BIGENCAP: JOINT BITEMPORAL GENERATIVE MODELING VIA RECIPROCAL PREDICTION FOR REMOTE SENSING IMAGE CHANGE CAPTIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote sensing image change captioning (RSICC) requires interpreting semantic changes between temporally separated images while preserving scene context and cross-temporal relationships. Existing methods mainly construct change representations through direct comparison and interaction of bitemporal visual features. Beyond such pairwise comparison, large-scale collections of bitemporal images contain diverse visual semantics and cross-temporal dependencies. Modeling these distributions can provide a generative prior for interpreting specific image pairs. We propose BiGenCap, which learns a joint generative prior over bitemporal images at scale and instantiates it for each image pair through reciprocal generative prediction. Specifically, we pretrain a Diffusion Transformer (DiT) with flow matching on approximately 230K image pairs from 21 public data sources, assigning independent flow times to the two temporal streams to jointly model within-stream semantics and cross-temporal dependencies. During captioning, reciprocal generative prediction alternately perturbs one temporal feature stream while keeping its counterpart clean. We exploit the resulting predictions in two complementary ways: generative discrepancy fusion (GDF) derives prediction-based discrepancy cues from the final DiT responses, while reciprocal relation modeling (RRM) uses intermediate representations to model local cross-temporal relations and aggregate visual content from the temporal counterpart. BiGenCap achieves aggregate captioning scores S_{avg} of 82.55 and 94.17 on LEVIR-CC and WHU-CDC, respectively. Further analyses highlight the value of genuine bitemporal pairing and the complementarity of discrepancy and relational cues.

1 INTRODUCTION

Remote sensing image change captioning (RSICC) aims to describe semantic changes between two images of the same geographical area acquired at different times in natural language (Liu et al., 2022a; Chang & Ghamisi, 2023). This requires preserving the semantics of both images while identifying what changes, where it changes, and in which temporal direction. For example, describing a newly constructed building requires recognizing the building in the later image and identifying the earlier content of the same region. Effective change representations should therefore capture both scene semantics and cross-temporal change information.

Existing RSICC methods predominantly construct change representations through direct comparison and interaction of bitemporal visual features, using feature differencing, cross-attention, and spatial-temporal interaction (Zou et al., 2025; Wang et al., 2026a). These methods learn effective discriminative cues from the observed image pair. Beyond such pairwise comparison, large collections of bitemporal images contain diverse visual semantics within individual observations and cross-temporal dependencies between paired observations. Modeling their joint distribution can provide a generative prior over paired visual features, allowing changes in a specific image pair to be interpreted with reference to distribution-level information learned from diverse paired scenes.

To this end, we propose BiGenCap, which performs large-scale joint generative modeling of bitemporal visual features. Specifically, we pretrain a Diffusion Transformer (DiT) (Peebles & Xie, 2023) with flow matching (Lipman et al., 2022) on about 230K remote sensing image pairs collected from

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.

21 public data sources. Each temporal feature stream is assigned an independent flow time, forming a two-dimensional corruption space spanning jointly perturbed states and asymmetric states in which one stream remains clean. Predicting features across these states allows the model to combine retained information from the perturbed stream, its semantic context, and information from the temporal counterpart, thereby learning a generative prior over bitemporal visual features.

The learned prior provides distribution-level information, whereas change captioning ultimately requires instance-specific cues for a given image pair. We bridge this gap through *reciprocal generative prediction*. Given a bitemporal pair, we instantiate the learned prior under two role-reversed asymmetric states, alternately perturbing one temporal feature stream while keeping the other clean. At the output level, the final velocity responses yield clean-feature estimates and context-conditioned prediction residuals relative to the observed features. At the representation level, intermediate DiT features encode cross-temporal interactions formed during prediction and provide relational cues between the two temporal states. We exploit these complementary cues through *generative discrepancy fusion* (GDF) and *reciprocal relation modeling* (RRM). GDF fuses prediction-derived discrepancies with the original bitemporal features for visual comparison, while RRM uses intermediate DiT representations to model local cross-temporal relations and aggregate visual content from the temporal counterpart. Together, GDF and RRM incorporate the generative prior into change representation.

Experiments on two large-scale RSICC benchmarks, LEVIR-CC and WHU-CDC, demonstrate strong performance against previous RSICC methods. Further analyses highlight the importance of genuine bitemporal pairing and the complementary contributions of prediction-derived discrepancies and intermediate cross-temporal relations.

