

STAND: Semantic Anchoring Constraint with Dual-Granularity Disambiguation for Remote Sensing Image Change Captioning

Yanpei Gong¹, Beichen Zhang^{1*}, Hao Wang¹, Xuhang Fu¹,
Zhaobo Qi¹, Xinyan Liu¹, Yuanrong Xu¹, Weigang Zhang¹

¹Harbin Institute of Technology, Weihai, China
2023211640@stu.hit.edu.cn, beiczhang@hit.edu.cn

Abstract

Remote sensing image change captioning (RSICC) aims to describe the difference between two remote sensing images. While recent methods have explored video modeling, they largely overlook the inherent ambiguities in viewpoint, scale, and prior knowledge, lacking effective constraints on the encoder. In this paper, we present STAND, a Semantic Anchoring Constraint with Dual-Granularity Disambiguation for RSICC, to progressively resolve these ambiguities. Specifically, to establish a reliable feature foundation, we first introduce an interpretable constraint to regularize temporal representations. Operating on these purified features, a dual-granularity disambiguation module resolves spatial uncertainties by coupling macro-level global context aggregation for viewpoint confusion with micro-level frequency-refocused attention for small-object scale enhancement. Ultimately, to translate these visually disambiguated features into precise text, a semantic concept anchoring module leverages language categorical priors to tackle knowledge ambiguity during decoding. Extensive experiments verify the superiority of STAND and its effectiveness in addressing viewpoint, scale, and knowledge ambiguities. The code will be publicly released on GitHub.

1 Introduction

Remote sensing image change captioning (RSICC) aims to describe the difference between two bi-temporal images. With increasing demand in urban monitoring and environmental assessment, this task has gained growing attention in both the remote sensing (RS) and vision-language communities (Shi et al. 2024; Liu et al. 2022a).

Unlike natural image change captioning, RSICC is more challenging due to the inherent ambiguities within RS images. We attribute these ambiguities to three factors: viewpoint, scale, and knowledge. (1) Viewpoint ambiguity (fig. 1 (a) (b)): since distinct objects may have highly similar appearances under a top-down and remote viewpoint, RS images are often confusing. (2) Scale ambiguity (fig. 1 (c)): a 256*256 image often corresponds to a region with hundreds of square meters (Liu et al. 2022a; Shi et al. 2024), while changed objects occupy only a small portion of pixels making the changes difficult to distinguish and recognize from the background. (3) Knowledge ambiguity (fig. 1 (d)): In contrast to natural image settings (Park, Darrell, and Rohrbach 2019), RSICC requires precise entity-level identification, making

*Corresponding author.

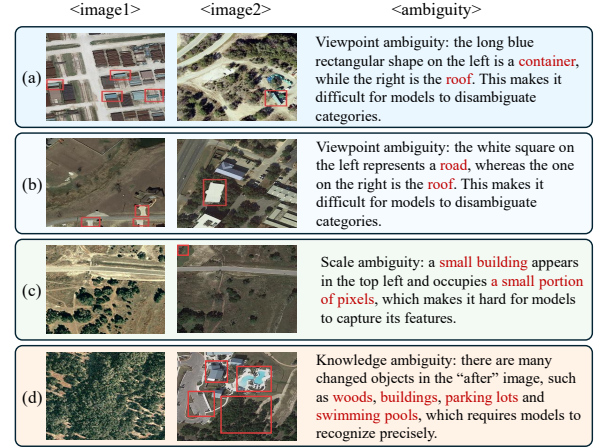


Figure 1: Typical examples of ambiguities in the remote sensing images. We attribute these ambiguities to three main factors: viewpoint, scale, and knowledge.

it heavily dependent on domain-specific prior knowledge, especially when multiple entities appear.

To generate accurate change descriptions, most previous studies (Chang and Ghamisi 2023; Zhou et al. 2024; Liu et al. 2022a) adopt a static image encoder for bi-temporal modeling. Nevertheless, image encoders mainly focus on local feature extraction while neglecting long-range context (Yang et al. 2024; Zhu et al. 2025). Building upon such encoders, difference-centric methods (Chen, Wang, and Yap 2024; Chang and Ghamisi 2023; Liu et al. 2022a, 2023a) assume that visual discrepancies directly correspond to semantic changes. However, this assumption often fails in RS due to the aforementioned ambiguity, where distinct objects may have a similar appearance from top-down viewpoint. Therefore, we argue that resolving such an issue requires understanding the overall scene rather than difference features alone. Recently, mask-based methods (Li et al. 2025b; Liu et al. 2024) attempted to highlight changed regions by hard filtering. Yet, this operation ignores contextual information that is crucial for disambiguating look-alike objects, often amplifying the ambiguity. More recently, video-based models utilizing learnable perception frames were proposed to extract temporal features (Zhu et al. 2025). While they

capture spatio-temporal dynamics, they naively overlook the inherent ambiguities common in remote sensing and lack effective constraints on the encoder, failing to systematically decouple these domain-specific challenges.

Based on these observations, we argue that high-quality change captioning requires progressively resolving these ambiguities. Accordingly, we propose STAND, a **S**emantic **A**nchoring **C**oNstraint with **D**ual-Granularity Disambiguation for RSICC. First, to capture temporal change dynamics, STAND models the before-mask-after triplet as a short video clip and encodes it using a video encoder. To further regularize the temporal transition, an Interpretable Transition Constraint (ITC) module encourages consistent relations among before, change, and after representations. Building upon these representations, we design a Dual-Granularity Target Disambiguation (DGTd) module to address scale and viewpoint ambiguities. Specifically, a macro-level Context-Aware Viewpoint Disambiguation (CAVD) module incorporates global semantics to distinguish visually similar entities, while a micro-level Frequency-Refocused Complementary Attention (FRCA) enhances small-object responses to mitigate scale ambiguity. Finally, to resolve knowledge ambiguity, a Semantic Concept Anchoring (SCA) module leverages language categorical priors to guide the decoding process.

Overall, STAND addresses the three ambiguities through a progressive pipeline and outperforms state-of-the-art methods on extensive RS datasets.

