# DTE-CAP: LEARNING DIRECTIONAL TEMPORAL EVIDENCE FOR REMOTE SENSING CHANGE CAPTIONING

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

## ABSTRACT

Remote Sensing Image Change Captioning (RSICC) aims to understand changes between bi-temporal remote sensing images and describe them in natural language. However, existing methods remain limited in modeling bi-temporal change representations. First, they lack explicit modeling of directional cross-temporal correspondences, making their predictions dependent on the input order established by the dataset and prone to errors when this order is reversed. Second, changes of different natures are often compressed into a single representation, hindering the language model from accessing structured change evidence with clear semantics. To address these issues, we propose DTE-Cap, a remote sensing change captioning framework for direction-aware temporal relation modeling and structured event representation. Specifically, we design a Directional Temporal Correspondence Modeling module that constructs mutually conditioned semantic references between the two observations, transforming conventional difference representations into relational representations with explicit temporal directionality. Building upon this, we propose a Transport-guided Event Structuring module that formulates cross-temporal semantic organization as an optimal transport problem. It explicitly derives persistent, appearing, and disappearing events from a unified transport coupling and integrates them into the base visual representation to guide caption generation. Extensive experiments and qualitative visualizations on three RSICC benchmarks demonstrate that DTE-Cap consistently outperforms existing methods, producing more accurate and detailed descriptions of visual changes.

## 1 INTRODUCTION

With the continued growth in the number of Earth observation satellites, large-scale multi-temporal remote sensing imagery is steadily accumulating Cong et al. (2022); Cao et al. (2023); Zhao & Yu (2025). Accurately understanding surface changes from these temporal observations is important for applications such as urban evolution analysis, land-use monitoring, and disaster assessment Lv et al. (2022); Yuan et al. (2021); Krejcar & Namazi (2026); Zhang et al. (2024). Remote Sensing Image Change Captioning (RSICC) aims to understand visual changes between bi-temporal remote sensing images and describe them in natural language. Compared with conventional change detection, which primarily focuses on localizing changed regions or identifying change categories Wu et al. (2024); Bai et al. (2023); Cheng et al. (2024); Sicong et al. (2026), RSICC further requires models to understand the semantic content of changes and generate corresponding descriptions, providing a more intuitive means of change understanding.

Existing RSICC methods mainly focus on visual change perception and language generation. Early methods typically use CNNs to extract visual features, followed by feature fusion and RNN- or SVM-based caption generation Chouaf et al. (2021); Hoxha et al. (2022). With the development of Transformers, RSICCformer$_c$ Liu et al. (2022) uses difference information to guide interactions between two feature branches. PSNet Liu et al. (2023a) models changes through progressive difference perception and multi-scale features. Sparse Focus Sun et al. (2024) employs sparse attention to capture changed regions with reduced computational complexity. SEN Zhou et al. (2024) adopts a single-stream extractor with contrastive pre-training to learn bi-temporal visual representations. More recent studies further expand the modeling paradigms. Diffusion-rscc Yu et al. (2025)



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

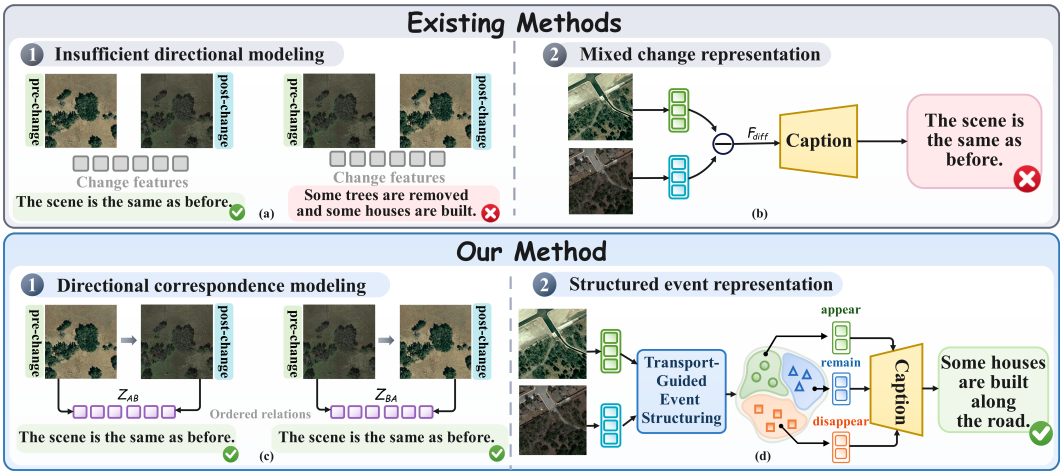


Figure 1: Limitations of existing methods and our solutions. (a) Reversing a no-change image pair leads to erroneous change predictions. (b) Unified change modeling overlooks local semantic changes. (c) Modeling directional cross-temporal correspondences improves temporal evolution understanding. (d) Structuring persistent, appearing, and disappearing events preserves distinct change semantics.

introduces diffusion-based generation to produce captions in a non-autoregressive manner, while Change3D Zhu et al. (2025) uses a video encoder to capture changes between bi-temporal images.

Despite this progress, existing RSICC methods still face two key challenges in modeling bi-temporal changes. First, they insufficiently model directional cross-temporal correspondences and do not adequately distinguish the two observations as pre-change and post-change states. This limits their understanding of temporal order. As illustrated in Figure 1(a), a model can correctly recognize a no-change image pair in the original input order but generate an erroneous change description when the order is reversed. This reveals a dependence on the established input order and an incomplete understanding of the temporal relations between the two observations. Second, as shown in Figure 1(b), existing methods typically compress changes of different natures into a single change representation. Persistent content, newly appearing objects, and disappearing objects are consequently mixed together. The language decoder must then implicitly recover these event roles from entangled visual features, leading to missed local changes and overly general or inaccurate descriptions.

To address these issues, we propose DTE-Cap, which enhances bi-temporal change understanding from two complementary perspectives: Directional Temporal Correspondence Modeling (DTCM) and Transport-guided Event Structuring (TES). Specifically, as illustrated in Figure 1(c), DTCM constructs mutually conditioned semantic references through sliding-window temporal attention. It separately characterizes semantic deviations of the pre-change and post-change observations from their respective cross-temporal references, and encodes them with ordered temporal roles to form direction-aware relational representations while retaining shared cross-temporal semantics. Building upon this, as shown in Figure 1(d), TES formulates cross-temporal semantic organization as an optimal transport problem. As change captioning requires identifying whether entities remain unchanged, newly appear, or disappear across temporal observations, TES explicitly derives persistent, appearing, and disappearing events from a unified transport coupling and integrates them into the base visual representation. Together, the two modules provide change evidence with explicit temporal directionality and distinct event roles to support accurate caption generation.

To comprehensively evaluate DTE-Cap, we conduct extensive experiments on three widely used RSICC benchmark datasets. Experimental results demonstrate that DTE-Cap achieves competitive performance across multiple evaluation metrics and consistently outperforms existing methods. The main contributions of this work are summarized as follows:

- We propose DTE-Cap, a direction-aware temporal relation modeling framework for remote sensing change captioning, which jointly models ordered cross-temporal correspondences and structured temporal event evidence to improve bi-temporal change understanding.

- We design a Directional Temporal Correspondence Modeling module that constructs bidirectional yet ordered cross-temporal semantic references, explicitly preserving directional correspondences between the two temporal states and endowing conventional difference representations with clear temporal orientation.
- We propose a Transport-guided Event Structuring module that formulates cross-temporal semantic organization as an optimal transport problem and explicitly derives persistent, appearing, and disappearing events from a unified transport coupling, providing structured change evidence with clear temporal semantics for caption generation.

