_{ii} / \sum_{i=0}^N \sum_{j=1}^N q_{ij}. \end{aligned} \quad (19)$$

The final F_{scd} is the harmonic mean of these two components:

$$F_{scd} = 2 \times P_{scd} \times R_{scd} / (P_{scd} + R_{scd}). \quad (20)$$

Table 7. Detailed information about change detection datasets used for experiments.

Dataset	Resolution	Image Size	Image Number	Evaluation Task	Class
WHU-CD	0.075m	32507×15354	1	BCD	Building
S2Looking	0.5-0.8m	1024×1024	5000	BCD	Building
LEVIR-CD+	0.5m	1024×1024	985	BCD	Building
SECOND	0.5-3m	512×512	4662	SCD	Building, Low Vegetation, Tree, Water, Playground, Bare Ground

The performance of UniChange on the SCD task is comprehensively evaluated using BCD metrics: IoU and F1 (F_{bcd}) and SCD metrics: mIoU, F_{scd} and SeK.

B. Visual Comparisons

To qualitatively evaluate the performance of our proposed UniChange, we provide comprehensive visual comparisons against other state-of-the-art (SOTA) methods on both binary and semantic change detection tasks.

Fig. 5 illustrates the visual results for binary change detection (BCD). The comparisons on the WHU-CD, S2Looking, and LEVIR-CD+ datasets all show that UniChange performs better than competing methods. Its generated change masks are more complete and have fewer false positives (FP) and false negatives (FN).

Furthermore, Fig. 6 presents the qualitative results for semantic change detection (SCD). As shown in the figure, UniChange not only accurately locates the changed regions but also exhibits a strong capability in correctly identifying the specific semantic categories of the changes. These visual results underscore the superior performance and generalisation ability of our unified model.

C. Dataset Details

To comprehensively evaluate our model, we utilise four diverse, publicly available remote sensing datasets, covering both binary change detection (BCD) and semantic change detection (SCD) tasks. The detailed specifications of these datasets are summarised in Tab. 7.

For the BCD task, we utilise three datasets, all of which focus on identifying building changes. The WHU-CD dataset provides a single, massive-scale image pair (32507×15354 pixels), captured at an ultra-high 0.075m resolution. It is particularly notable for its coverage, capturing a 20.5 km² area in Christchurch, New Zealand. The dual-temporal images, captured in 2012 and 2016, document the region’s reconstruction following a major earthquake. This provides an excellent case for studying large-scale urban recovery. In addition, we use S2Looking, a large-scale dataset consisting of 5000 dual-temporal image pairs, which are 1024×1024 pixels in size and have a resolution of 0.5-0.8m. Its defining characteristic is the use of

satellite side-looking images captured at various off-nadir angles, a sharp contrast to typical near-nadir (or top-down) imagery. This dataset focusses on globally distributed rural areas. It challenges models with large illumination variances and complex rural scenes. It also features geometric distortions from the oblique angles. Finally, we utilise LEVIR-CD+, an expanded version of the LEVIR-CD dataset. It contains 985 image pairs of 0.5m resolution Google Earth imagery. Each image pair is 1024×1024 pixels. In contrast to S2Looking, LEVIR-CD+ features near-nadir images and focusses on building changes in urban and suburban environments, primarily covering 20 different regions in Texas, USA.

For the more complex SCD task, we use the SECOND dataset. This is a large-scale benchmark containing 4,662 pairs of aerial images, each 512×512 pixels in size. The images come from several platforms and sensors. They cover major cities in China, including Hangzhou, Chengdu, and Shanghai. Unlike the BCD datasets, SECOND requires the model to identify semantic transitions among six distinct land-cover classes: building, low vegetation, tree, water, playground, and bare ground. A defining feature of this dataset is its annotation method; it provides land-cover map pairs and nonchange masks. This structure is specifically designed to challenge models. It tests their ability to detect changes that occur between the same land-cover class. For example, a model must find where an old playground is removed and a new one is built in its place. This is a critical capability that many other datasets cannot evaluate.