The main contributions are summarized as follows:

- We propose STAND, a spatio-temporal disambiguation framework for RSICC that jointly addresses viewpoint, scale, and knowledge ambiguities. Its transition constraint regularizes representations to capture change dynamics.
- We develop a progressive disambiguation strategy that couples dual-granularity spatial reasoning with semantic concept anchoring. DGTd combines global context with small-object refinement, while SCA grounds visual changes in structured language priors.
- We conduct extensive experiments together with ambiguity-specific evaluations, demonstrating consistent improvements under small-scale changes, viewpoint ambiguity, and knowledge ambiguity.

2 Related Work

2.1 Change Captioning

Change captioning is an essential task in the field of vision language processing (Wang et al. 2026; Tu et al. 2024; Zhu et al. 2025). Depending on the image domain, existing research in this area can be broadly divided into two main domains based on image type: natural images and remote sensing images (Chang and Ghamisi 2023).

In the domain of natural images, early works (Li et al. 2025a; Tu et al. 2021; Hosseinzadeh and Wang 2021; Lv, Wang, and Jing 2025; Tu et al. 2023a; Zhong et al. 2025) have explored various methods for localizing and describing changes. A typical approach (Kim et al. 2021; Yue et al. 2023; Shi et al. 2020) is to isolate the change by computing the image similarity and removing the shared components.

Another approach (Tu et al. 2023b) adopts contrastive learning for better robustness. These advances in natural images have offered valuable insights, yet their applicability to other domains remains limited due to domain-specific ambiguities.

In view of this, remote sensing image change captioning (RSICC) was proposed for satellite images. Unlike natural scenes, remote sensing (RS) images contain more complex, dynamic landscapes, substantial visual noise and unrelated changes like illumination and seasonal effects (Chang and Ghamisi 2023; Liu et al. 2025). Therefore, the goal is to describe meaningful changes while ignoring unimportant visual variations. Recent studies (Zhou et al. 2024; Li et al. 2024b; Liu et al. 2023a; Chen, Wang, and Yap 2024; Chang and Ghamisi 2023) employ various attention mechanisms or revisit the task from a video-modeling perspective (Zhu et al. 2025) to enhance difference perception. In parallel, multi-task frameworks (Wang et al. 2024; Shi et al. 2024; Karaca et al. 2025; Sun et al. 2025) leverage complementary supervision from related tasks to improve representation learning. More recently, works (Li et al. 2025b; Liu et al. 2024) have introduced change masks to highlight the changed areas and achieved promising results. Yet, their methods rely on explicit mask filtering, which risks disrupting contextual integrity and causes irreversible information loss. Furthermore, most RS methods (Shi et al. 2024; Liu et al. 2022a; Tu et al. 2024; Chang and Ghamisi 2023) adopt Siamese encoders that process the two images separately. However, this parallel processing architecture focuses on local feature extraction while often neglecting crucial scene-level long-range context (Zhu et al. 2025; Yang et al. 2024).

2.2 Video Understanding

Video understanding provides an effective paradigm for modeling spatio-temporal dynamics (Tang et al. 2025). By treating a sequence of frames as a unified input, video understanding models excel at capturing spatio-temporal relationships and dynamic events. Early works (Tran et al. 2015, 2018; Carreira and Zisserman 2017; Simonyan and Zisserman 2014) have extended 2D CNNs to 3D CNNs, achieving notable progress but suffering from high computational cost. To achieve a better trade-off between accuracy and efficiency, (Feichtenhofer 2020) proposed X3D to expand 2D image backbone along several network axes. This design enables powerful spatio-temporal feature learning while keeping computational complexity low, making it well suited for compact video-like modeling scenarios.

3 Methodology

As illustrated in fig. 2, STAND follows a two-stage pipeline. In the first stage, a pretrained video encoder encodes the short video clip, the ITC regularizes the temporal representations, and the DGTd module resolves scale and viewpoint ambiguities. In the second stage, the SCA module grounds the disambiguated visual representations in language priors and guides the decoder to generate change descriptions.

3.1 Video Clip Encoding

Formally, given a pair of bi-temporal images I_{bef} and I_{aft} , along with the “change mask” I_{mask} , STAND first concate-

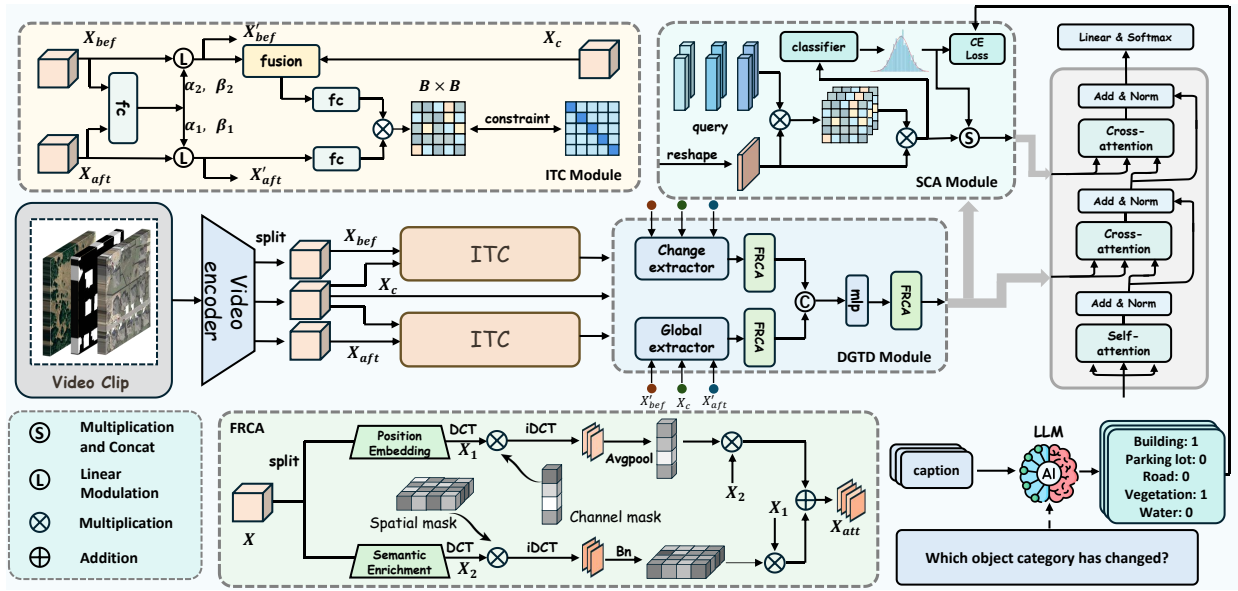


Figure 2: The architecture of STAND, including an interpretable transition constraint (ITC), dual-granularity target disambiguation (DGTD) module, and semantic concept anchoring (SCA) module. B and fc denote batch size and fully connected layer.