## 2 Related Work

**Remote Sensing Image Change Captioning.** RSICC aims to understand semantic changes between bi-temporal remote sensing images and generate corresponding natural language descriptions. Chouaf et al. (2021) first introduced the RSICC task and developed an early encoder-decoder framework based on VGG16 Simonyan & Zisserman (2014) and an RNN Elman (1990) decoder for change description generation. Subsequently, Hoxha et al. (2022) constructed several RSICC datasets and explored different feature fusion strategies to improve change description generation. With the development of Transformer architectures Vaswani et al. (2017), recent studies have increasingly focused on cross-temporal feature interaction and change representation learning. Liu et al. (2022) proposed RSICCformer$_c$, which leverages difference information to guide the interaction between two temporal branches for enhanced change perception. PSNet Liu et al. (2023a) further introduced progressive difference modeling and multi-scale feature fusion to improve the representation of changes at different scales. More recently, with the emergence of pretrained models and advanced generative architectures, RSICC studies have explored more diverse modeling paradigms. Prompt-CC Liu et al. (2023b) incorporates pretrained vision models and large language models through prompt learning to enhance change understanding. Diffusion-rscc Yu et al. (2025) introduces diffusion models for change description generation, while Change3D Zhu et al. (2025) reformulates RSICC from a video modeling perspective and utilizes video encoders to capture variations between bi-temporal observations. However, existing methods generally represent bi-temporal relations as unified change features, lacking explicit modeling of directional temporal relations and structured change events.

**Optimal Transport Learning.** Optimal Transport (OT) has recently been widely explored for learning structured correspondences between different data distributions Gabriel & Marco (2019). Unlike conventional matching strategies based on fixed assignments, OT seeks an optimal transport plan between two distributions, enabling flexible soft correspondences and capturing underlying structural relationships Cuturi (2013). In vision-language learning, OT has been applied to enhance fine-grained cross-modal correspondence modeling. For example, UNITER employs OT to establish region-word alignments between visual and textual representations, improving multimodal representation learning Chen et al. (2020). Sarlin et al. (2020) combines graph neural networks with OT to learn correspondences between local features across images. More recently, OT-CLIP incorporates OT into the CLIP framework by reformulating image-text matching as a transport problem to enhance cross-modal understanding Shi et al. (2024). Han et al. (2025) further exploit OT to address noisy correspondence in vision-language models. Inspired by these advances, we exploit OT as a structured correspondence modeling tool to establish semantic transitions between bi-temporal observations and derive temporally meaningful event evidence for remote sensing change captioning.

## 3 Method

### 3.1 Overview

Given bi-temporal images $I_A$ (pre-change) and $I_B$ (post-change) acquired in chronological order, remote sensing change captioning aims to generate a textual sequence $Y = \{y_t\}_{t=1}^{T}$ that describes the semantic changes between the two observations. DTE-Cap models bi-temporal changes through directional temporal correspondence and structured event organization. As illustrated in Figure 2, the visual encoder extracts features $F_A, F_B$. Directional Temporal Correspondence Modeling constructs a directional relation representation $Z_{AB}$ and shared semantic context $P_{AB}$, which jointly

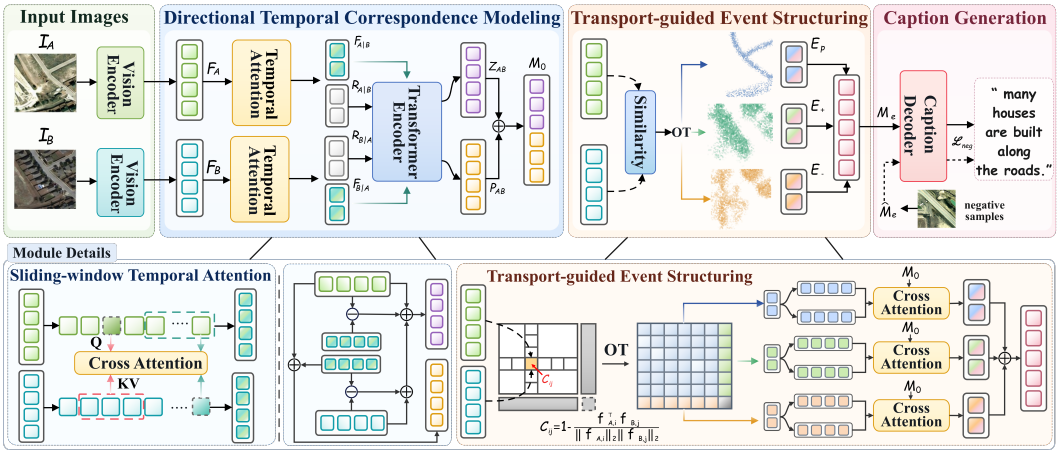


Figure 2: Overview of DTE-Cap. DTCM establishes directional correspondences through sliding-window temporal attention. TES constructs the semantic cost matrix $C$ from cross-temporal feature similarity and uses optimal transport to derive persistent ($E_p$), appearing ($E_+$), and disappearing ($E_-$) evidence, visualized as event-specific feature clusters. During training, mismatched event evidence is introduced as negative supervision to ground caption generation in valid temporal changes.

form the base visual representation $M_0$. Transport-guided Event Structuring further derives persistent, appearing, and disappearing evidence $E_p$, $E_+$, and $E_-$ from a unified transport coupling and incorporates them into $M_0$ to obtain the event-aware representation $M_e$. Finally, Evidence-Grounded Caption Generation uses $M_e$ as visual evidence to generate the change caption.

### 3.2 Directional Temporal Correspondence Modeling

We propose Directional Temporal Correspondence Modeling (DTCM) to capture directional temporal relations between bi-temporal observations. Given visual features $F_A, F_B \in \mathbb{R}^{N\times d}$ extracted from the two observations, DTCM establishes ordered cross-temporal correspondences and learns directional relation representations.

**Sliding-window Temporal Attention.** Bi-temporal remote sensing observations exhibit strong spatial correlation. We therefore propose sliding-window temporal attention to aggregate semantic references within a sliding window around the corresponding position in the other observation. Taking the post-change observation $I_B$ as the query, its pre-change reference is computed as

$$F_{B|A} = \text{Softmax}\left(\frac{Q_B K_A^\top}{\sqrt{d_k}} + B_w\right) V_A, \tag{1}$$

where $Q_B = F_B W_Q$, $K_A = F_A W_K$, and $V_A = F_A W_V$. For each query position $i$, the window mask $B_w$ retains only pre-change positions within the $w \times w$ neighborhood $\mathcal{N}w(i)$ centered at its corresponding location, and masks all others with negative infinity. The window thus shifts with the query position, yielding cross-temporal correspondence. Exchanging the query and key-value roles yields $F_{A|B}$. Please refer to Appendix A.1 for more details.

**Directional Relation Representation.** To characterize changes associated with the pre-change and post-change states, we compute directional residuals between each observation and its cross-temporal semantic reference:

$$R_{B|A} = F_B - F_{B|A}, \qquad R_{A|B} = F_A - F_{A|B}. \tag{2}$$

Here, $R_{B|A}$ and $R_{A|B}$ encode post-change and pre-change semantic deviations relative to their respective references. These complementary residuals provide change evidence associated with distinct temporal roles.

We concatenate the original features and directional residuals in a fixed order and jointly encode them as $Z_{AB} = \mathcal{E}([F_A, F_B, R_{B|A}, R_{A|B}])$, where brackets denote feature concatenation and $\mathcal{E}$ is a

Transformer-based encoding function. This ordered encoding integrates visual content and semantic deviations into a direction-aware relation representation. To complement the directional evidence with shared scene context, we construct the additive representation $\tilde{P}_{AB} = F_A + F_{A|B}$ and refine it through feature fusion and confidence weighting to obtain $P_{AB}$. Finally, we concatenate the shared context and directional representation along the sequence dimension to form the base visual representation $M_0 = [P_{AB}; Z_{AB}]$.

### 3.3 Transport-guided Event Structuring

Matching visual content across the two observations provides a basis for distinguishing temporal events: reliably matched content indicates persistence, whereas content supported only by the pre-change or post-change observation provides evidence for disappearance or appearance, respectively. We therefore propose Transport-guided Event Structuring (TES), which leverages optimal transport (OT) to derive structured event evidence from a unified coupling.

