TMCD-RS: Trustworthy Multimodal Change Detection for Bi-Temporal Remote Sensing

Anonymous Author(s)

Abstract

Bi-temporal remote sensing change detection is essential for urban monitoring, disaster response, and infrastructure assessment. However, existing change detection models often rely on task-specific dense supervision and are highly sensitive to temporal misalignment, background clutter, and cross-domain distribution shifts. To address these limitations, we propose **TMCD-RS**, a lightweight vision–language framework that reformulates bi-temporal change detection as text-guided structural change reasoning.

Built upon a frozen CLIP backbone, TMCD-RS adopts a shared-weight Siamese visual encoder to process pre-event and post-event images jointly. Instead of relying solely on raw image differencing, the proposed method performs temporal reasoning in the feature space by combining absolute bi-temporal representations with residual change evidence derived from their feature discrepancies. To improve robustness under imperfect temporal correspondence, we introduce a **reliability-aware temporal fusion** module that predicts a confidence score from the global temporal discrepancy and uses it to adaptively modulate both image-level and pixel-level fusion. In parallel, a learnable multi-scale fusion module aggregates text-guided anomaly maps from multiple intermediate layers, enabling fine-grained localization of changed structures.

TMCD-RS follows a two-stage training strategy. In Stage 1, lightweight text adapters are optimized to learn disentangled normal and abnormal textual prototypes for change-aware semantic alignment. In Stage 2, the text branch is fixed, while image adapters and temporal fusion modules are optimized to capture domain-invariant structural change cues from paired observations. This design preserves the open-vocabulary generalization of CLIP while introducing only a small number of trainable parameters.

Experiments on remote sensing building-change benchmarks demonstrate that TMCD-RS achieves strong localization and image-level discrimination, while showing improved robustness to pre-existing structures and better transferability across datasets. These results suggest that confidence-aware bi-temporal feature fusion provides an effective and practical direction for trustworthy remote sensing change reasoning.

Keywords

vision-language models, remote sensing change detection, Siamese temporal adaptation, confidence-aware fusion, trustworthy multimodal learning

Conference'17, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Remote sensing change detection is a fundamental problem in Earth observation, with applications in urban expansion monitoring, disaster response, infrastructure assessment, and environmental management [6, 9, 17]. Among these applications, building-related change analysis is particularly important because newly constructed, damaged, or altered structures often serve as direct indicators of socioeconomic activity and post-event impact. Recent progress in deep learning has substantially advanced bi-temporal change detection by learning discriminative representations from paired observations acquired before and after an event [1, 4, 5, 7]. However, despite strong benchmark performance, many existing methods remain highly sensitive to temporal misalignment, illumination variation, background clutter, and cross-domain distribution shifts across sensors, regions, and acquisition conditions [6, 10, 18].

A central limitation of conventional change detection pipelines is that they are typically optimized as task-specific supervised architectures. Most approaches explicitly compare pre-event and post-event images through Siamese encoders, feature differencing, or cross-temporal interaction modules, and then train the full model on a target dataset with dense pixel-level annotations [1, 4, 5, 7]. While effective in-domain, such designs often entangle genuine structural change with dataset-specific appearance statistics, making them vulnerable under zero-shot transfer or cross-dataset deployment [10, 18]. This issue is particularly problematic in high-stakes remote sensing scenarios, where performance under imperfect inputs matters as much as average in-domain accuracy.

In parallel, vision–language models (VLMs), especially CLIP, have demonstrated strong open-vocabulary semantic generalization by aligning images and text in a shared representation space [15, 16]. This property is attractive for remote sensing, where dense annotations are expensive and deployment conditions are often more variable than those assumed by standard supervised benchmarks. Recent works such as RemoteCLIP, GeoCLIP, GRAFT, and open-vocabulary remote sensing segmentation further suggest that vision–language pretraining can provide transferable priors for geospatial understanding without exhaustive downstream retraining [2, 3, 11, 13]. However, directly applying CLIP to remote sensing change detection is nontrivial. CLIP is not explicitly designed for paired temporal reasoning, and naive extensions based on raw image subtraction or uncalibrated feature differencing may amplify spurious changes caused by misregistration, seasonal variation, or irrelevant background fluctuations.

Recent anomaly-aware CLIP frameworks provide an appealing starting point for this problem. Methods such as AA-CLIP show that lightweight adapters and anomaly-aware semantic prototypes can improve dense localization while preserving the pretrained backbone [12]. RSAD-CLIP further demonstrates the promise of zero-shot anomaly reasoning in remote sensing scenes [19]. Nevertheless, most existing CLIP-based anomaly localization methods

are primarily developed for single-image reasoning and do not explicitly address the reliability of bi-temporal evidence. As a result, visually plausible temporal discrepancies may still be over-amplified even when they are not trustworthy indicators of structural change.

In this work, we argue that trustworthy remote sensing change reasoning should satisfy two requirements. First, it should preserve the transferable semantic prior of pretrained VLMs instead of over-specializing to a single source domain. Second, it should explicitly model the reliability of temporal discrepancy rather than assuming that all observed differences are equally informative. Motivated by these principles, we propose **TMCD-RS** (**S**iamese **T**emporal **A**dapter and **R**eliability-aware fusion for **R**emote **S**ensing), a lightweight bi-temporal vision–language framework for text-guided structural change reasoning.