notes them to form a short video clip. The concatenated sequence is then fed into a video encoder and the encoder outputs are subsequently split into three feature tokens, denoted as X_{bef} , X_c , and X_{aft} . This process is defined by:

$$(X_{bef}, X_c, X_{aft}) = \text{Split}(\mathcal{F}_{\text{enc}}([I_{bef}; I_{\text{mask}}; I_{aft}))), \quad (1)$$

where $\mathcal{F}_{\text{enc}}(\cdot)$ is a video encoder, and $[\cdot; \cdot]$ denotes concatenation. $I \in \mathcal{R}^{3 \times H \times W}$, $X \in \{\mathcal{R}^{C_i \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}\}_{i=0}^3$. X_i^t indicates the feature of the t -th frame from the i -th encoder layer. I_{mask} is obtained from a pre-trained change detector (Zhu et al. 2025) for explicit change guidance.

3.2 Interpretable Transition Constraint

After video clip encoding, we apply FiLM-style linear modulation (Perez et al. 2018) between X_{bef} and X_{aft} , with feed-forward networks predicting element-wise scale and bias to suppress feature entanglement from 3D convolutions.

To supervise this modulation, we synthesize a pseudo-after feature from the corresponding before and change features:

$$X_{\text{p-aft}}^{(k)} = \mathcal{F}_{\text{fus}}(X_{\text{bef}}^{(k)}, X_c^{(k)}), \quad (2)$$

This explicitly models the transition: the composition should recover its paired after state in a sample-specific manner, rather than merely enforcing similarity between unspecified latent representations. We use AFF (Dai et al. 2021) for feature fusion. The pair $(X_{\text{aft}}^{(k)}, X_{\text{p-aft}}^{(k)})$ is treated as a positive sample, while all other features $X_{\text{p-aft}}^{(j)}$ ($j \neq k$) serve as negatives. The model projects $X_{\text{aft}}^{(k)}$ and $X_{\text{p-aft}}^{(k)}$ into a shared embedding space, denoted as $\tilde{X}_{\text{aft}}^{(k)}$ and $\tilde{X}_{\text{p-aft}}^{(k)}$.

We adopt a bidirectional InfoNCE loss (Oord, Li, and Vinyals 2018) to align with the true after state within the

batch, while pushing away all mismatched pairs:

$$\begin{aligned} \mathcal{L}_{p2a} &= -\frac{1}{B} \sum_{k=1}^B \log \frac{e^{\text{sim}(\tilde{X}_{\text{aft}}^{(k)}, \tilde{X}_{\text{p-aft}}^{(k)})/\tau}}{\sum_{j=1}^B e^{\text{sim}(\tilde{X}_{\text{aft}}^{(k)}, \tilde{X}_{\text{p-aft}}^{(j)})/\tau}}, \\ \mathcal{L}_{a2p} &= -\frac{1}{B} \sum_{k=1}^B \log \frac{e^{\text{sim}(\tilde{X}_{\text{p-aft}}^{(k)}, \tilde{X}_{\text{aft}}^{(k)})/\tau}}{\sum_{j=1}^B e^{\text{sim}(\tilde{X}_{\text{p-aft}}^{(k)}, \tilde{X}_{\text{aft}}^{(j)})/\tau}}, \\ \mathcal{L}_{\text{ca}} &= \mathcal{L}_{p2a} + \mathcal{L}_{a2p}, \end{aligned} \quad (3)$$

where “sim” is the dot-product to measure the similarity between two features. τ is a temperature hyper-parameter.

3.3 Dual-Granularity Target Disambiguation Module

Although ITC aligns temporal features, viewpoint and scale ambiguities remain in top-down RS images. DGTD addresses them through macro-level context aggregation and micro-level frequency attention, respectively.

Micro-Level: Frequency-Refocused Complementary Attention Continuous down-sampling weakens the high-frequency details that distinguish small changed objects, allowing low-frequency background responses to dominate. To mitigate this scale ambiguity, Frequency-Refocused Complementary Attention (FRCA) adaptively reweights spatial and channel frequency responses and combines them through complementary gating. Given $X \in \mathcal{R}^{C \times H \times W}$, we split its channels at a ratio of $\gamma : (1 - \gamma)$ and project the two parts back to C channels, obtaining X_s and X_c , respectively. Here, γ is a hyper-parameter governing channel allocation.

For frequency refinement, we construct a radial spatial mask $M_s \in \mathcal{R}^{1 \times H \times W}$ and a learnable channel-frequency mask $M_c \in \mathcal{R}^{C \times 1 \times 1}$. Let $d_{x,y}$ denote the normalized radial distance from frequency coordinate (x, y) to the DC

component. The masks are defined as:

$$\begin{aligned} d_{x,y} &= \frac{1}{\sqrt{2}} \sqrt{\left(\frac{x}{H-1}\right)^2 + \left(\frac{y}{W-1}\right)^2}, \\ \text{HighPass}_{x,y} &= \sigma(\kappa(d_{x,y} - t)), \\ \text{LowPass}_{x,y} &= 1 - \text{HighPass}_{x,y}, \\ M_s &= \text{HighPass} + (1 - \beta)\text{LowPass}, \\ M_c &= \sigma(\theta_c), \quad \theta_c \in \mathcal{R}^{C \times 1 \times 1}, \end{aligned} \quad (4)$$

where $\sigma(\cdot)$ is the sigmoid function, κ is a hyper-parameter controlling the filter sharpness, $t \in (0, 1)$ is the cutoff threshold, and β and θ_c are learnable. θ_c is initialized such that M_c is close to one. Within each FRCA instance, both masks are shared across samples. We then reweight frequency responses in the spatial and channel dimensions:

$$\begin{aligned} X'_s &= \text{iDCT}_{HW}(\text{DCT}_{HW}(X_s) \odot M_s), \\ X'_c &= \text{iDCT}_C(\text{DCT}_C(X_c) \odot M_c), \end{aligned} \quad (5)$$

where DCT_{HW} and DCT_C denote the 2D spatial DCT and 1D channel-wise DCT, respectively; iDCT_{HW} and iDCT_C are their inverses.