**Temporal Semantic Transport.** Given the bi-temporal features $F_A, F_B$, we construct a semantic cross-temporal cost matrix $C$ in a shared transport space, as illustrated in Fig. 2. Lower costs indicate stronger semantic consistency. To accommodate appearing and disappearing content, we augment each transport domain with an auxiliary unmatched component that absorbs semantic mass without reliable correspondence. Matched and unmatched assignments are jointly determined by the entropy-regularized optimal transport problem:

$$\Pi^* = \underset{\Pi \in U(\tilde{a}, \tilde{b})}{\arg\min} \langle \Pi, \widetilde{C} \rangle - \varepsilon H(\Pi), \tag{3}$$

where $\Pi$ denotes the transport coupling, and $U(\tilde{a}, \tilde{b})$ is the set of feasible couplings over the augmented domains. $\widetilde{C}$ augments $C$ with costs associated with the unmatched components, $H$ denotes entropy, and $\varepsilon$ denotes the transport temperature, which controls the strength of entropy regularization. The entropy term encourages a smooth transport structure and enables differentiable optimization through Sinkhorn iterations Cuturi (2013). The marginal constraints jointly regulate semantic assignments across tokens, while the unmatched components prevent unsupported content from being forced into unrelated correspondences. Please refer to Appendix A.2 for more details.

**Event Evidence Decomposition.** The optimal transport coupling has the following block structure:

$$\Pi^* = \begin{bmatrix} \Pi_p & \pi_d \\ \pi_a^\top & \pi_u \end{bmatrix}, \tag{4}$$

where $\Pi_p \in \mathbb{R}^{N \times N}$ represents transport between real visual tokens, $\pi_d \in \mathbb{R}^N$ is the pre-change mass assigned to the post-change unmatched component, and $\pi_a \in \mathbb{R}^N$ is the post-change mass received from the pre-change unmatched component. The scalar $\pi_u$ maintains mass balance in the augmented space. This structure provides a shared correspondence basis for the three event roles. The matched component $\Pi_p$ captures semantics preserved across observations. The disappearance component $\pi_d$ captures content unsupported by the post-change observation, providing disappearance evidence. Conversely, the appearance component $\pi_a$ captures post-change content unexplained by the pre-change observation, providing appearance evidence. The three transport components are then used to construct event-specific visual evidence. The matched transport $\Pi_p$ aggregates mutually supported features from the two observations to form persistent evidence $E_p$. The appearance mass $\pi_a$ weights features in $F_B$ that are insufficiently explained by the pre-change observation, yielding appearing evidence $E_+$. Similarly, the disappearance mass $\pi_d$ weights features in $F_A$ that are unsupported by the post-change observation, yielding disappearing evidence $E_-$. Thus, all three evidence banks are derived from the same transport coupling while preserving distinct temporal roles. Please refer to Appendix A.2 for more details.

**Event Evidence Fusion.** To make distinct event roles directly accessible to caption generation, we retain $E_p, E_+, E_-$ as separate evidence banks and associate each with a learnable event-role embedding. Using $M_0$ as the query, we apply multi-head cross-attention to each bank separately, with role-embedded evidence as keys and original evidence as values. We sum the resulting responses and add them to $M_0$ through a residual connection to obtain $M_e$. The resulting representation $M_e$

integrates DTCM-based temporal relations with TES-derived event evidence and serves as the final visual representation for caption generation.

### 3.4 Evidence-Grounded Caption Generation

To enhance the dependence of caption generation on event evidence, we propose an evidence grounding loss to optimize the caption decoder based on the standard captioning loss $\mathcal{L}_{\text{cap}}$. This loss introduces contrastive supervision between correct and mismatched event evidence, encouraging the decoder to generate captions grounded in valid change evidence. Specifically, for each training sample, we select candidate negatives with opposite event polarities from the same mini-batch and filter out potential false negatives. The selected negative samples provide mismatched event evidence $\hat{M}_e$, while the original event evidence is denoted as $M_e$. The evidence grounding loss is formulated as:

$$\mathcal{L}_{\text{neg}} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \left[ m - S(Y_i, M_{e,i}) + S(Y_i, \widehat{M}_{e,i}) \right]_+ . \tag{5}$$

where $S(Y, M_e)$ denotes the caption matching score between the target caption and evidence $M_e$, computed as the negative mean token-level cross-entropy. $\mathcal{A}$ represents the set of changed samples with valid negative evidence, $m$ is the predefined margin, and $[x]_+ = \max(0, x)$. When no valid negative evidence is available in the current mini-batch, $\mathcal{L}_{\text{neg}}$ is set to zero. This loss encourages higher consistency between the correct event evidence and the target caption while suppressing caption generation induced by mismatched evidence. During inference, the decoder directly utilizes the event evidence $M_e$ from the input image pair to generate the final change caption.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets.** We evaluate DTE-Cap on three widely used RSICC benchmarks: LEVIR-CC Liu et al. (2022), DUBAI-CC Hoxha et al. (2022), and WHU-CDC Ji et al. (2018). LEVIR-CC contains 10,077 bi-temporal image pairs covering diverse urban changes. DUBAI-CC includes 500 image pairs depicting urban evolution, while WHU-CDC contains 7,434 pairs of high-resolution aerial images. Each image pair in these datasets is annotated with five change descriptions.

**Baselines.** To comprehensively evaluate the effectiveness of DTE-Cap, we compare it with a broad set of representative RSICC methods, including Capt-Rep-Diff, Capt-Att, Capt-Dual-Att and DUDA Park et al. (2019); CC-RNN-sub and CC-SVM-sub Hoxha et al. (2022); MCCFormer-S and MCCFormer-D Qiu et al. (2021); RSICCformer$_c$ Liu et al. (2022); PSNet Liu et al. (2023a); Prompt-CC Liu et al. (2023b); Sparse Focus Sun et al. (2024); SEN Zhou et al. (2024); Change3D Zhu et al. (2025) and Diffusion-rscc Yu et al. (2025). These baselines span early encoder-decoder architectures, attention- and Transformer-based methods, and recent approaches using pretrained models, diffusion models, and video encoders.

**Implementation Details.** We implement DTE-Cap in PyTorch with a shared pretrained ResNet-50 encoder and a 512-dimensional visual embedding space. DTCM uses a matching dimension of 128 and a $5 \times 5$ sliding window. TES uses a 128-dimensional transport space with 7 Sinkhorn iterations and a transport temperature of 0.03. The caption decoder consists of one Transformer layer with 8 attention heads and a dropout rate of 0.5. We train the model using AdamW for up to 48 epochs. The initial learning rates are set to $3 \times 10^{-5}$ for the visual encoder and $10^{-4}$ for the temporal modules and caption decoder. Baseline results are adopted from previously reported evaluations under the corresponding benchmark settings when available. For methods without released model weights, we retrain them using the authors' released code and the training protocol described in the original papers. For a consistent decoding setting, the beam size is set to 1 for DTE-Cap and all autoregressive baselines during inference. All experiments are conducted on an NVIDIA RTX 4090 GPU. Please refer to Appendix B for additional hyperparameter analysis.