Built upon a frozen CLIP backbone, TMCD-RS processes pre-event and post-event images using a shared-weight Siamese visual encoder and introduces only a small number of trainable parameters through lightweight adapters. The framework first learns disentangled normal and abnormal semantic prototypes through a semantic alignment stage, and then performs temporal reasoning in feature space by combining absolute bi-temporal evidence with residual discrepancy cues. To improve robustness under noisy temporal correspondence, TMCD-RS further introduces a confidence-aware fusion mechanism that calibrates both image-level and pixel-level predictions according to the estimated reliability of temporal discrepancy. This design is conceptually aligned with the broader need for reliable and calibrated visual decision making under imperfect inputs [14].

Our contributions are threefold:

- We reformulate remote sensing building change detection as a trustworthy bi-temporal structural change reasoning problem within a frozen vision–language framework.
- We propose a shared-weight Siamese temporal adaptation architecture with reliability-aware fusion, which combines absolute temporal evidence and residual feature discrepancy in a lightweight yet principled manner.
- We position confidence-aware temporal fusion as a practical design principle for robust remote sensing VLMs under domain shift and imperfect temporal correspondence.

2 Related Work

2.1 Remote Sensing Change Detection

Remote sensing change detection has been extensively studied for applications such as urban monitoring, disaster assessment, and land-use analysis [6, 9, 17]. Early deep learning methods commonly relied on Siamese convolutional architectures that extracted features from pre-event and post-event images and compared them through differencing or joint decoding [7]. Subsequent approaches incorporated stronger temporal interaction mechanisms, including spatial–temporal attention and transformer-based encoders, to improve the localization of changed regions [1, 4, 5]. Although these models have demonstrated strong performance on benchmark datasets, they are still typically trained and evaluated in supervised in-domain settings, where image registration and data distribution are relatively controlled.

2.2 Change Detection Under Domain Shift

A long-standing challenge in remote sensing is that data distributions vary substantially across sensors, cities, seasons, and acquisition conditions. This issue has motivated extensive research on domain adaptation and domain generalization for remote sensing classification and segmentation [10, 18]. For change detection, such variation is particularly harmful because the model must distinguish true structural change from nuisance factors such as illumination difference, seasonal appearance shifts, or background clutter [6]. As a result, high in-domain performance does not necessarily translate into reliable behavior on unseen target domains. In this paper, we address this challenge by retaining a frozen vision–language backbone and introducing lightweight temporal adaptation, with the goal of preserving transferable semantics while reducing over-specialization to source-domain appearance statistics.

2.3 Vision–Language Models for Remote Sensing

The success of CLIP has inspired a growing body of work on vision–language modeling for remote sensing [15, 16]. Recent studies such as RemoteCLIP, GeoCLIP, and GRAFT demonstrate that large-scale image–text pretraining can provide useful priors for geospatial understanding, while open-vocabulary remote sensing segmentation further extends this paradigm to dense prediction [2, 3, 11, 13]. These works collectively suggest that VLMs are promising for remote sensing applications where exhaustive annotation is expensive or impractical. However, most existing remote sensing VLM research focuses on semantic understanding, retrieval, geolocalization, or single-image dense prediction, rather than explicit bi-temporal reasoning under unreliable temporal correspondence.

2.4 CLIP-Based Anomaly Reasoning

Anomaly-aware CLIP adaptation has recently emerged as an effective direction for zero-shot anomaly localization. AA-CLIP improves zero-shot anomaly detection by introducing anomaly-aware semantic disentanglement and staged adaptation between text and image branches [12]. In the remote sensing domain, RSAD-CLIP demonstrates that pretrained vision–language models can also be used for zero-shot anomaly reasoning on Earth surface imagery [19]. These studies are highly relevant because building change can be interpreted as a structurally meaningful anomaly relative to the surrounding scene. However, existing CLIP-based anomaly localization methods are largely designed for single-image inference and do not explicitly model the reliability of paired temporal evidence. By contrast, our method incorporates pre-event and post-event observations into a shared feature space and uses confidence-aware fusion to calibrate how strongly temporal discrepancy should affect the final decision.

2.5 Parameter-Efficient Adaptation and Reliable Fusion

Full fine-tuning of large pretrained multimodal models is computationally expensive and may weaken their pretrained generalization ability. Parameter-efficient transfer methods provide a practical alternative by introducing a small number of trainable parameters

while keeping the backbone frozen. This principle is especially suitable for remote sensing, where datasets are comparatively limited and distribution shift is common. Our design is particularly inspired by the staged adaptation philosophy of anomaly-aware CLIP frameworks [12], but extends it to bi-temporal change reasoning.

A second key issue is how to fuse temporal evidence reliably. Naive pixel-wise subtraction or uncalibrated feature differencing may over-amplify nuisance variation rather than genuine structural change. Recent studies in both remote sensing and computer vision have highlighted the importance of robust representation design and calibration under imperfect inputs [14]. In our setting, temporal evidence should be treated as conditionally reliable: it may strongly indicate change in some cases, but be corrupted or ambiguous in others. This motivates our reliability-aware temporal fusion strategy, which preserves absolute structural evidence from each temporal observation while using a confidence head to modulate the contribution of residual discrepancy features.