Finally, following complementary interaction (Xiao et al. 2025), the channel summary of X'_c modulates X_s , while the spatial response of X'_s modulates X_c :

$$X_{\text{att}} = X_s \odot \sigma(\text{AvgPool}(X'_c)) + X_c \odot \sigma(\text{BN}(\text{Conv}(X'_s))), \quad (6)$$

where $\text{BN}(\cdot)$ denotes BatchNorm and Conv reduces the channel dimension to 1.

Macro-Level: Context-Aware Viewpoint Disambiguation
Micro-level refinement strengthens small-object responses, but local evidence remains ambiguous for objects with similar top-down appearances. CAVD therefore uses the change feature to retrieve scene-wide evidence from both temporal features, flattened into HW tokens.

CAVD forms an ordered difference representation and a temporally balanced context representation:

$$\begin{aligned} D_{\text{bef}} &= \mathcal{A}_{d,1}(X_c, X_{\text{bef}}, X_{\text{bef}}), \\ C_{\text{diff}} &= \mathcal{A}_{d,2}(D_{\text{bef}}, X_{\text{aft}}, X_{\text{aft}}), \\ C_t &= \mathcal{A}_g(X_c, X_t, X_t), \quad t \in \{\text{bef}, \text{aft}\}, \end{aligned} \quad (7)$$

where the three arguments of each multi-head cross-attention operation denote query, key, and value, respectively. $\mathcal{A}_{d,1}$ and $\mathcal{A}_{d,2}$ sequentially preserve the before-to-after order, while $\mathcal{A}_g$ shares parameters across the two temporal inputs.

We combine C_{bef} and C_{aft} through a feature-wise gate:

$$\begin{aligned} G &= \sigma(\phi_g([C_{\text{bef}}; C_{\text{aft}}])), \\ C_{\text{global}} &= \phi_o([G \odot \phi_v(C_{\text{bef}}); (1 - G) \odot \phi_v(C_{\text{aft}})]), \end{aligned} \quad (8)$$

where ϕ_g , ϕ_v , and ϕ_o are linear projections, ϕ_v is shared by both inputs, and $[\cdot; \cdot]$ denotes channel concatenation.

The ordered branch retains directional difference cues, while the gated branch supplies scene-wide context. FRCA serves as the micro-level refinement operator when combining these two representations:

$$\begin{aligned} C'_{\text{diff}} &= \text{FRCA}_d(C_{\text{diff}}), \quad C'_{\text{global}} = \text{FRCA}_g(C_{\text{global}}), \\ C_{\text{truth}} &= \text{FRCA}_o(\phi_f([C'_{\text{diff}}; C'_{\text{global}}])), \end{aligned} \quad (9)$$

where ϕ_f is a linear projection, and each subscript denotes an independently parameterized FRCA instance.

3.4 Semantic Concept Anchoring Module

Although DGTD resolves spatial ambiguity, visual features alone may not reliably identify the changed entities. SCA therefore converts C_{truth} into category-indexed visual tokens using category-change labels obtained only from the training set; label construction is described in the experimental setup. Let $\tilde{C}_{\text{truth}} \in \mathcal{R}^{HW \times C}$ be the flattened visual feature. For category i , a learnable query set $Q_i \in \mathcal{R}^{n \times C}$ aggregates category-specific evidence:

$$\begin{aligned} A_i &= \text{Softmax}_{HW} \left(\frac{Q_i \tilde{C}_{\text{truth}}^\top}{\tau \sqrt{C}} \right), \\ z_i &= \text{LN} \left(\text{AvgPool}_n(A_i \tilde{C}_{\text{truth}}) \right), \end{aligned} \quad (10)$$

where $A_i \in \mathcal{R}^{n \times HW}$ is normalized over spatial locations, n is the number of queries per category, and τ is a learnable temperature. LN and AvgPool_n denote LayerNorm and averaging over the n queries, respectively.

To estimate whether category i has changed, z_i is concatenated with its learnable category embedding e_i and classified into change or no-change. The resulting probability $\mathbf{p}_i \in \mathcal{R}^2$ weights the category token:

$$\begin{aligned} \mathbf{p}_i &= \text{Softmax}(f_{\text{cls}}([z_i; e_i])), \\ \bar{z}_i &= \text{sg}(p_{i,1}) z_i, \quad C_{\text{object}} = \text{Stack}(\bar{z}_1, \dots, \bar{z}_N), \end{aligned} \quad (11)$$

where $p_{i,1}$ is the change probability, sg denotes stop-gradient, and $C_{\text{object}} \in \mathcal{R}^{N \times C}$ contains the N category tokens.

As the final step of concept-guided decoding, each decoder layer attends first to C_{truth} and then to C_{object} :

$$\begin{aligned} H'_l &= \text{MHCA}(\text{MHSA}(H_{l-1}), C_{\text{truth}}), \\ H_l &= \text{MHCA}(H'_l, C_{\text{object}}), \end{aligned} \quad (12)$$

Here, MHSA denotes multi-head self-attention, and the two-input MHCA uses its second input as both key and value. H_{l-1} and H_l denote the input and output of layer l , respectively. The top-layer representation is projected to vocabulary probabilities for caption generation.