Table 1: Comparison with representative RSICC methods on the DUBAI-CC test set.

| Method | B-1 | B-2 | B-3 | B-4 | M | R | C | $S_m^*$ |
|---|---|---|---|---|---|---|---|---|
| CC-RNN-sub | 67.19 | 52.23 | 40.00 | 28.54 | 25.51 | 51.78 | 69.73 | 43.89 |
| CC-SVM-sub | 70.71 | 57.58 | 46.10 | 35.50 | <u>27.60</u> | 56.61 | 82.96 | 50.67 |
| DUDA | 58.82 | 43.59 | 33.63 | 25.39 | 22.05 | 48.34 | 62.78 | 39.64 |
| MCCFormer-S | 52.97 | 37.02 | 27.62 | 22.57 | 18.64 | 43.29 | 53.81 | 34.58 |
| MCCFormer-D | 64.65 | 50.45 | 39.36 | 29.48 | 25.09 | 51.27 | 66.51 | 43.09 |
| RSICCformer$_c$ | 67.92 | 53.61 | 41.37 | 31.28 | 25.41 | 51.96 | 66.54 | 43.80 |
| Prompt-CC | 70.03 | <u>58.41</u> | <u>49.44</u> | <u>40.32</u> | 26.48 | 55.82 | 85.44 | 52.01 |
| Sparse Focus | 67.30 | 55.97 | 47.00 | 37.30 | 26.32 | 56.38 | 91.59 | <u>52.90</u> |
| Change3D | 70.54 | 57.30 | 46.12 | 36.14 | 26.25 | <u>57.95</u> | 82.23 | 50.64 |
| Diffusion-rscc | - | - | - | 33.30 | 27.40 | 56.50 | 88.70 | 51.48 |
| SEN | <u>70.95</u> | 57.28 | 45.81 | 36.25 | 26.62 | 55.95 | <u>91.77</u> | 52.65 |
| **DTE-Cap (Ours)** | **77.82** | **65.43** | **53.47** | **42.19** | **30.44** | **61.44** | **111.17** | **61.31** |

Table 2: Comparison with representative RSICC methods on the WHU-CDC test set.

| Method | B-1 | B-2 | B-3 | B-4 | M | R | C | $S_m^*$ |
|---|---|---|---|---|---|---|---|---|
| Capt-Rep-Diff | 70.59 | 59.02 | 50.70 | 45.33 | 32.29 | 63.44 | 108.07 | 62.28 |
| Capt-Att | 75.62 | 64.68 | 56.89 | 50.62 | 33.97 | 67.17 | 118.33 | 67.52 |
| Capt-Dual-Att | 77.15 | 68.47 | 60.81 | 55.42 | 34.22 | 68.04 | 121.45 | 69.78 |
| DUDA | 79.04 | 69.53 | 61.57 | 55.64 | 34.29 | 68.98 | 121.85 | 70.19 |
| MCCFormer-S | 81.12 | <u>75.04</u> | <u>69.95</u> | 65.34 | 42.11 | <u>78.52</u> | <u>147.09</u> | <u>83.27</u> |
| MCCFormer-D | 73.29 | 67.88 | 64.03 | 60.96 | 39.69 | 73.67 | 134.92 | 77.31 |
| RSICCformer$_c$ | 80.05 | 74.24 | 69.61 | <u>66.54</u> | <u>42.65</u> | 73.91 | 133.44 | 79.14 |
| PSNet | <u>81.26</u> | 73.25 | 65.78 | 60.32 | 36.97 | 71.60 | 130.52 | 74.85 |
| Prompt-CC | 81.12 | 73.96 | 67.22 | 61.45 | 36.99 | 71.88 | 134.50 | 76.21 |
| Sparse Focus | 81.17 | 72.90 | 66.06 | 60.27 | 37.34 | 72.63 | 134.64 | 76.22 |
| Diffusion-rscc | 75.32 | 70.15 | 66.40 | 63.76 | 40.18 | 73.80 | 127.96 | 76.43 |
| SEN | 80.60 | 74.64 | 67.69 | 61.97 | 36.76 | 71.70 | 133.57 | 76.00 |
| **DTE-Cap (Ours)** | **84.09** | **78.78** | **74.45** | **71.26** | **45.33** | **79.75** | **149.71** | **86.51** |

**Metrics.** Following the standard RSICC evaluation protocol Liu et al. (2022), we adopt BLEU-1 to BLEU-4, METEOR, ROUGE-L, and CIDEr as evaluation metrics, denoted as B-1 to B-4, M, R, and C, respectively. We further report $S_m^*$, the arithmetic mean of BLEU-4, METEOR, ROUGE-L, and CIDEr. Higher scores indicate better performance.

### 4.2 Performance Comparison

**Results on DUBAI-CC.** As shown in Table 1, DTE-Cap achieves the best performance across all metrics, reaching 42.19 BLEU-4, 30.44 METEOR, 61.44 ROUGE-L, and 111.17 CIDEr, with an overall score of $S_m^* = 61.31$. **Results on WHU-CDC.** Table 2 shows that our method consistently outperforms previous approaches across all evaluation metrics, achieving 71.26 BLEU-4, 45.33 METEOR, 79.75 ROUGE-L, and 149.71 CIDEr, with an overall score of $S_m^* = 86.51$. **Results on LEVIR-CC.** As shown in Table 3, our method obtains the best overall score of $S_m^* = 79.46$ and achieves the best performance on most metrics, including 65.71 BLEU-4, 40.14 METEOR, and 75.05 ROUGE-L. Overall, DTE-Cap achieves consistent improvements across the three benchmarks, demonstrating its effectiveness for remote sensing change captioning.

### 4.3 Ablation Studies

We evaluate the contribution of each component in Table 4. **Directional Temporal Correspondence Modeling.** Removing $Z_{AB}$ reduces BLEU-4 and CIDEr by 4.52 and 12.93 points, respectively, demonstrating the importance of explicitly encoding directional temporal relations. Remov-

Table 3: Comparison with representative RSICC methods on the LEVIR-CC test set.

| Method | B-1 | B-2 | B-3 | B-4 | M | R | C | $S_m^*$ |
|---|---|---|---|---|---|---|---|---|
| Capt-Rep-Diff | 72.90 | 61.98 | 53.62 | 47.41 | 34.47 | 65.64 | 110.57 | 64.52 |
| Capt-Att | 77.64 | 67.40 | 59.24 | 53.15 | 36.58 | 69.73 | 121.22 | 70.17 |
| Capt-Dual-Att | 79.51 | 70.57 | 63.23 | 57.46 | 36.56 | 70.69 | 124.42 | 72.28 |
| DUDA | 81.44 | 72.22 | 64.24 | 57.79 | 37.15 | 71.04 | 124.32 | 72.58 |
| MCCFormer-S | 79.90 | 70.26 | 62.68 | 56.68 | 36.17 | 69.46 | 120.39 | 70.68 |
| MCCFormer-D | 80.42 | 70.87 | 62.86 | 56.38 | 37.29 | 70.32 | 124.44 | 72.11 |
| RSICCformer$_c$ | 83.09 | 74.32 | 66.66 | 60.44 | 38.76 | 72.63 | 130.00 | 75.46 |
| PSNet | 83.86 | 75.13 | 67.89 | 62.11 | 38.80 | 73.60 | 132.62 | 76.78 |
| Prompt-CC | 83.66 | 75.73 | 69.10 | 63.54 | 38.83 | 73.72 | 136.44 | 78.13 |
| Sparse Focus | 84.56 | 75.87 | 68.64 | 62.87 | <u>39.93</u> | <u>74.69</u> | **137.05** | <u>78.64</u> |
| Change3D | <u>85.66</u> | <u>77.25</u> | 69.76 | 63.53 | 39.50 | 74.41 | 134.53 | 77.99 |
| Diffusion-rscc | - | - | - | 60.90 | 37.80 | 71.50 | 125.60 | 73.95 |
| SEN | 85.10 | 77.05 | <u>70.01</u> | <u>64.09</u> | 39.59 | 74.57 | 136.02 | 78.57 |
| **DTE-Cap (Ours)** | **85.70** | **77.70** | **71.04** | **65.71** | **40.14** | **75.05** | <u>136.92</u> | **79.46** |

Table 4: Ablation results on the LEVIR-CC validation set.