2.6 Positioning of the Proposed Method

The proposed TMCD-RS framework sits at the intersection of remote sensing change detection, vision–language adaptation, and trust-oriented multimodal inference. Compared with conventional supervised change detection networks, TMCD-RS retains a frozen vision–language backbone and emphasizes transferability under domain shift. Compared with existing CLIP-based anomaly localization methods, it explicitly handles paired temporal inputs and models the reliability of temporal discrepancy. Compared with naive bi-temporal differencing pipelines, it performs temporal reasoning in feature space and calibrates fusion through an intrinsic confidence mechanism. As a result, TMCD-RS provides a unified formulation for trustworthy bi-temporal structural change reasoning in remote sensing.

3 Proposed Method

3.1 Problem Formulation

Given a pre-event image $x^{(1)} \in \mathbb{R}^{3\times H\times W}$ and a post-event image $x^{(2)} \in \mathbb{R}^{3\times H\times W}$, our goal is to predict a pixel-level change mask $m \in \{0,1\}^{H\times W}$ and an image-level change label $y \in \{0,1\}$. Unlike conventional fully supervised change detection pipelines that rely on task-specific temporal encoders, we formulate the task as *text-guided structural change reasoning* using a frozen vision–language model. The key idea is to preserve CLIP's open-vocabulary semantic prior while introducing lightweight temporal adaptation for paired remote sensing observations.

For each semantic category c, we define a binary textual prototype matrix

$$T_c = [t_c^n,\ t_c^a] \in \mathbb{R}^{d\times 2},$$

where t_c^n and t_c^a denote the normal and abnormal semantic prototypes, respectively, and d is the embedding dimension.

3.2 Overall Architecture

Our framework, termed **TMCD-RS**, is built upon a frozen CLIP ViT-L/14 backbone. To capture structural evidence at multiple resolutions, we extract intermediate patch tokens from transformer layers

$$\mathcal{L} = \{6, 12, 18, 24\}.$$

Given an input image, the visual encoder outputs multi-scale patch features

$$F_\ell \in \mathbb{R}^{B\times N\times d}, \quad \ell \in \mathcal{L},$$

where N is the number of spatial patches, together with a global detection descriptor obtained from the deepest visual representation.

To enable task-specific adaptation while maintaining the generalization ability of the pretrained backbone, we insert lightweight adapters into both the text and image branches. The text branch is adapted in Stage 1 for semantic disentanglement, while the image branch is adapted in Stage 2 for structural change reasoning.

3.3 Stage 1: Semantic Alignment via Text Adapters

We first optimize the text branch to construct stable anomaly-aware semantic prototypes. For each class c, we use prompt ensembling to describe both normal and changed states, e.g., "unchanged building," "newly built building," and "building change." Let $\mathcal{P}_c^n$ and $\mathcal{P}_c^a$ denote the sets of normal and abnormal prompts. Their corresponding text embeddings are averaged to form

$$t_c^s = \mathrm{Norm}\left(\frac{1}{|\mathcal{P}_c^s|}\sum_{p\in\mathcal{P}_c^s}\mathrm{Norm}(E_{\text{text}}(p))\right), \quad s \in \{n, a\},$$

where $E_{\text{text}}(\cdot)$ denotes the adapted CLIP text encoder.

To explicitly separate the two semantic states, we impose an orthogonality regularizer

$$\mathcal{L}_{\text{orth}} = \left((t_c^n)^\top t_c^a\right)^2.$$

During Stage 1, only the text adapters are optimized, while the image branch remains frozen. This stage stabilizes class-level change semantics before temporal visual adaptation.

3.4 Stage 2: Siamese Temporal Visual Adaptation

In Stage 2, the text prototypes are fixed, and a shared-weight Siamese visual encoder is used to process $x^{(1)}$ and $x^{(2)}$:

$$\{F_\ell^{(1)}\}_{\ell\in\mathcal{L}},\ d^{(1)} = E_{\text{img}}(x^{(1)}), \qquad \{F_\ell^{(2)}\}_{\ell\in\mathcal{L}},\ d^{(2)} = E_{\text{img}}(x^{(2)}),$$

where $F_\ell^{(1)}$ and $F_\ell^{(2)}$ are the multi-scale patch embeddings for the two temporal observations, and $d^{(1)}, d^{(2)} \in \mathbb{R}^{B\times d}$ are global detection descriptors.

The two branches share all encoder and adapter weights. This design encourages the model to learn *change-sensitive yet domain-consistent* representations, rather than memorizing temporal order-specific appearance patterns.

3.5 Reliability-Aware Temporal Fusion

A central challenge in bi-temporal remote sensing is that naive differencing may amplify noise caused by misregistration, illumination variation, or background clutter. Instead of directly training on raw $T_2 - T_1$ images, we perform residual reasoning in the feature space.

We first compute a global temporal discrepancy

$$\Delta_d = |d^{(2)} - d^{(1)}|.$$

A lightweight confidence head predicts a sample-wise reliability score

$$r = \sigma(\mathrm{MLP}(\Delta_d)) \in [0, 1].$$

We then combine absolute bi-temporal evidence and residual temporal evidence through

$$\tilde{d} = \mathrm{Norm}\left(\frac{d^{(1)} + d^{(2)}}{2} + r\,\beta\,|d^{(2)} - d^{(1)}|\right),$$

where β is a learnable temporal residual scale.

For each scale ℓ, patch-level temporal fusion is performed analogously:

$$\tilde{F}_\ell = \mathrm{Norm}\left(\frac{F_\ell^{(1)} + F_\ell^{(2)}}{2} + r\,\beta\,|F_\ell^{(2)} - F_\ell^{(1)}|\right).$$

This formulation allows the model to emphasize residual change evidence when temporal discrepancy is reliable, while falling back to more conservative absolute structural evidence when uncertainty is high.