3.5 Joint Training

Given the ground-truth caption $(w_1^*, \dots, w_m^*)$, the decoder minimizes the negative log-likelihood:

$$\mathcal{L}_{\text{cap}}(\theta) = - \sum_{t=1}^m \log p_\theta(w_t^* | w_{<t}^*), \quad (13)$$

where θ denotes learnable parameters. Category-change classification uses:

$$\mathcal{L}_{\text{cls}} = - \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^2 w_{i,j} y_{i,j} \log p_{i,j}, \quad (14)$$

where $y_{i,j}$ is the binary target for category i and $w_{i,j}$ balances the classes. The total objective is:

$$\mathcal{L} = \mathcal{L}_{\text{cap}} + \lambda_{\text{ca}} \mathcal{L}_{\text{ca}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}, \quad (15)$$

where λ_{ca} and λ_{cls} are hyper-parameters.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE _L	CIDEr
Chg2Cap (Chang and Ghamisi 2023)	86.14	78.08	70.66	64.39	40.03	75.12	136.61
MV-CC (Liu et al. 2024)	86.37	79.01	72.03	66.22	40.20	75.73	138.28
Change3D (Zhu et al. 2025)	85.81	77.81	70.57	64.38	40.38	75.12	138.29
KCFI (Yang et al. 2025)	86.34	77.31	70.89	65.30	39.42	75.47	138.25
CD4C (Li et al. 2025b)	86.74	78.71	71.48	65.41	40.56	75.72	138.00
SCNet (Sun et al. 2026)	86.14	78.19	71.44	65.82	40.51	75.37	140.23
RMNet (Cao et al. 2026)	86.09	78.43	71.68	66.11	40.94	75.57	139.43
D3-Net (Peng et al. 2026)	86.70	78.77	71.60	65.68	40.74	75.82	140.52
CTM (Bai and Wang 2025)	86.78	79.24	72.65	65.48	40.63	75.72	138.78
DGAT (Qin et al. 2025)	87.09	78.93	71.35	65.30	40.98	76.41	141.27
STAND (ours)	88.24	80.33	73.15	67.11	41.55	77.25	143.39

Table 1: Comparison with other methods on LEVIR-CC.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE _L	CIDEr
RSICCFFormer (Liu et al. 2022a)	80.05	74.24	69.61	66.54	42.65	73.91	133.44
CTMTNet (Shi et al. 2024)	83.56	77.66	72.76	69.00	45.39	79.23	149.40
CTM (Bai and Wang 2025)	85.36	79.49	75.36	72.36	46.98	80.97	153.29
SEN (Zhou et al. 2024)	85.12	80.28	76.84	74.71	46.81	79.45	148.70
CTSD-Net (Wu et al. 2025)	85.49	80.07	76.45	73.84	47.75	80.99	153.96
D3-Net (Peng et al. 2026)	85.42	80.02	76.23	73.15	47.13	81.47	154.33
DualFocus-CapNet (Meng et al. 2025)	85.85	80.67	76.65	73.59	48.03	81.28	157.71
FST-Net (Zou et al. 2025)	88.15	83.46	79.55	76.78	48.78	82.91	160.01
STAND (ours)	88.66	84.04	80.05	76.91	49.08	83.36	160.78

Table 2: Comparison with other methods on WHU-CDC.

4 Experiments

4.1 Experimental Setup

Datasets. LEVIR-CC (Liu et al. 2022a) consists of 10,077 image pairs with 5,038 change pairs. Each pair is annotated with five sentences. WHU-CDC (Shi et al. 2024) comprises 7,434 image pairs and 37,170 corresponding change descriptions. The sentences range from 3 to 24 words and are drawn from a vocabulary of 327 tokens. Both datasets adopt the same five annotated object categories, including *building*, *road*, *parking lot*, *vegetation*, and *water*.

Evaluation metrics. Following prior work (Liu et al. 2022a), we evaluate the captions with seven common metrics, *i.e.*, BLEU-N ($N = 1, 2, 3, 4$) (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE_L (Lin 2004), and CIDEr-D (Vedantam, Lawrence Zitnick, and Parikh 2015).

Implementation details. The proposed method adopts X3D-L (Feichtenhofer 2020) as the baseline. All experiments are conducted on an RTX 4090 GPU. To extract category change labels for the training set, we construct a dual-agent verification: Agent A (Qwen2.5-72B-Instruct) derives labels from only the training set, while Agent B (GPT-4o) performs an inspection. In addition, the change masks are obtained from a pre-trained change detector. **Importantly, label extraction and detector pretraining procedures are conducted strictly within the training data, thereby avoiding data leakage.**

4.2 Performance Comparison

Results on the LEVIR-CC Dataset. We compare STAND with SOTA methods, including Chg2Cap (Chang and Ghamisi 2023), MV-CC (Liu et al. 2024), CD4C (Li et al. 2025b), Change3D (Zhu et al. 2025), CTM (Bai and Wang 2025), KCFI (Yang et al. 2025), SCNet (Sun et al. 2026), RMNet (Cao et al. 2026), D3-Net (Peng et al. 2026), and DGAT (Qin et al. 2025).

As shown in table 1, STAND achieves the best performance, with notable improvements in METEOR (+0.57), ROUGE_L (+0.84), and CIDEr (+2.12) over the second-best method (DGAT). Meanwhile, MV-CC (Liu et al. 2024) and CD4C (Li et al. 2025b) are the methods that apply masks via simple filtering, whereas STAND leverages the mask as a guidance to focus on changes. STAND outperforms both approaches indicating the superiority of this guidance method.

Results on the WHU-CDC Dataset. We compare STAND with state-of-the-art methods, including RSICCFFormer (Liu et al. 2022a), CTMTNet (Shi et al. 2024), CTM (Bai and Wang 2025), SEN (Zhou et al. 2024), CTSD-Net (Wu et al. 2025), FST-Net (Zou et al. 2025), D3-Net (Peng et al. 2026), and DualFocus-CapNet (Meng et al. 2025).

Existing methods already achieve high scores on WHU-CDC, leaving limited room for further improvement. Nevertheless, STAND obtains the highest reported scores across all seven metrics in table 2, with modest but consistent gains over FST-Net and CIDEr improvements of 6.45 and 3.07 points over the recent D3-Net and DualFocus-CapNet, respectively.