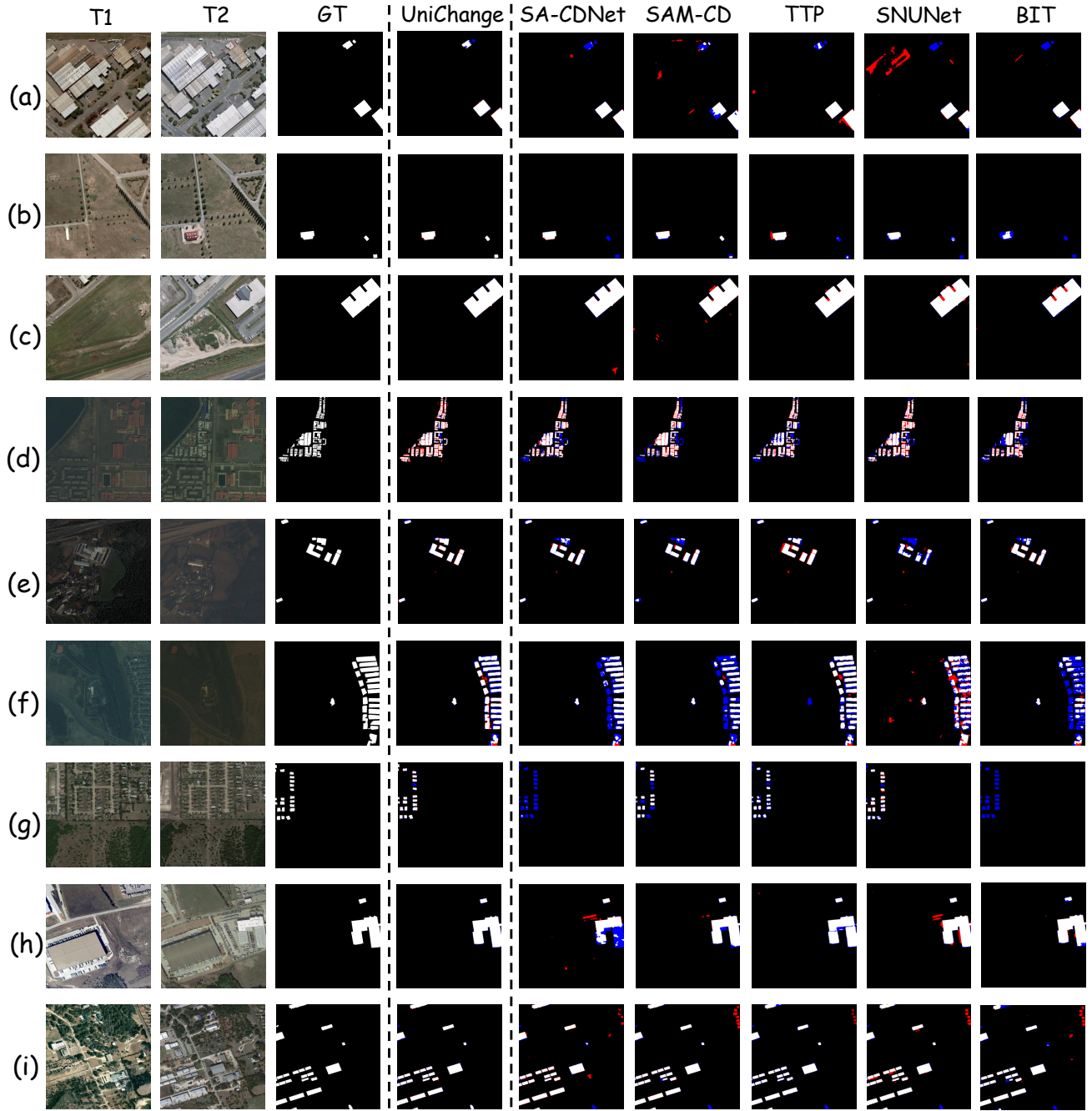


Figure 5. Visual comparisons of the UniChange with other state-of-the-art methods for binary change detection. **Red** means false positives (FP), while **Blue** denotes false negatives (FN). Samples (a) (b) (c) are from the WHU-CD dataset, (d) (e) (f) are from the S2Looking dataset, and (g) (h) (i) are from the LEVIR-CD+ dataset.