3.6 Text-Guided Dense Prediction and Multi-Scale Fusion

Given the fused patch features $\tilde{F}_\ell$ and textual prototype matrix T_c, we compute class-aware similarity maps at each scale:

$$S_\ell = \mathrm{Sim}(\tilde{F}_\ell, T_c),$$

where $\mathrm{Sim}(\cdot,\cdot)$ denotes cosine-similarity-based matching between visual features and the normal/abnormal text embeddings. These maps are reshaped into dense 2-channel predictions over the spatial grid.

To aggregate complementary evidence from multiple transformer depths, we introduce a learnable multi-scale fusion module:

$$M_{\mathrm{fuse}} = \Phi_{\mathrm{ms}}(S_6, S_{12}, S_{18}, S_{24}),$$

where Φ_{ms} denotes the fusion network. To improve robustness, the final pixel-level prediction is confidence-calibrated as

$$\hat{M} = r\,M_{\mathrm{fuse}} + (1 - r)\,\bar{M},$$

where $\bar{M}$ is the simple average of the scale-specific predictions. This prevents unstable scales from dominating the output in uncertain temporal cases.

For image-level prediction, the fused global descriptor $\tilde{d}$ is matched with the text prototypes, followed by a softmax over the normal/abnormal states to obtain the final change probability.

3.7 Training Objective

The overall training is conducted in two stages. Stage 1 optimizes the text branch with semantic alignment and orthogonality regularization. Stage 2 optimizes the image adapters, temporal residual scale, confidence head, and multi-scale fusion module using both image-level and pixel-level supervision:

$$\mathcal{L}_{\mathrm{stage2}} = \mathcal{L}_{\mathrm{cls}} + \lambda \mathcal{L}_{\mathrm{seg}},$$

where $\mathcal{L}_{\mathrm{cls}}$ is the image-level cross-entropy loss over the normal/abnormal states, and $\mathcal{L}_{\mathrm{seg}}$ is the dense segmentation loss computed from the fused pixel predictions.

This two-stage optimization preserves the semantic prior of CLIP while enabling efficient temporal adaptation with only a small number of trainable parameters.

4 Experiments

4.1 Datasets and Evaluation Protocol

We evaluate **TMCD-RS** on remote sensing building-change benchmarks under a *trust-oriented bi-temporal transfer* setting. Specifically, we use paired pre-event and post-event observations (T_1, T_2) and assess the model from two complementary perspectives: (i) standard change localization and image-level discrimination, and (ii) robustness under domain shift and imperfect temporal correspondence [6, 10, 14, 18].

Following the revised bi-temporal formulation, LEVIR-CD is used as the primary source-domain dataset for training, where paired images and change masks provide supervision for structural change reasoning [5]. To evaluate cross-domain transferability, we further test on the WHU Building dataset without target-domain fine-tuning [8]. This protocol is intentionally challenging: the source and target domains differ in geographic layout, building morphology, and imaging characteristics, thereby providing a realistic testbed for trustworthy remote sensing VLM adaptation [6, 10, 18].

In contrast to the previous single-temporal formulation, our revised setting explicitly incorporates both temporal observations during inference. This allows us to evaluate not only whether the model can localize changed structures, but also whether it can suppress spurious activations caused by pre-existing buildings, ambiguous boundaries, or unreliable temporal discrepancy.

4.2 Implementation Details

All experiments are implemented in PyTorch with a frozen CLIP ViT-L/14-336 backbone [15, 16]. Input images are resized to 518×518, and multi-scale patch tokens are extracted from transformer layers $\{6, 12, 18, 24\}$. To preserve the pretrained open-vocabulary prior while introducing task-specific temporal reasoning, we optimize TMCD-RS in two stages, following the general staged adaptation philosophy of anomaly-aware CLIP frameworks [12].

In Stage 1, only the text adapters are trained for semantic alignment. The text branch is optimized for 5 epochs using AdamW with a learning rate of 1×10^{-5} and a batch size of 16. In Stage 2, the text prototypes are fixed, while the image adapters, the multi-scale fusion module, the confidence head, and the temporal residual scale are optimized for 20 epochs using AdamW with a learning rate of 5×10^{-4} and a batch size of 2. Unless otherwise specified, all experiments use the shared-weight Siamese setting with paired (T_1, T_2) inputs.

4.3 Evaluation Metrics

To comprehensively evaluate both localization quality and decision reliability, we report pixel-level and image-level metrics. For dense prediction, we report Pixel AUC (P-AUC), Pixel Average Precision