(a) Progressive component ablation						
#	Model	BLEU-4	METEOR	ROUGE _L	CIDEr	Δ CIDEr
B0	Video Encoder (base)	64.38	40.38	75.12	138.29	–
B1	+DGTD (CAVD+FRCA)	65.24	40.31	75.81	139.29	+1.00
B2	+DGTD (CAVD+FRCA)+SCA	65.82	41.03	76.78	142.11	+3.82
B3	+ ITC + DGTD + SCA (Full)	67.11	41.55	77.25	143.39	+5.10

(b) DGTD component ablation						
Setting	CAVD	FRCA	BLEU-4	METEOR	ROUGE _L	CIDEr
w/o DGTD	×	×	64.38	40.25	75.61	139.66
Macro only (CAVD)	✓	×	65.52	40.41	75.83	140.21
Micro only (FRCA)	×	✓	64.79	40.55	75.58	140.17
Dual (CAVD+FRCA)	✓	✓	65.47	40.60	76.39	141.22

Table 3: Ablation studies on LEVIR-CC: (a) progressive component ablation; (b) DGTD ablation with ITC enabled and SCA excluded.

4.3 Ablation Study and Analysis

Progressive Ablation on LEVIR-CC. As shown in table 3(a), the model performance steadily improves as modules are progressively added. Among them, incorporating the SCA module brings the most significant performance gain, particularly in CIDEr. Adding ITC leads to improvements in absolute scores, and we conduct five other independent runs on B2 & B3 which demonstrate it stabilizes training, reducing the averaged metric variance from 0.26 to 0.18 across all evaluation metrics. These results verify the effectiveness and stability of the proposed progressive refinement pipeline.

Analysis of DGTD. Viewpoint and scale ambiguities are inherently spatial in nature, fundamentally different from knowledge ambiguity. To isolate their effects, we evaluate CAVD and FRCA separately in table 3(b), while keeping ITC enabled and excluding SCA for clarity. Using either CAVD or FRCA alone brings limited gains, indicating that resolving spatial ambiguity from a single perspective is insufficient. In contrast, combining both components yields the best performance on most metrics, demonstrating that macro-level contextual conditioning and micro-level feature refinement are complementary for tackling spatial ambiguity.

Effect of the Encoder Architecture. We further investigate image- and video-based backbones, including ResNet-101 (He et al. 2016), I3D (Carreira and Zisserman 2017), Swin-T (Liu et al. 2022b), and X3D (Feichtenhofer 2020). Among them, X3D-L achieves the best overall balance across BLEU, METEOR, and ROUGE. Considering its stable and competitive performance across metrics, we adopt X3D-L as the default backbone in subsequent experiments.

Effect of the Change Mask. To verify the necessity of high-quality change guidance, we replace the change mask with a naive pixel-wise difference mask. Specifically, this mask is generated by computing the absolute pixel difference between two images and applying a binary threshold. As shown in table 4, while this replacement leads to a metric drop, our method equipped with this naive mask *still* outperforms all *non-mask-based* models in table 1. This demonstrates that the superiority of STAND does not merely rely on the exter-

Setting	BLEU-4	METEOR	ROUGE _L	CIDEr
<i>Encoder Architecture</i>				
ResNet-101 (Image)	65.13	40.45	75.50	139.81
I3D	66.71	41.00	76.02	142.36
Swin-T	67.03	41.23	76.98	143.66
X3D-M	64.52	40.21	75.42	138.53
X3D-L	67.11	41.55	77.25	143.39
<i>Change guidance</i>				
Naive Diff. Mask	66.15	41.02	75.84	142.01
Pre-trained Mask	67.11	41.55	77.25	143.39

Table 4: Ablation studies on LEVIR-CC: encoder and masks.

nal mask, but fundamentally stems from the disambiguation capabilities of our core pipeline.

Ambiguity-Specific Evaluation. Existing RSICC works lack targeted evaluation for ambiguity. To validate whether STAND mitigates ambiguities, we design metrics for different ambiguity types, as shown in table 5.

For *scale ambiguity*, we evaluate results on a small-scale subset (selected by LLM), where BLEU-4 and METEOR reflect the model’s ability to describe subtle changes. For *viewpoint ambiguity*, we report the false positive rate (FPR) and precision, which measure whether the model avoids hallucinated changes when entities share similar appearance. For *knowledge ambiguity*, we further report entity-level recall and F1 score, which quantify the completeness and balance of detected changed categories. All entity-level metrics (FPR, Prec, Rec, F1) are computed by extracting change labels from the generated captions and comparing them with ground-truth labels. These are standard evaluation metrics in captioning and machine learning. Compared with Change3D and the targeted ablations, the full model performs best, confirming the roles of FRCA, CAVD, and SCA in scale, viewpoint, and knowledge disambiguation, respectively.

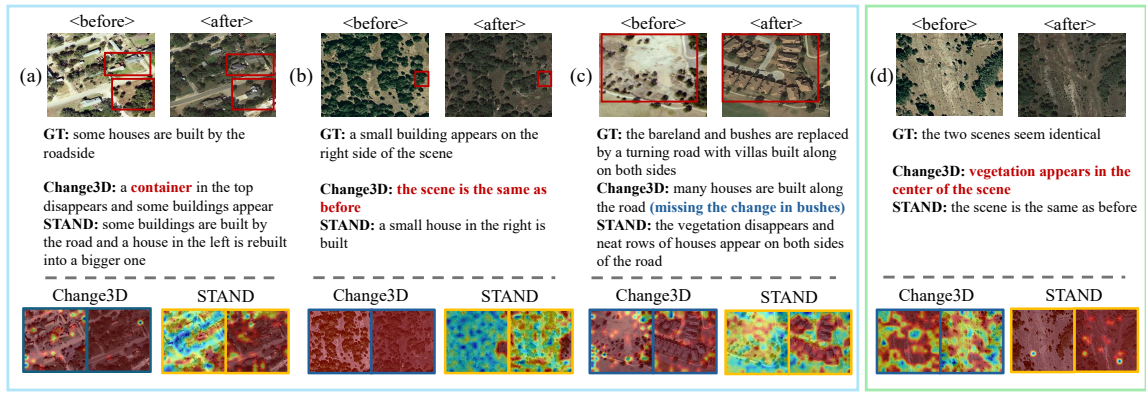


Figure 3: Visualization of baseline Change3D (Zhu et al. 2025) and STAND on LEVIR-CC. (a), (b), and (c) are change image pairs, while (d) is a no-change pair.