| No. | Method | B-1 | B-2 | B-3 | B-4 | M | R | C |
|---|---|---|---|---|---|---|---|---|
| 1 | w/o $Z_{AB}$ | 81.78 | 73.32 | 66.69 | 61.27 | 37.58 | 71.06 | 124.96 |
| 2 | w/o $P_{AB}$ | 83.72 | 76.63 | 70.54 | 65.17 | 39.57 | 74.07 | 135.72 |
| 3 | w/o DTCM | 80.14 | 72.15 | 66.00 | 60.91 | 37.44 | 70.47 | 125.55 |
| 4 | w/o TES | 83.85 | 76.21 | 69.59 | 64.02 | 39.50 | 73.98 | 135.50 |
| 5 | w/o $\mathcal{L}_{\text{neg}}$ | 84.38 | 76.78 | 70.44 | 65.05 | 39.87 | 74.37 | 136.97 |
| **DTE-Cap (Ours)** | | **84.43** | **77.23** | **71.06** | **65.79** | **40.36** | **74.89** | **137.89** |

ing $P_{AB}$ results in a CIDEr decrease of 2.17 points, indicating that shared scene context provides complementary information. When the entire DTCM is removed, BLEU-4 and CIDEr decrease by 4.88 and 12.34 points, further validating the effectiveness of directional temporal correspondence modeling. **Transport-guided Event Structuring.** Without TES, BLEU-4 decreases from 65.79 to 64.02 and CIDEr drops from 137.89 to 135.50. These results demonstrate the benefit of explicitly organizing persistent, appearing, and disappearing evidence before caption generation. **Negative-sample Constraint.** Removing $\mathcal{L}_{\text{neg}}$ decreases BLEU-4 and CIDEr by 0.74 and 0.92 points, respectively, indicating that mismatched event evidence helps strengthen evidence-dependent caption generation.

### 4.4 Analysis of Temporal Relations and Event Evidence

**#1: Temporal Direction.** Figure 3 evaluates temporal modeling through reversal robustness on no-change samples and directional discrimination on changed samples. For no-change samples, swapping the observations should preserve their semantics. As shown in Figure 3 (a,b), the full model outperforms SEN under both original and reversed temporal orders, with a 33.53-point BLEU-4 advantage under reversal. Removing $R_{A|B}$ and $R_{B|A}$, individually or jointly, reduces performance under reversal, suggesting that both directional residuals help alleviate dependence on the dataset-established input order. For changed samples, Figure 3 (c) further examines whether the representations encode directional event semantics. The full representation $Z_{AB}$ achieves the best or tied-best results across all five metrics. Removing $R_{A|B}$ degrades some metrics, while removing $R_{B|A}$ causes larger drops, particularly in PBC. Here, A-R and D-R denote appearance and disappearance recall, respectively; Acc and M-F1 measure classification performance, while PBC measures whether event predictions are consistently reversed after swapping the temporal order. These results support the joint contribution of both residuals to temporal modeling: explicitly capturing semantic deviations of the pre-change and post-change states from their cross-temporal references improves both reversal robustness on no-change samples and directional event discrimination on changed samples.

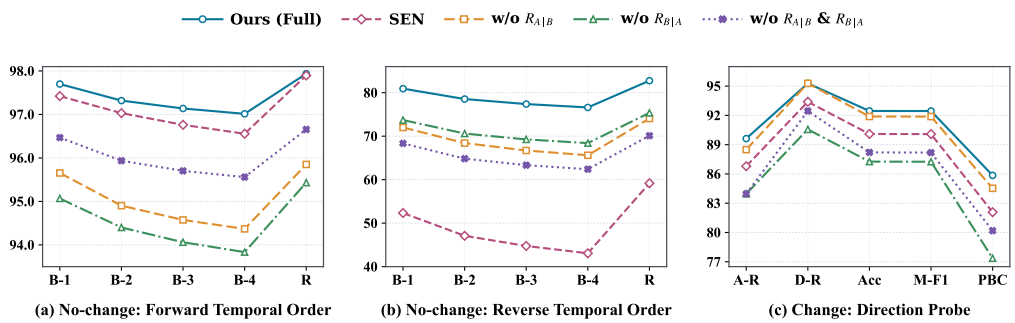


Figure 3: Analysis of temporal direction. (a) and (b) Comparison of change captioning results on no-change samples under the original and reversed temporal orders, respectively. (c) Comparison of appearance/disappearance discrimination using the learned visual representations, where our method leverages the directional representation $Z_{AB}$ produced by DTCM.

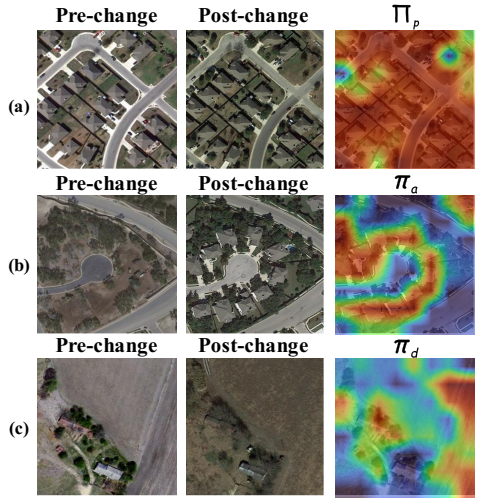


Figure 4: Visualization of transport-induced event evidence. $\Pi_p$, $\pi_a$, and $\pi_d$ highlight persistent, appearing, and disappearing content, respectively.

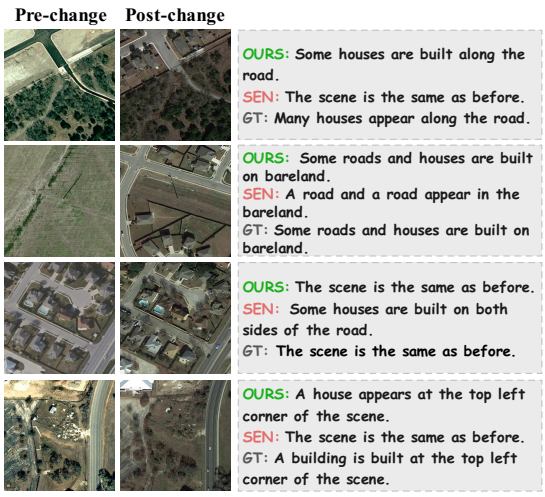


Figure 5: Qualitative comparison of change captions. Ours captures real changes more accurately while reducing hallucinated changes in no-change scenes.

**#2: Event Evidence.** Figure 4 visualizes the transport components induced by TES under different temporal events. The matched transport $\Pi_p$ mainly responds to regions preserved across the two observations, while $\pi_a$ and $\pi_d$ focus on newly appearing and disappearing content, respectively. These visualizations show that TES organizes the transport coupling into event-specific evidence with distinct temporal semantics.

### 4.5 Case Study

Figure 5 presents qualitative results. In changed scenes, our method better captures newly appearing objects and generates descriptions closer to the ground truth, while in no-change scenes it effectively avoids hallucinated changes. These examples demonstrate the reliability of our method across both change and no-change scenarios. Please refer to Appendix C for additional qualitative results.

## 5 Conclusion

We propose DTE-Cap for remote sensing change captioning, which combines directional temporal correspondence modeling with transport-guided event structuring to preserve ordered tempo-

ral relations and organize persistent, appearing, and disappearing evidence. Experiments on three benchmarks demonstrate consistent improvements over existing methods, highlighting the value of explicit temporal direction and structured event evidence for remote sensing change understanding. By converting implicit bi-temporal differences into directional relations and structured event evidence, DTE-Cap provides a more effective way to capture semantic evolution between remote sensing observations.

### AI USE STATEMENT

We used generative AI tools to assist with language editing to improve the clarity and readability of the manuscript; to support relevant literature searches for a better understanding of developments and prior work in remote sensing change captioning, temporal relation modeling, and related areas; and to help organize the research content through discussions of research ideas and suggestions for conceptual illustrations, thereby improving the presentation of the proposed method. The authors reviewed and revised all AI-assisted content, verified the accuracy and relevance of the references, and checked the consistency of the text and figures with the actual method. The authors take full responsibility for the accuracy, originality, and academic integrity of the final manuscript.

### ETHICS STATEMENT

This study does not involve human subjects, personally identifiable information, or sensitive personal data. All experiments are conducted using publicly available remote sensing change captioning benchmark datasets, including LEVIR-CC, DUBAI-CC, and WHU-CDC. No additional human-subject data collection or processing of personal information is involved. The data are used solely for academic research and model evaluation related to remote sensing image change understanding and caption generation. We follow appropriate research practices and principles of research integrity throughout data use, experimental design, result analysis, and manuscript preparation.