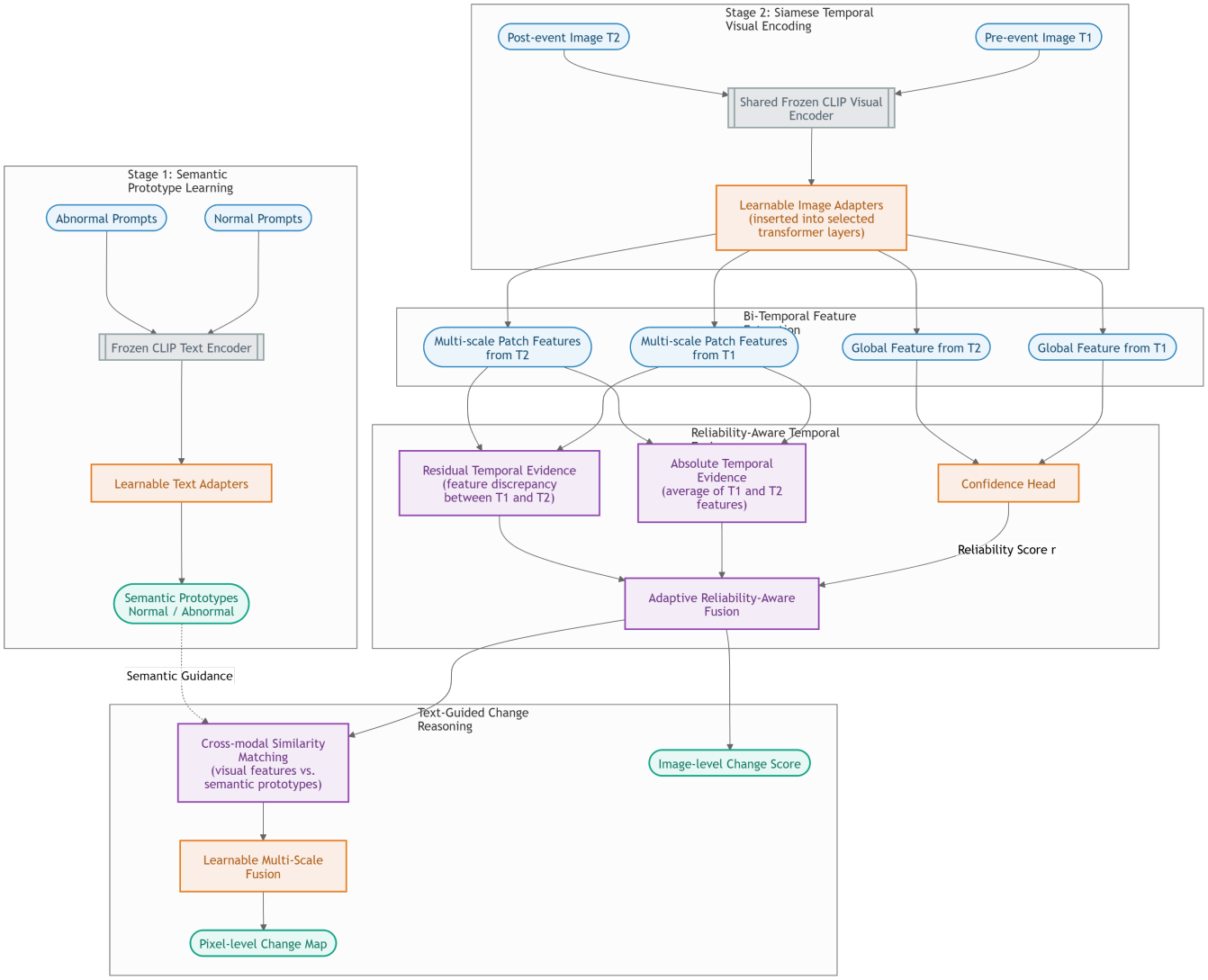

Figure 1: Overview of the proposed TMCD-RS framework for trustworthy bi-temporal remote sensing change reasoning. In Stage 1, a frozen CLIP text encoder with learnable text adapters constructs disentangled normal and abnormal semantic prototypes. In Stage 2, paired pre-event and post-event images (T_1 and T_2) are processed by a shared frozen CLIP visual encoder with lightweight image adapters. The resulting bi-temporal features are decomposed into absolute temporal evidence and residual discrepancy evidence, which are adaptively calibrated by a confidence head through reliability-aware fusion. Finally, the fused visual representation is aligned with the learned semantic prototypes for pixel-level change localization and image-level change prediction.

(P-AP), Pixel F1-score (P-F1), and Pixel IoU (P-IoU). For image-level prediction, we report Image AUC (I-AUC) and Image Average Precision (I-AP).

Since remote sensing change detection is highly imbalanced and sensitive to threshold choice, P-F1 and P-IoU are computed at the threshold that maximizes F1 on the evaluation set. Compared with reporting AUC alone, these threshold-dependent metrics better reflect whether the model produces operationally usable change masks.

4.4 Trustworthiness-Oriented Evaluation

To better align with trustworthy VLM evaluation, we analyze TMCD-RS beyond conventional benchmark accuracy. In particular, we study three aspects: **(1)** cross-dataset generalization from source to unseen target domains, **(2)** robustness to imperfect temporal evidence, such as noisy alignment or ambiguous structural difference, and **(3)** confidence-aware behavior, i.e., whether the model becomes more conservative when temporal discrepancy is unreliable [10, 14, 18].

Table 1: Cross-dataset performance comparison of TMCD-RS for bi-temporal remote sensing change reasoning. Best results are in bold.

2*Target Dataset	2*Source Setting	Image-level Change Detection		Pixel-level Localization			
		AUROC	AP	AUROC	AP	F1	IoU
1*WHU	LEVIR-CD → WHU	**98.88**	**99.63**	**96.12**	**74.18**	**63.47**	**47.09**
1*LEVIR-CD	WHU → LEVIR-CD	93.18	99.34	91.06	35.84	32.91	20.42

This evaluation perspective is important because, in real-world remote sensing deployment, temporal difference is not always a trustworthy indicator of structural change. A model that merely amplifies all discrepancy signals may achieve visually plausible outputs while remaining unreliable under domain shift or temporal corruption. Accordingly, we treat reliability under imperfect inputs as a first-class evaluation objective rather than a secondary diagnostic.