(a) Scale (Small-Scale Subset)			(b) Viewpoint Ambiguity			(c) Knowledge Ambiguity			
Setting	BLEU-4	METEOR	Setting	FPR↓	Prec↑	Setting	Prec↑	Rec↑	F1↑
Change3D (Baseline)	30.06	20.12	Change3D (Baseline)	0.06	0.70	Change3D (Baseline)	0.70	0.72	0.71
Ours w/o FRCA	30.72	20.86	Ours w/o CAVD	0.04	0.74	Ours w/o SCA	0.73	0.77	0.75
Ours	31.39	21.91	Ours	0.02	0.78	Ours	0.78	0.85	0.81

Table 5: Ambiguity-specific evaluation on LEVIR-CC.

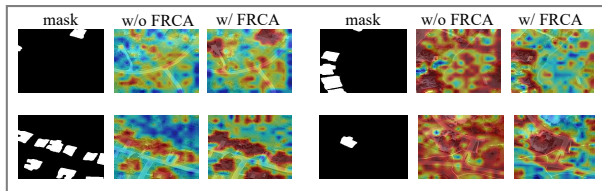


Figure 4: Visualization of the heatmap, without (left) and with (right) FRCA.

4.4 Visualization

We further compare STAND with Change3D (Zhu et al. 2025) and present four representative cases. As shown in fig. 3, we can observe: (a) Under the top-down remote-sensing viewpoint, the baseline misreads the building as a container, whereas our approach correctly recognizes the change. (b) When changes occur at a small scale, our model reliably captures differences, whereas Change3D predicts no change. (c) For captions involving multiple objects, STAND effectively exploits prior knowledge to improve semantic precision. In contrast, the caption generated by Change3D misses some of the changed objects. (d) For image pairs with no actual change, our attention remains uniformly distributed and is robust to seasonal effects. In addition, as shown in fig. 4, the proposed FRCA module refocuses attention from low-frequency regions to true change areas, improving the localization of real changes.

4.5 Comparison with MLLM Methods

To further examine whether general MLLMs can replace task-specific architectures and achieve competitive performance, we adapt off-the-shelf MLLMs (Zhu et al. 2023; Liu

Method	Model Size	BLEU-4	METEOR	ROUGE _L	CIDEr
MiniGPT4	1.1B + 7B	62.55	39.60	74.22	136.10
LLaVA-1.5	0.3B + 7B	62.75	39.85	75.10	136.56
LLaVA-ov	0.4B + 0.5B	64.55	40.77	75.32	138.24
LLaVA-ov	0.4B + 7B	63.66	41.23	75.78	139.61
STAND	9.2M + 6.5M	67.11	41.55	77.25	143.39

Table 6: Comparison with MLLM methods on LEVIR-CC. Model size represents encoder + decoder.

et al. 2023b; Li et al. 2024a) to RS change captioning and perform instruction tuning on the training split. As shown in table 6, our STAND clearly surpasses the tuned MLLMs. This suggests that MLLMs still lack sufficient domain awareness for this vertical remote sensing task, making their generic priors difficult to transfer to RSICC. Therefore, in the era of foundation MLLMs, carefully designed compact models still remain indispensable for RSICC, especially when domain-specific (e.g., remote sensing) data for pretraining or fine-tuning are limited in practice.

5 Conclusion

In this paper, we proposed STAND to model changes in a spatio-temporal manner, thereby resolving the ambiguities of viewpoint, scale, and knowledge in RS images. In the future, we plan to further explore the applications of RSICC to high-resolution images and more complex scenes (e.g., deserts, rainforests, coastal zones). At the same time, we hope to develop a closer integration with MLLMs and remote sensing priors to generate more high-level semantics. We believe that our work has the potential to inspire more studies on RSICC with spatio-temporal modeling.

References

- Bai, Q.; and Wang, X. 2025. Cross-Temporal Remote Sensing Image Change Captioning: A Manifold Mapping and Bayesian Diffusion Approach for Land Use Monitoring. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.
- Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.
- Cao, X.; Dong, W.; Qu, J.; and Li, Y. 2026. RMNet: Dual-Dimensional Difference Recalibration-Guided CNN-VMamba Synergistic Network for Remote Sensing Image Change Captioning. *IEEE Transactions on Geoscience and Remote Sensing*.
- Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.
- Chang, S.; and Ghamisi, P. 2023. Changes to captions: An attentive network for remote sensing change captioning. *IEEE Transactions on Image Processing*, 32: 6047–6060.
- Chen, C.; Wang, Y.; and Yap, K.-H. 2024. Multi-scale attentive fusion network for remote sensing image change captioning. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–5. IEEE.
- Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; and Barnard, K. 2021. Attentional feature fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3560–3569.
- Feichtenhofer, C. 2020. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 203–213.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Hosseinzadeh, M.; and Wang, Y. 2021. Image change captioning by learning from an auxiliary task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2725–2734.
- Karaca, A. C.; Ozelbas, E.; Berber, S.; Karimli, O.; Yildirim, T.; and Amasyali, M. F. 2025. Robust change captioning in remote sensing: Second-cc dataset and mmodalcc framework. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.
- Kim, H.; Kim, J.; Lee, H.; Park, H.; and Kim, G. 2021. Agnostic change captioning with cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2095–2104.
- Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. arXiv:2408.03326.
- Li, R.; Li, L.; Zhang, J.; Zhao, Q.; Wang, H.; and Yan, C. 2025a. Region-aware Difference Distilling with Attribute-guided Contrastive Regularization for Change Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4887–4895.
- Li, X.; Sun, B.; Wu, Z.; Li, S.; and Guo, H. 2025b. Cd4c: Change detection for remote sensing image change captioning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.
- Li, Y.; Zhang, X.; Cheng, X.; Chen, P.; and Jiao, L. 2024b. Inter-temporal interaction and symmetric difference learning for remote sensing image change captioning. *IEEE Transactions on Geoscience and Remote Sensing*.
- Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.
- Liu, C.; Yang, J.; Qi, Z.; Zou, Z.; and Shi, Z. 2023a. Progressive scale-aware network for remote sensing image change captioning. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, 6668–6671. IEEE.
- Liu, C.; Zhang, J.; Chen, K.; Wang, M.; Zou, Z.; and Shi, Z. 2025. Remote Sensing Spatiotemporal Vision–Language Models: A comprehensive survey. *IEEE Geoscience and Remote Sensing Magazine*.
- Liu, C.; Zhao, R.; Chen, H.; Zou, Z.; and Shi, Z. 2022a. 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.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.
- Liu, R.; Li, K.; Song, J.; Sun, D.; and Cao, X. 2024. Mvcc: Mask enhanced video model for remote sensing change caption. arXiv:2410.23946.
- Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022b. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3202–3211.
- Lv, F.; Wang, R.; and Jing, L. 2025. Revisiting Change Captioning from Self-supervised Global-Part Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5892–5900.
- Meng, X.; Zhao, Y.; Cao, K.; Wang, Q.; Wang, J.; and Hu, N. 2025. DualFocus-CapNet: A Dual-Stream Network for Real Change and Interscale Relationship-Aware Change Captioning in Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 19: 2045–2059.
- Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. arXiv:1807.03748.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.
- Park, D. H.; Darrell, T.; and Rohrbach, A. 2019. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4624–4633.