### REPRODUCIBILITY STATEMENT

This paper provides the necessary methodological and experimental information for reproducing the reported results throughout the main text and supplementary materials. Section 3 presents the proposed DTE-Cap framework in detail, including directional temporal correspondence modeling, transport-guided event structuring, and evidence-grounded caption generation. Section 4.1 describes the experimental datasets, evaluation metrics, baseline configurations, and implementation details, including model settings, training strategies, and inference procedures. In addition, the supplementary materials provide further details of key components to facilitate a better understanding and reproduction of the proposed method. All experiments are conducted on publicly available remote sensing change captioning datasets, and the corresponding data processing procedures and experimental settings are described in the paper. These descriptions aim to support researchers in reproducing our method and validating the reported experimental results.

## REFERENCES

Ting Bai, Le Wang, Dameng Yin, Kaimin Sun, Yepei Chen, Wenzhuo Li, and Deren Li. Deep learning for change detection in remote sensing: a review. *Geo-spatial Information Science*, 26 (3):262–288, 2023.

Yinxia Cao, Xin Huang, and Qihao Weng. A multi-scale weakly supervised learning method with adaptive online noise correction for high-resolution change detection of built-up areas. *Remote Sensing of Environment*, 297:113779, 2023.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

Guangliang Cheng, Yunmeng Huang, Xiangtai Li, Shuchang Lyu, Zhaoyang Xu, Hongbo Zhao, Qi Zhao, and Shiming Xiang. Change detection methods for remote sensing in the last decade: A comprehensive review. *Remote Sensing*, 16(13):2355, 2024.

Seloua Chouaf, Genc Hoxha, Youcef Smara, and Farid Melgani. Captioning changes in bi-temporal remote sensing images. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 2891–2894. IEEE, 2021.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Haochen Han, Alex Jinpeng Wang, Peijun Ye, and Fangming Liu. Unlearning the noisy correspondence makes clip more robust. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4518–4528. IEEE, 2025.

Genc Hoxha, Seloua Chouaf, Farid Melgani, and Youcef Smara. Change captioning: A new paradigm for multitemporal remote sensing image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

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, 2018.

Ondrej Krejcar and Hamidreza Namazi. Ai in remote sensing and satellite image processing-a review. *Environmental Earth Sciences*, 85(3):78, 2026.

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

Chenyang Liu, Jiajun Yang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. Progressive scale-aware network for remote sensing image change captioning. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6668–6671. IEEE, 2023a.

Chenyang Liu, Rui Zhao, Jianqi Chen, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. A decoupling paradigm with prompt learning for remote sensing image change captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18, 2023b.

ZhiYong Lv, HaiTao Huang, Xinghua Li, MingHua Zhao, Jon Atli Benediktsson, WeiWei Sun, and Nicola Falco. Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective. *Proceedings of the IEEE*, 110(12):1976–1991, 2022.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4623–4632. IEEE, 2019.

Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. Describing and localizing multiple changes with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1951–1960. IEEE, 2021.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4937–4946. IEEE, 2020.

Liangliang Shi, Jack Fan, and Junchi Yan. Ot-clip: Understanding and generalizing clip via optimal transport. In *Forty-first International Conference on Machine Learning*, 2024.

LIU Sicong, DU Kecheng, ZHENG Yongjie, CHEN Jin, DU Peijun, and TONG Xiaohua. Remote sensing change detection technology in the era of artificial intelligence: Inheritance, development and challenges. *National Remote Sensing Bulletin*, 27(9):1975–1987, 2026.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Dongwei Sun, Yajie Bao, Junmin Liu, and Xiangyong Cao. A lightweight sparse focus transformer for remote sensing image change captioning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:18727–18738, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Chen Wu, Liangpei Zhang, Bo Du, Hongruixuan Chen, Jingxuan Wang, and Huan Zhong. Unet-like remote sensing change detection: A review of current models and research directions. *IEEE Geoscience and Remote Sensing Magazine*, 12(4):305–334, 2024.

Xiaofei Yu, Yitong Li, Jie Ma, Chang Li, and Hanlin Wu. Diffusion-rscc: Diffusion probabilistic model for change captioning in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025.

Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert systems with applications*, 169:114417, 2021.

Xiao Zhang, Tingting Zhao, Hong Xu, Wendi Liu, Jinqing Wang, Xidong Chen, and Liangyun Liu. Glc_fcs30d: the first global 30 m land-cover dynamics monitoring product with a fine classification system for the period from 1985 to 2022 generated using dense-time-series landsat imagery and the continuous change-detection method. *Earth System Science Data*, 16(3):1353–1381, 2024.

Qiang Zhao and Le Yu. Advancing sustainable development goals through earth observation satellite data: Current insights and future directions. *Journal of Remote Sensing*, 5:0403, 2025.

Qing Zhou, Junyu Gao, Yuan Yuan, and Qi Wang. Single-stream extractor network with contrastive pre-training for remote-sensing change captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

Duowang Zhu, Xiaohu Huang, Haiyan Huang, Hao Zhou, and Zhenfeng Shao. Change3d: Revisiting change detection and captioning from a video modeling perspective. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24011–24022. IEEE, 2025.

# A Appendix

## A.1 Details of Sliding-window Temporal Attention

This section details the local neighborhood, window mask, attention weights, and semantic reference aggregation used in sliding-window temporal attention. Let $F_A, F_B \in \mathbb{R}^{N\times d}$ denote the pre-change and post-change visual features, respectively, where each visual token corresponds to a spatial position in the feature map, $N$ is the number of positions, and $d$ is the feature dimension.

**Spatial Neighborhood and Window Mask.** For a query position $i$ in the post-change observation, let $\mathbf{p}_i = (x_i, y_i)$ denote its spatial coordinates. Given corresponding spatial grids across the two observations, we define a local neighborhood of width $w$ centered at the same coordinates in the pre-change feature map:

$$\mathcal{N}^w(i) = \{j : |x_j - x_i| \le r,\ |y_j - y_i| \le r\}, \tag{6}$$

where $r = (w-1)/2$, $w$ is odd, and $j$ ranges over valid pre-change positions. We use $w = 5$, so each query attends to at most 25 local candidates in the other observation. Only valid positions are retained at feature-map boundaries; padded positions are also masked when padding is used to extract neighborhoods. The additive window mask $B_w \in \mathbb{R}^{N\times N}$ is defined as

$$(B_w)_{ij} = \begin{cases} 0, & j \in \mathcal{N}^w(i), \\ -\infty, & j \notin \mathcal{N}^w(i). \end{cases} \tag{7}$$

Scores within the neighborhood remain unchanged, whereas those outside it are set to negative infinity and therefore receive zero weight after Softmax. The window center moves with the query position, and neighborhoods of adjacent queries may overlap. This local search preserves the spatial correspondence prior of bi-temporal imagery while allowing semantic references to be retrieved from nearby positions.

**Attention Weights and Reference Aggregation.** Using the post-change observation $I_B$ as the query, we retrieve semantic references from the pre-change observation $I_A$. Queries, keys, and values are obtained through the linear projections $Q_B = F_B W_Q$, $K_A = F_A W_K$, and $V_A = F_A W_V$, respectively. Queries and keys lie in a $d_k$-dimensional matching space, while values carry the visual information to be aggregated. Denoting the corresponding vectors by $q_i^B$, $k_j^A$, and $v_j^A$, the local attention weights are

$$\alpha_{ij}^{B\leftarrow A} = \begin{cases} \dfrac{\exp\big((q_i^B)^\top k_j^A/\sqrt{d_k}\big)}{\sum_{\ell\in\mathcal{N}^w(i)}\exp\big((q_i^B)^\top k_\ell^A/\sqrt{d_k}\big)}, & j\in\mathcal{N}^w(i), \\ 0, & j\notin\mathcal{N}^w(i). \end{cases} \tag{8}$$

The dot product measures query–key compatibility, and $1/\sqrt{d_k}$ controls the score scale. Softmax is applied over candidate positions, so the weights sum to one for each query. Pre-change values are then aggregated as $f_{B|A,i} = \sum_{j\in\mathcal{N}^w(i)} \alpha_{ij}^{B\leftarrow A} v_j^A$. Collecting the outputs at all positions yields the matrix form used in the main text:

$$F_{B|A} = \mathrm{Softmax}\left(\frac{Q_B K_A^\top}{\sqrt{d_k}} + B_w\right) V_A. \tag{9}$$

The spatial indices of $F_{B|A}$ follow the post-change features $F_B$, whereas its aggregated content comes from the pre-change features $F_A$. It therefore represents pre-change semantic references indexed by post-change positions. The conditional notation in the subscript denotes this reference relationship. We set the matching dimension to $d_k = 128$.