5 Results

5.1 Main Quantitative Results

Table 1 summarizes the main cross-dataset results of TMCD-RS. Overall, the proposed framework achieves strong performance in both pixel-level localization and image-level discrimination, indicating that the integration of paired temporal evidence with text-guided semantic prototypes is effective for remote sensing structural change reasoning.

More specifically, the transfer behavior is clearly asymmetric. When trained on LEVIR-CD and evaluated on WHU, TMCD-RS achieves the strongest overall results, including **98.88** I-AUC, **99.63** I-AP, **96.12** P-AUC, **74.18** P-AP, **63.47** P-F1, and **47.09** P-IoU. In contrast, the reverse setting, WHU→LEVIR-CD, drops to 93.18 I-AUC, 99.34 I-AP, 91.06 P-AUC, 35.84 P-AP, 32.91 P-F1, and 20.42 P-IoU. This large gap indicates that the proposed framework transfers more effectively when trained on the structurally richer LEVIR-CD source domain, suggesting that source-domain diversity plays an important role in trustworthy cross-dataset generalization [6, 10, 18].

Compared with post-event-only reasoning, the revised bi-temporal TMCD-RS formulation provides a more faithful basis for change-specific reasoning, since the model can explicitly compare pre-event and post-event structural evidence rather than inferring change from post-event appearance alone. More importantly, the new formulation is better aligned with trustworthy deployment: it reduces the tendency to misclassify pre-existing man-made structures as anomalous simply because they are visually salient.

5.2 Cross-Dataset Generalization

A key result is that TMCD-RS maintains strong transferability under cross-dataset evaluation. When trained on the source domain and directly evaluated on the target domain without fine-tuning, the model remains competitive across both localization and image-level metrics. This suggests that the frozen CLIP backbone retains semantically transferable priors [15, 16], while the lightweight temporal adapters provide sufficient flexibility to model structural change without over-specializing to source-domain appearance statistics [10, 12, 18].

Table 2: Ablation study of TMCD-RS components on cross-dataset bi-temporal remote sensing change reasoning. Best results are in bold.

2*Target Dataset	2*Source Dataset	2*Method Variant	Image-level Change Detection		Pixel-level Localization			
			AUROC	AP	AUROC	AP	F1	IoU
6*WHU	6*LEVIR-CD	Frozen CLIP baseline	88.50	91.50	82.52	39.84	33.81	20.54
		+ Text Adapter only	94.82	96.25	88.15	46.72	38.65	24.91
		+ Image Adapter only	96.45	97.82	92.48	63.21	54.22	37.85
		+ Siamese temporal fusion	97.56	98.41	94.63	68.94	58.16	41.26
		+ Residual temporal branch	98.21	99.07	95.54	72.03	61.24	44.65
		+ Confidence-aware fusion (Full)	**98.88**	**99.63**	**96.12**	**74.18**	**63.47**	**47.09**

Importantly, the goal here is not merely to maximize in-domain fitting, but to preserve reliable behavior under unseen data distributions. From a trustworthiness perspective, such cross-domain robustness is especially meaningful: it indicates that the model is not simply memorizing source-domain textures or spatial layouts, but is instead learning a more stable notion of building-related structural change.

5.3 Reliability Under Imperfect Temporal Evidence

Beyond aggregate performance, the revised TMCD-RS framework is designed to behave more conservatively when temporal evidence is unreliable. Because the model combines absolute temporal representations with residual discrepancy cues, it is less dependent on a single differencing mechanism. The confidence-aware fusion module further modulates how strongly temporal discrepancy influences the final decision, which is consistent with recent findings on robustness and calibration under imperfect inputs [14].

This behavior is also reflected in the robustness study of Table 3. Under clean paired inputs, the full model achieves 63.47 P-F1 and 47.09 P-IoU. After introducing temporal shift, the full model degrades to 59.08 P-F1 and 42.96 P-IoU, while retaining a high 97.94 I-AUC. However, removing confidence-aware fusion leads to a larger drop, reducing performance further to 54.21 P-F1, 38.87 P-IoU, and 96.85 I-AUC. These results indicate that confidence-aware fusion improves robustness by preventing unreliable temporal discrepancy from dominating the final prediction [14].

5.4 Qualitative Analysis

Figure 2 presents representative qualitative examples. In typical cases, TMCD-RS produces compact and spatially coherent anomaly maps that align well with building-level change regions. Compared with post-event-only reasoning, the addition of explicit temporal context helps the model focus on newly emerged or structurally modified regions, while suppressing false activations on unchanged but visually prominent structures.

Failure cases remain in scenes with extremely small buildings, heavy shadow interference, or severe temporal ambiguity. However, even in these difficult examples, the model tends to produce more localized and interpretable responses rather than large chaotic false positives. This behavior is desirable from a trust perspective, since it indicates that the model fails in a more constrained and diagnosable manner.