Our contributions are summarized as follows: (1) We introduce large-scale joint generative pretraining for RSICC. A flow-matching DiT pretrained on 230K image pairs from 21 public data sources learns a generative prior over bitemporal visual features. (2) We propose reciprocal generative prediction to instantiate the learned prior for a specific image pair through two role-reversed asymmetric states, yielding complementary final predictive responses and intermediate representations. (3) We propose GDF and RRM to incorporate prediction-derived discrepancies and intermediate cross-temporal relations into change representation learning.

2 Related Work

Most RSICC methods construct change-aware representations by explicitly comparing and integrating bitemporal visual features before language decoding. RSICCformer (Liu et al., 2022a) introduces difference-guided cross-encoding and multistage feature fusion, while Chg2Cap (Chang & Ghamisi, 2023) combines intra-image and inter-image attention. RSCaMa (Liu et al., 2024a) employs spatial difference-aware and temporal-traversing state-space modeling, and HiSem (Wang et al., 2026a) further separates change existence from fine-grained semantics through difference-guided interaction and hierarchical modeling. Despite different architectures, these methods mainly improve how semantic differences between two images are encoded and fused for captioning.

Beyond architectural designs for bitemporal interaction, recent studies introduce additional supervision or pretrained knowledge to strengthen change representations. SEN (Zhou et al., 2024) performs contrastive learning on bitemporal images, while Change-Agent (Liu et al., 2024b), CTMT-Net (Shi et al., 2024), ChangeMinds (Wang et al., 2024b), and FST-Net (Zou et al., 2025) incorporate change detection into joint or multi-task learning. CD4C (Li et al., 2025) uses predicted binary masks to guide foreground feature extraction. Pretrained foundation and language models provide another source of knowledge: PromptCC (Liu et al., 2023) combines prompt learning with a pretrained language model, Semantic-CC (Zhu et al., 2024) introduces foundation-model features and semantic guidance, and KCFI (Yang et al., 2025) couples change-detection supervision with an instruction-tuned language decoder. More recently, CDChat (Noman et al., 2025) and RSICCLLM (Wang et al., 2026b) adapt multimodal large language models to RSICC through instruction tuning and task-specific post-training. These approaches enrich change captioning with auxiliary supervision and transferable visual-language knowledge.

Generative modeling has also been explored in RSICC, but existing methods mainly introduce generation in the language or auxiliary prediction space. Diffusion-RSCC (Yu et al., 2025) generates captions by denoising continuous language embeddings conditioned on bitemporal features, whereas

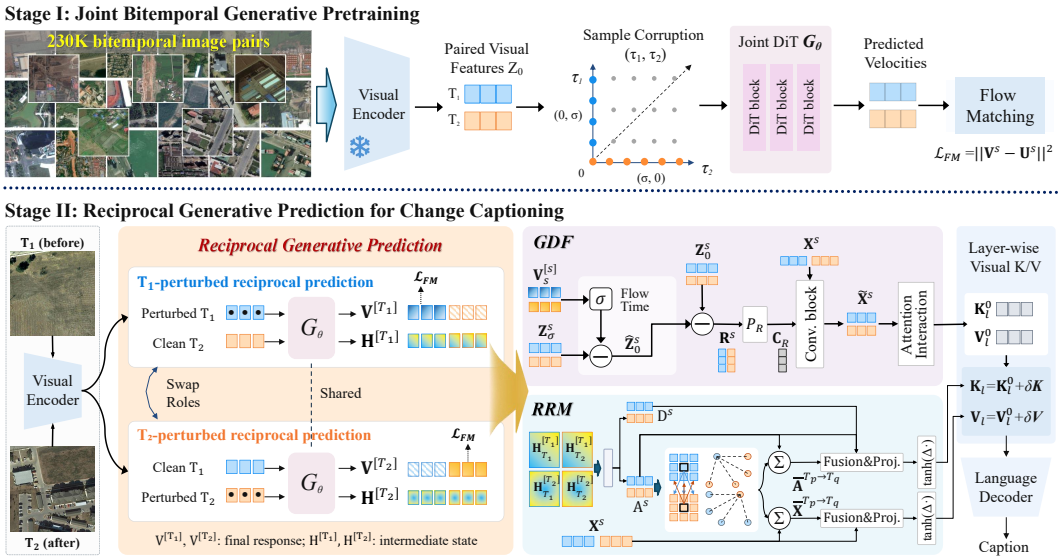


Figure 1: Overview of BiGenCap. Stage I learns a generative prior over paired bitemporal features through joint flow-matching pretraining. Stage II instantiates this prior via reciprocal generative prediction, with GDF using final responses and RRM using intermediate representations for captioning.