- Peng, C.; Wu, F.; Song, W.; and Wang, Z. 2026. Frequency-Spatial Semantic Decoupling for Remote Sensing Image Change Captioning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.
- Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Qin, P.; Liu, J.; Li, L.; Tian, C.; and Cao, X. 2025. DGAT: Dynamic Gaussian Attenuate Transformer for Remote Sensing Image Change Captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–16.
- Shi, J.; Zhang, M.; Hou, Y.; Zhi, R.; and Liu, J. 2024. A multi-task network and two large scale datasets for change detection and captioning in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*.
- Shi, X.; Yang, X.; Gu, J.; Joty, S.; and Cai, J. 2020. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *European conference on computer vision*, 574–590. Springer.
- Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27.
- Sun, D.; Wang, Y.; Yao, J.; Yu, W.; Cao, X.; and Ghamisi, P. 2026. SCNet: Lightweight Spatial-Channel Attention Network for Remote Sensing Change Captioning. *IEEE Transactions on Geoscience and Remote Sensing*.
- Sun, D.; Yao, J.; Xue, W.; Zhou, C.; Ghamisi, P.; and Cao, X. 2025. Mask approximation net: A novel diffusion model approach for remote sensing change captioning. *IEEE Transactions on Geoscience and Remote Sensing*.
- Tang, Y.; Bi, J.; Xu, S.; Song, L.; Liang, S.; Wang, T.; Zhang, D.; An, J.; Lin, J.; Zhu, R.; et al. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.
- Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 6450–6459.
- Tu, Y.; Li, L.; Su, L.; Du, J.; Lu, K.; and Huang, Q. 2023a. Adaptive representation disentanglement network for change captioning. *IEEE Transactions on Image Processing*, 32: 2620–2635.
- Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; Yan, C.; and Huang, Q. 2023b. Self-supervised cross-view representation reconstruction for change captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2805–2815.
- Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; Yan, C.; and Huang, Q. 2024. Context-aware Difference Distilling for Multi-change Captioning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7941–7956.
- Tu, Y.; Yao, T.; Li, L.; Lou, J.; Gao, S.; Yu, Z.; and Yan, C. 2021. Semantic relation-aware difference representation learning for change captioning. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 63–73.
- Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.
- Wang, H.; Zhang, B.; Gong, Y.; Fang, S.; Qi, Z.; Xu, Y.; Liu, X.; and Zhang, W. 2026. AIFIND: Artifact-Aware Interpreting Fine-Grained Alignment for Incremental Face Forgery Detection. arXiv:2604.16207.
- Wang, Y.; Yu, W.; Kopp, M.; and Ghamisi, P. 2024. Change-minds: Multi-task framework for detecting and describing changes in remote sensing. arXiv:2410.10047.
- Wu, R.; Ye, H.; Liu, X.; Li, Z.; Sun, C.; and Wu, J. 2025. A Cross-Spatial Differential Localization Network for Remote Sensing Change Captioning. *Remote Sensing*, 17(13): 2285.
- Xiao, Y.; Xu, T.; Xin, Y.; and Li, J. 2025. FBRT-YOLO: Faster and Better for Real-Time Aerial Image Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8673–8681.
- Yang, C.; Li, Z.; Jiao, H.; Gao, Z.; and Zhang, L. 2025. Enhancing perception of key changes in remote sensing image change captioning. *IEEE Transactions on Image Processing*.
- Yang, Y.; Liu, T.; Pu, Y.; Liu, L.; Zhao, Q.; and Wan, Q. 2024. Remote sensing image change captioning using multi-attentive network with diffusion model. *Remote Sensing*, 16(21): 4083.
- Yue, S.; Tu, Y.; Li, L.; Yang, Y.; Gao, S.; and Yu, Z. 2023. I3n: Intra-and inter-representation interaction network for change captioning. *IEEE Transactions on Multimedia*, 25: 8828–8841.
- Zhong, G.; Hu, J.; Chen, J.; Yuan, J.; and Pan, W. 2025. DECIDER: Difference-aware Contrastive Diffusion Model with Adversarial Perturbations for Image Change Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10662–10670.
- Zhou, Q.; Gao, J.; Yuan, Y.; and Wang, Q. 2024. Single-stream extractor network with contrastive pre-training for remote-sensing change captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–14.
- Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigtpt-4: Enhancing vision-language understanding with advanced large language models. arXiv:2304.10592.
- Zhu, D.; Huang, X.; Huang, H.; Zhou, H.; and Shao, Z. 2025. Change3D: Revisiting Change Detection and Captioning from A Video Modeling Perspective. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24011–24022.
- Zou, S.; Wei, Y.; Xie, Y.; and Luan, X. 2025. Frequency-Spatial-Temporal Domain Fusion Network for Remote Sensing Image Change Captioning. *Remote Sensing*, 17(8): 1463.