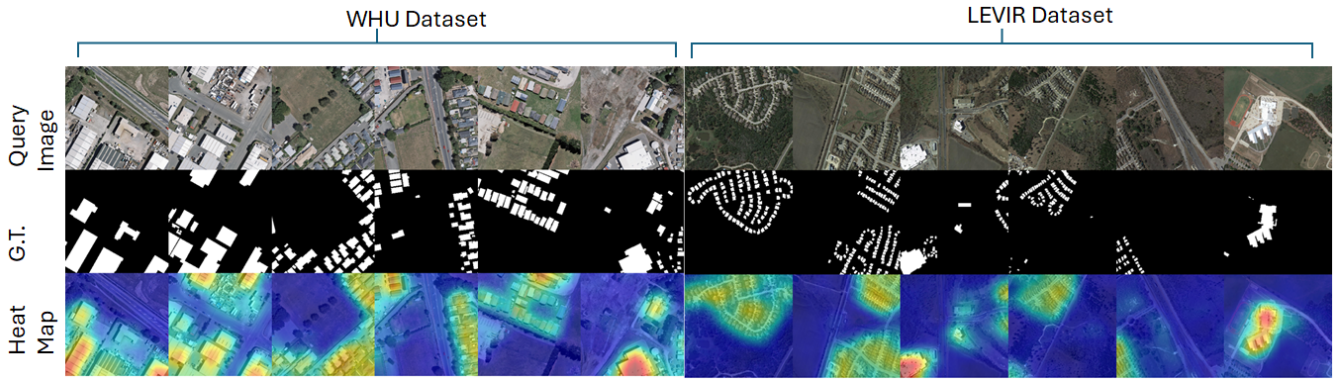

Figure 2: Qualitative results of TMCD-RS under cross-dataset bi-temporal change reasoning. From top to bottom, each example shows the paired input images (T_1 and T_2), the ground-truth change mask, and the predicted change heatmap. Results are illustrated on LEVIR-CD and WHU to demonstrate the ability of TMCD-RS to localize structurally changed regions under zero-shot cross-dataset transfer.

Table 3: Robustness study of TMCD-RS under imperfect temporal inputs. Best results are in bold.

2*Target Dataset	2*Source Setting	Image-level Change Detection		Pixel-level Localization			
		AUROC	AP	AUROC	AP	F1	IoU
3*WHU	Full model (clean pair)	**98.88**	**99.63**	**96.12**	**74.18**	**63.47**	**47.09**
	Full model + temporal shift	97.94	99.18	94.83	70.52	59.08	42.96
	No confidence fusion + temporal shift	96.85	98.74	93.27	66.41	54.21	38.87

6 Ablation Study

6.1 Component-Wise Ablation

To quantify the contribution of each design choice, we conduct ablations that directly reflect the revised TMCD-RS pipeline. Specifically, we evaluate: **(1)** frozen CLIP baseline, **(2)** + Text Adapter only, **(3)** + Image Adapter only, **(4)** + Siamese temporal fusion, **(5)** + Residual temporal branch, and **(6)** + Confidence-aware fusion (Full).

As shown in Table 2, the performance gain is progressive and internally consistent with the proposed pipeline. Starting from the frozen CLIP baseline, TMCD-RS improves from 88.50 to 98.88 in I-AUC, from 82.52 to 96.12 in P-AUC, from 33.81 to 63.47 in P-F1, and from 20.54 to 47.09 in P-IoU. Adding the text adapter first improves semantic disentanglement, raising I-AUC from 88.50 to 94.82 and P-F1 from 33.81 to 38.65. Introducing the image adapter then produces the largest single jump in dense localization, increasing P-F1 from 38.65 to 54.22 and P-IoU from 24.91 to 37.85, which is consistent with the role of learnable visual adaptation in enhancing spatial sensitivity [12].

Temporal reasoning contributes further gains. Relative to the image-adapter-only variant, Siamese temporal fusion improves P-F1 from 54.22 to 58.16 and P-IoU from 37.85 to 41.26, showing that explicit paired temporal evidence provides substantial benefit over single-branch reasoning. Adding the residual temporal branch further raises performance to 61.24 P-F1 and 44.65 P-IoU, indicating that feature-level discrepancy modeling contributes additional change-sensitive evidence. Finally, confidence-aware fusion yields the best overall performance, improving over the residual-branch variant from 98.21 to 98.88 in I-AUC, from 95.54 to 96.12 in P-AUC, from 61.24 to 63.47 in P-F1, and from 44.65 to 47.09 in P-IoU.

6.2 Why Confidence-Aware Fusion Matters

A standard bi-temporal model may still behave unreliably if it assumes that all temporal difference is meaningful. Our ablation is therefore not only about performance gain, but also about *calibration of evidence usage*. Table 2 already shows that the full model consistently outperforms the confidence-free residual variant, while Table 3 confirms that this advantage becomes more pronounced under temporal corruption. This supports the claim that reliability-aware fusion is not merely an architectural add-on, but a useful mechanism for trustworthy remote sensing reasoning under imperfect inputs [14].

This point is particularly important for TrustVLM-style evaluation. A model that achieves similar average AUC but produces fewer unstable false positives under noisy temporal evidence is arguably more trustworthy than one that attains slightly higher peak scores while behaving erratically in difficult cases.

6.3 Optional Robustness Ablation Under Temporal Corruption

The robustness study in Table 3 provides direct evidence that the confidence-aware branch improves stability under imperfect temporal inputs. Injecting synthetic perturbations into T_1, such as translation, blur, or brightness shift, allows the evaluation to move beyond raw accuracy and toward reliability under corrupted temporal correspondence.