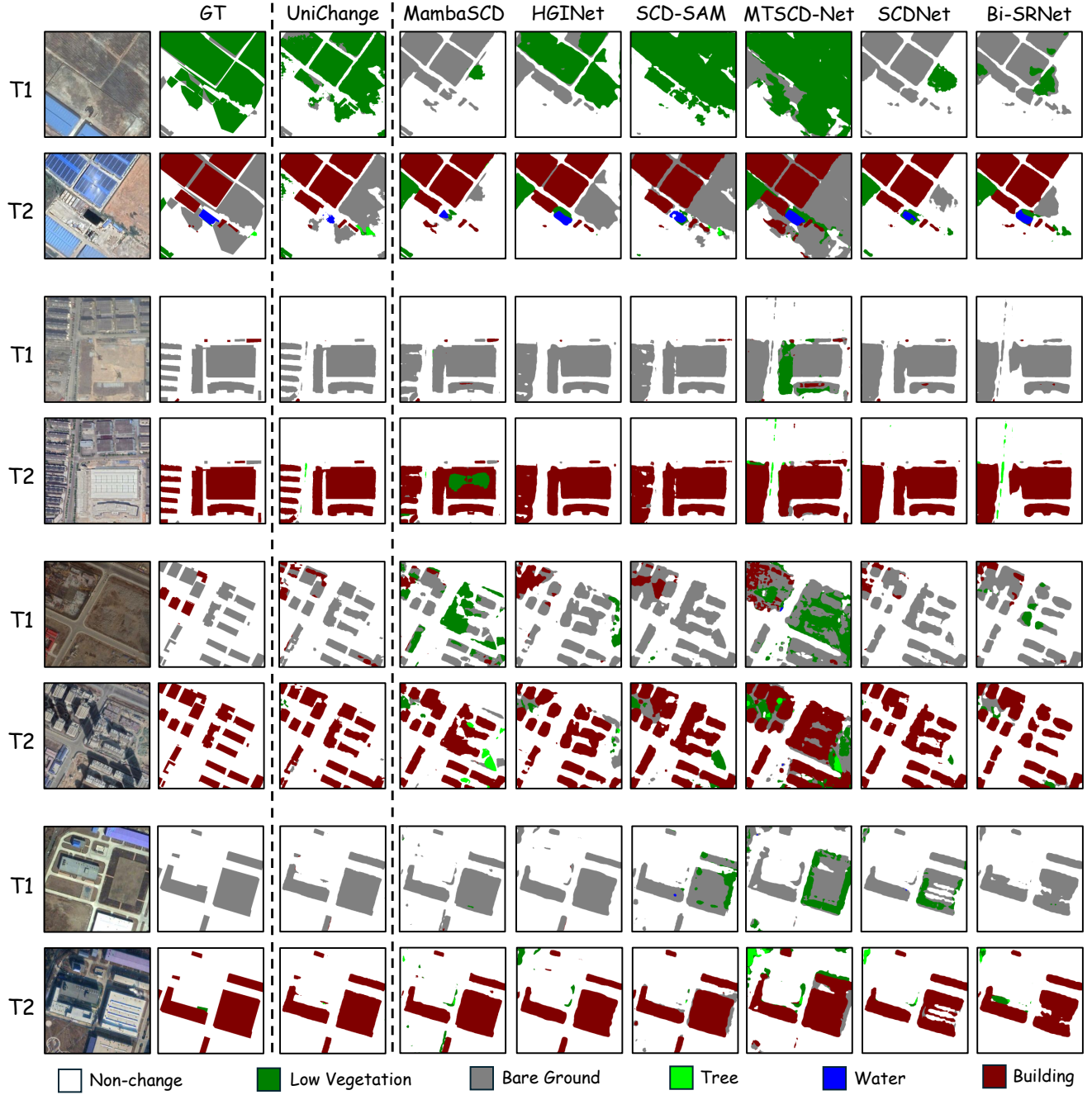


Figure 6. Visual comparisons of the UniChange with other state-of-the-art methods for semantic change detection. The colour legend is as follows: **White** represents Non-change, **dark green** represents Low Vegetation, **grey** represents Bare Ground, **bright green** represents Tree, **blue** represents Water, and **dark red** represents Building.