### A.2 Details of Temporal Semantic Transport

This section explains the semantic cost matrix, unmatched augmentation, marginal constraints, and entropy regularization in temporal semantic transport, together with its optimization and event decomposition. Throughout this section, rows of the transport matrix correspond to the pre-change observation $I_A$, and columns correspond to the post-change observation $I_B$.

**Cross-temporal Semantic Cost.** We project the visual features from both observations into a shared transport space and denote the projected features by $u_i^A, u_j^B \in \mathbb{R}^{d_t}$. Following the cosine-based definition illustrated in Figure 2, the entries of the cost matrix $C \in \mathbb{R}^{N\times N}$ are

$$C_{ij} = 1 - \frac{(u_i^A)^\top u_j^B}{\|u_i^A\|_2 \|u_j^B\|_2}. \tag{10}$$

We use $d_t = 128$. Higher semantic similarity produces a lower cost. The matrix $C$ specifies the cost of each token pairing, whereas the transport coupling $\Pi$ specifies the mass assigned to that pairing, with mass denoting a nonnegative allocation weight.

**Unmatched Components and Cost Augmentation.** Balanced transport restricted to real visual tokens requires all mass to be assigned between the two sets of features. For appearing or disappearing content, this may force unsupported features into unrelated correspondences. We therefore introduce an auxiliary unmatched slot in each domain, denoted by $\mathrm{u}_A$ and $\mathrm{u}_B$, where u stands for unmatched. The augmented cost matrix is

$$\widetilde{C} = \begin{bmatrix} C & c_d \\ c_a^\top & c_u \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}. \tag{11}$$

Here, $c_d, c_a \in \mathbb{R}^N$ and $c_u \in \mathbb{R}$. The vector $c_d$ forms the additional column, with $(c_d)_i = \widetilde{C}_{i,N+1}$ specifying the cost of assigning pre-change token $i$ to the post-change unmatched slot $\mathrm{u}_B$. Similarly,

$c_a^\top$ forms the additional row, where $(c_a)_j = \widetilde{C}_{N+1,j}$ specifies the cost of assigning mass from the pre-change unmatched slot $\mathrm{u}_A$ to post-change token $j$. The scalar $c_u = \widetilde{C}_{N+1,N+1}$ defines the transport cost between the two auxiliary slots.

This augmentation incorporates real-token matches and unmatched assignments into the same optimization problem. Pre-change mass without post-change semantic support can be assigned to $\mathrm{u}_B$, while post-change content unexplained by the pre-change observation can receive mass from $\mathrm{u}_A$. Matched and unmatched assignments are thus determined jointly under the marginal constraints. Similar auxiliary unmatched slots have been used in visual matching Sarlin et al. (2020); TES further uses the resulting transport components to organize event evidence with explicit temporal roles.

**Marginals and Feasible Couplings.** Let $\widetilde{a} = [a_1, \ldots, a_N, a_{\mathrm{u}}]^\top$ and $\widetilde{b} = [b_1, \ldots, b_N, b_{\mathrm{u}}]^\top$ denote the augmented mass vectors. Their first $N$ entries specify the masses of real tokens, and their final entries specify the masses of the corresponding auxiliary slots. The two marginals must have equal total mass; for exposition, both can be normalized to unit total mass. The feasible coupling set is

$$\mathcal{U}(\widetilde{a}, \widetilde{b}) = \left\{ \Pi \in \mathbb{R}_+^{(N+1)\times(N+1)} : \Pi \mathbf{1}_{N+1} = \widetilde{a},\ \Pi^\top \mathbf{1}_{N+1} = \widetilde{b} \right\}, \tag{12}$$

where $\Pi$ is nonnegative and $\mathbf{1}_{N+1}$ is the all-ones vector. The two equalities constrain the total mass assigned by each row and received by each column, respectively, coupling the assignments across tokens. The marginal masses of the auxiliary slots also bound the mass that can pass through unmatched assignments. Together with the augmented costs, they regulate the balance between matched and unmatched transport.

**Matrix Inner Product and Entropy Regularization.** The optimization problem in the main text is

$$\Pi^* = \underset{\Pi \in \mathcal{U}(\widetilde{a}, \widetilde{b})}{\arg\min} \langle \Pi, \widetilde{C} \rangle - \varepsilon H(\Pi). \tag{13}$$

Here, $\arg\min$ returns the transport matrix minimizing the objective, $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product, and $H(\Pi)$ denotes the entropy of the coupling. The latter two terms are defined as

$$\begin{aligned} \langle \Pi, \widetilde{C} \rangle &= \sum_{i=1}^{N+1} \sum_{j=1}^{N+1} \Pi_{ij} \widetilde{C}_{ij}, \\ H(\Pi) &= -\sum_{i=1}^{N+1} \sum_{j=1}^{N+1} \Pi_{ij} \log \Pi_{ij}. \end{aligned} \tag{14}$$

The inner product sums the products of transport mass and cost at corresponding entries, yielding the total assignment cost. The entropy sums $-\Pi_{ij} \log \Pi_{ij}$ over all transport weights, where $\log$ is the natural logarithm and $0 \log 0 = 0$ by convention.

Because the objective contains $-\varepsilon H(\Pi)$, a positive $\varepsilon$ encourages smoother soft assignments under the marginal constraints. Smaller values place greater emphasis on low-cost correspondences, whereas larger values strengthen the smoothing effect. Substituting the entropy definition gives the equivalent objective $\sum_{i,j} \Pi_{ij} \widetilde{C}_{ij} + \varepsilon \sum_{i,j} \Pi_{ij} \log \Pi_{ij}$. An alternative entropy convention, $-\sum_{i,j} \Pi_{ij} (\log \Pi_{ij} - 1)$, differs only by a constant when total mass is fixed and therefore produces the same optimal coupling. These definitions follow standard entropy-regularized OT Cuturi (2013).

**Sinkhorn Optimization.** Let $K_{ij} = \exp(-\widetilde{C}_{ij}/\varepsilon)$ denote the elementwise exponential kernel. Lower costs yield larger kernel values. The optimal coupling admits the form $\Pi^* = \mathrm{diag}(s) K \,\mathrm{diag}(t)$, where $\mathrm{diag}(\cdot)$ places a vector on the diagonal of a matrix, and $s, t$ are nonnegative scaling vectors for the row and column masses. The marginal constraints yield the alternating updates

$$s^{(\ell+1)} = \widetilde{a} \oslash (K t^{(\ell)}), \qquad t^{(\ell+1)} = \widetilde{b} \oslash (K^\top s^{(\ell+1)}), \tag{15}$$

where $\ell$ indexes iterations and $\oslash$ denotes elementwise division. The scaling vectors can be initialized to all ones. The two updates adjust row and column sums in turn, progressively satisfying both marginal constraints Cuturi (2013).