7 Conclusion

We presented **TMCD-RS**, a lightweight bi-temporal vision–language framework for trustworthy remote sensing change reasoning. Built upon a frozen CLIP backbone, TMCD-RS combines shared-weight Siamese temporal encoding, anomaly-aware text prototype learning, residual temporal evidence modeling, and confidence-aware

multi-scale fusion. Unlike conventional pipelines that treat temporal differencing as inherently reliable, the proposed framework explicitly models the trustworthiness of temporal discrepancy and adaptively calibrates how strongly such evidence influences the final decision.

This design leads to two key benefits. First, TMCD-RS preserves the open-vocabulary semantic prior and cross-domain generalization ability of pretrained vision–language models through parameter-efficient adaptation. Second, it provides a more reliable basis for remote sensing change analysis under imperfect temporal correspondence, thereby reducing false responses to pre-existing structures and unstable discrepancy signals.

More broadly, our results suggest that trustworthy bi-temporal change detection should not be framed solely as an accuracy optimization problem. Instead, reliable remote sensing VLMs should be evaluated by how well they generalize across domains, how robustly they behave under imperfect inputs, and how conservatively they respond when temporal evidence is uncertain. We believe TMCD-RS offers a practical step in this direction and provides a useful foundation for future work on trustworthy multimodal reasoning in remote sensing.

References

[1] Wele Gedara Chaminda Bandara and Vishal M. Patel. 2022. A Transformer-Based Siamese Network for Change Detection. *CoRR* abs/2201.01293 (2022). doi:10.48550/arXiv.2201.01293

[2] Qinglong Cao, Yuntian Chen, Chao Ma, and Xiaokang Yang. 2024. Open-Vocabulary Remote Sensing Image Semantic Segmentation. *CoRR* abs/2409.07683 (2024). doi:10.48550/arXiv.2409.07683

[3] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization. *CoRR* abs/2309.16020 (2023). doi:10.48550/arXiv.2309.16020

[4] H. Chen, Z. Qi, and Z. Shi. 2022. Remote Sensing Image Change Detection With Transformers. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–14.

[5] Hao Chen and Zhenwei Shi. 2020. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sensing* 12, 10 (2020), 1662.

[6] G. Cheng, Y. Huang, X. Li, S. Lyu, Z. Xu, H. Zhao, Q. Zhao, and S. Xiang. 2024. Change Detection Methods for Remote Sensing in the Last Decade: A Comprehensive Review. *Remote Sensing* 16, 13 (2024), 2355.

[7] R. C. Daudt, B. L. Saux, and A. Boulch. 2018. Fully Convolutional Siamese Networks for Change Detection. *2018 25th IEEE International Conference on Image Processing (ICIP)*, 4063–4067.

[8] Shunping Ji, Shiqing Wei, and Meng Lu. 2018. Fully Convolutional Networks for Multi-Source Building Extraction From an Open Aerial and Satellite Imagery Dataset. *IEEE Transactions on Geoscience and Remote Sensing* 57, 1 (2018), 574–586. doi:10.1109/TGRS.2018.2858817

[9] H. Jiang, M. Peng, H. Zhong, Y.and Xie, Z. Hao, J. Lin, X. Ma, and X. Hu. 2022. A Survey on Deep Learning-Based Change Detection from High-Resolution Remote Sensing Images. *Remote Sensing* 14, 7 (2022), 1552.

[10] Chenbin Liang, Weibin Li, Yunyun Dong, and Wenlin Fu. 2024. Single Domain Generalization Method for Remote Sensing Image Segmentation via Category Consistency on Domain Randomization. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–16. doi:10.1109/TGRS.2024.3379669

[11] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. 2024. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–16. doi:10.1109/TGRS.2024.3390838

[12] W. Ma, X. Zhang, Y. Li, Q. Yao, F. Tang, C. Wu, R. Yan, Z. Jiang, and S. K. Zhou. 2025. AA-CLIP: Enhancing Zero-Shot Anomaly Detection via Anomaly-Aware CLIP. *arXiv preprint arXiv:2503.06661* (2025).

[13] Utkarsh Mall, Cheng Perng Phoo, Meilin Kelsey Liu, Carl Vondrick, Bharath Hariharan, and Kavita Bala. 2023. Remote Sensing Vision-Language Foundation Models without Annotations via Ground Remote Alignment. *CoRR* abs/2312.06960 (2023). doi:10.48550/arXiv.2312.06960

[14] M. Minderer, J. Djolonga, R. Romijnders, et al. 2021. Revisiting the calibration of modern neural networks. *NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems* 1200 (2021), 15682–15694.

[15] A. Radford, J. W. Kim, C. Hallacy, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning* 139 (2021), 8748–8763.

[16] A. Radford, J. W. Kim, C. Hallacy, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning, PMLR* 139 (2021), 8748–8763.

[17] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan. 2020. Change Detection Based on Artificial Intelligence: State-of-the-Art and Challenges. *Remote Sensing* 12, 10 (2020), 1688.

[18] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. 2016. Recent Advances in Domain Adaptation for the Classification of Remote Sensing Data. *IEEE Geoscience and Remote Sensing Magazine* 4, 2 (2016), 41–57. doi:10.1109/MGRS.2016.2548504

[19] Yu Zhang and Zhi Gao. 2025. RSAD-CLIP: Zero-Shot Remote Sensing Anomaly Detection of the Earth's Surface Based on Pre-Trained Vision-Language Model. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 1–5.