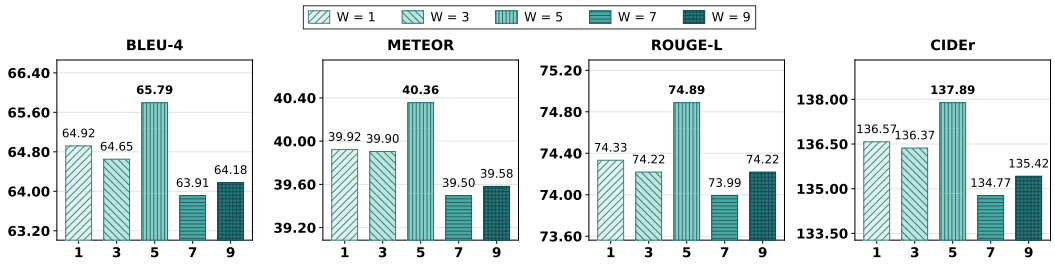


Figure 6: Performance comparison with different sliding-window sizes $w$. Here, $w$ denotes the side length of a $w \times w$ neighborhood. The four panels report BLEU-4, METEOR, ROUGE-L, and CIDEr, respectively.

Entropy regularization enters the solution through the exponential kernel and its scaling. Unrolling a finite number of iterations allows gradients to propagate through these operations to the costs and upstream features. We use seven Sinkhorn iterations to obtain a finite-iteration approximation to $\Pi^*$, which denotes the optimum of the stated problem. Equivalent updates can also be performed in the log domain to improve numerical stability.

**Transport Blocks and Event Roles.** The augmented coupling has the same block structure as the augmented cost matrix:

$$\Pi^* = \begin{bmatrix} \Pi_p & \pi_d \\ \pi_a^\top & \pi_u \end{bmatrix}. \tag{16}$$

Here, $\Pi_p$ describes correspondences between real pre-change and post-change tokens, providing persistence evidence. The vector $\pi_d$ describes pre-change mass assigned to the post-change unmatched slot, providing disappearance evidence, while $\pi_a$ describes mass assigned from the pre-change unmatched slot to post-change tokens, providing appearance evidence. The scalar $\pi_u$ maintains mass balance between the auxiliary slots.

The marginal constraints imply $\sum_{j=1}^{N}(\Pi_p)_{ij} + (\pi_d)_i = a_i$, meaning that each pre-change token's mass comprises a matched component and a post-change unmatched component. Similarly, $\sum_{i=1}^{N}(\Pi_p)_{ij} + (\pi_a)_j = b_j$ decomposes each post-change token's mass into a component supported by pre-change correspondences and a component unexplained by the pre-change observation. TES extracts these components from the same coupling to construct persistent, disappearing, and appearing evidence $E_p, E_-, E_+$, respectively, and integrates them into the base visual representation to support change caption generation.

## B Hyperparameter Sensitivity Analysis

We examine the effects of the sliding-window size $w$ and transport temperature $\varepsilon$ on model performance. Specifically, $w$ determines the local search range for cross-temporal semantic references, while $\varepsilon$ controls the strength of entropy regularization in optimal transport.

**Sliding-window Size.** The window size affects the spatial coverage and matching selectivity of local semantic references. As shown in Figure 6, $w = 5$ achieves the best results across all four metrics among the tested settings, reaching 65.79, 40.36, 74.89, and 137.89 on BLEU-4, METEOR, ROUGE-L, and CIDEr, respectively. Both smaller windows ($w = 1, 3$) and larger windows ($w = 7, 9$) yield lower performance. This may be because smaller windows restrict access to neighboring semantic information, whereas larger windows introduce more potentially irrelevant matching candidates. These results suggest that a $5 \times 5$ neighborhood provides a favorable balance between local context coverage and matching selectivity, supporting our choice of $w = 5$.

**Transport Temperature.** The transport temperature $\varepsilon$ controls the trade-off between favoring low-cost correspondences and smoothing transport assignments. As shown in Figure 7, $\varepsilon = 0.03$ achieves the best overall performance, yielding the highest BLEU-4, ROUGE-L, and CIDEr scores.

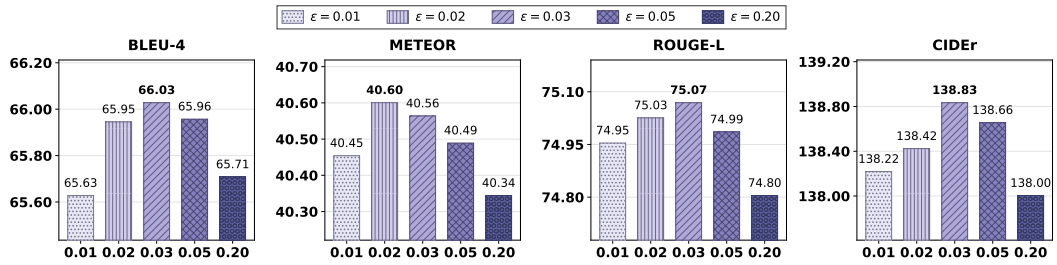


Figure 7: Sensitivity analysis of the transport temperature $\varepsilon$, which controls the strength of entropy regularization in optimal transport. The four panels report BLEU-4, METEOR, ROUGE-L, and CIDEr, respectively.

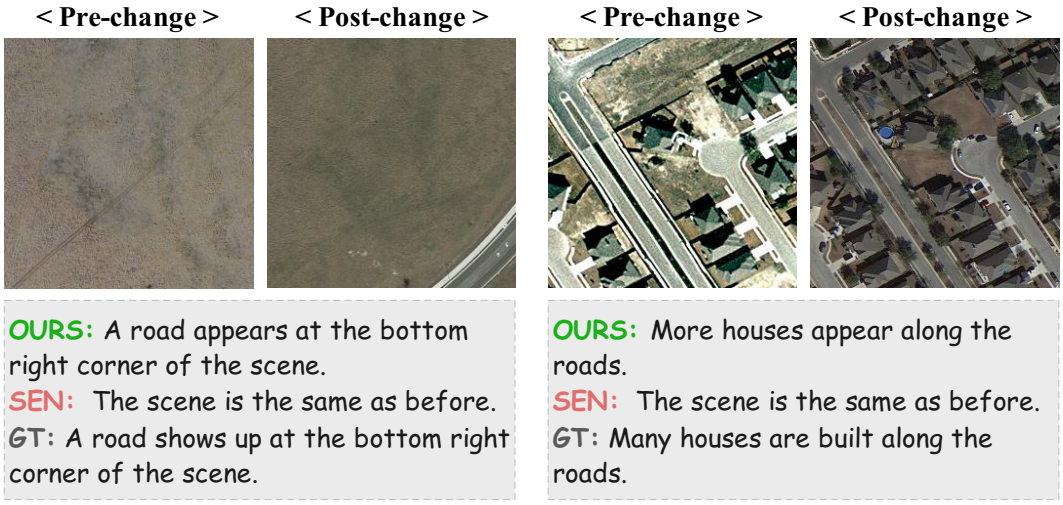


Figure 8: Additional qualitative comparisons of change captions. Each example presents the bi-temporal images, captions generated by our method and the baseline, and the ground-truth description (GT).

Results are similar across the tested temperatures between 0.02 and 0.05, whereas decreasing $\varepsilon$ to 0.01 or increasing it to 0.20 reduces all four metrics relative to $\varepsilon = 0.03$. This trend suggests that moderate entropy regularization benefits the task. A possible explanation is that an excessively low temperature produces overly concentrated assignments, while an excessively high temperature introduces too much smoothing and weakens the distinction between candidate correspondences. Considering the overall performance, we adopt $\varepsilon = 0.03$ as the default transport temperature. Results for the default configuration vary slightly across the two studies due to independent training runs.

## C Additional Case Studies

The additional qualitative results in Figures 8 and 9 illustrate the performance of DTE-Cap across different scenes. Our method captures local object changes and describes co-occurring appearance and disappearance events, generating captions that are semantically closer to the reference descriptions. It also reduces false change predictions caused by variations in illumination, viewpoint, and other factors, better distinguishing real semantic changes from irrelevant appearance differences.




Figure 9: Additional qualitative comparisons of change captions (continued). Each example presents the bi-temporal images, captions generated by our method and the baseline, and the ground-truth description (GT).