Mask Approximation Net (Sun et al., 2025) employs conditional diffusion to generate change masks for subsequent caption decoding. In contrast, our BiGenCap performs joint generative modeling directly for bitemporal visual features, learning a generative prior from large-scale paired imagery.

3 Method

3.1 Overview

Given a bitemporal image pair $\mathbf{I}^{T_1}$ and $\mathbf{I}^{T_2}$ of the same geographical area acquired at times $T_1 < T_2$, our goal is to generate a natural-language description of their semantic changes. As illustrated in Figure 1, BiGenCap consists of two stages: (i) joint bitemporal generative pretraining to learn a generative prior over paired visual features, and (ii) change captioning that instantiates this prior through reciprocal generative prediction. In the first stage, a shared visual encoder extracts two temporal feature streams, and a joint Diffusion Transformer (DiT) is pretrained with flow matching using independently sampled flow times for the two streams. In the second stage, reciprocal prediction is performed under two role-reversed asymmetric states, alternately perturbing one temporal stream while keeping its counterpart clean. Generative discrepancy fusion (GDF) derives prediction-based discrepancy cues from the final DiT responses and fuses them with the original bitemporal features, whereas reciprocal relation modeling (RRM) uses intermediate DiT representations to model local cross-temporal relations and aggregate visual content from the temporal counterpart. Together, they construct layer-wise visual keys and values for autoregressive caption generation.

3.2 Joint Bitemporal Generative Pretraining

Before change-caption training, we pretrain a joint generative model on paired remote sensing imagery to model the joint distribution of bitemporal visual features. We construct a corpus of 229,642 image pairs from 21 public remote sensing data sources, excluding the LEVIR-CC and WHU-CDC test images (Appendix B). A shared visual encoder, fixed during this stage, maps each pair to third-stage visual features $\mathbf{Z}_0^{T_1}$ and $\mathbf{Z}_0^{T_2}$, which are treated as samples from the joint feature distribution $p_{\text{pair}}(\mathbf{Z}_0^{T_1}, \mathbf{Z}_0^{T_2})$. We use the third-stage features to preserve spatial granularity for subsequent reciprocal prediction and local relation modeling. We then learn a joint flow field over the two temporal feature streams, allowing feature prediction to exploit both within-stream semantics and cross-temporal information from the temporal counterpart.

3.2.1 JOINT BITEMPORAL FLOW-MATCHING FORMULATION

For $q \in \{1, 2\}$, let $\mathbf{Z}_0^{T_q} \in \mathbb{R}^{N \times C_z}$ denote the clean visual features of image T_q. We independently sample a flow time $\tau_q \in [0, 1]$ for each temporal stream and define the linear probability path

$$\mathbf{Z}_{\tau_q}^{T_q} = (1 - \tau_q)\mathbf{Z}_0^{T_q} + \tau_q \boldsymbol{\epsilon}^{T_q}, \qquad \mathbf{U}^{T_q} = \boldsymbol{\epsilon}^{T_q} - \mathbf{Z}_0^{T_q}, \qquad \boldsymbol{\epsilon}^{T_q} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $\mathbf{U}^{T_q}$ denotes the target velocity. Independently controlling the two flow times yields a two-dimensional bitemporal corruption space: $(0, 0)$ corresponds to the clean paired features, $(1, 1)$ to Gaussian endpoints, and $(\tau, 0)$ or $(0, \tau)$ to an asymmetric state in which one stream is perturbed while its temporal counterpart remains clean. A shared DiT (Peebles & Xie, 2023) jointly processes the two temporal token groups and parameterizes the joint velocity field:

$$\left(\mathbf{V}_\theta^{T_1}, \mathbf{V}_\theta^{T_2}\right) = G_\theta\left(\mathbf{Z}_{\tau_1}^{T_1}, \mathbf{Z}_{\tau_2}^{T_2}, \tau_1, \tau_2\right). \tag{2}$$

The two streams interact through joint self-attention, with spatial position, temporal role, and flow time explicitly encoded. Each velocity prediction can therefore exploit information retained in its own corrupted stream together with context from the temporal counterpart. Architectural and conditioning details are provided in Appendix A.1.

3.2.2 CORRUPTION-SPACE SAMPLING AND PRETRAINING OBJECTIVE

To cover complementary regions of the bitemporal corruption space, we sample (τ_1, τ_2) from the mixture $q_{\text{plane}} = \frac{1}{4}q_{\text{diag}} + \frac{1}{4}q_{\text{int}} + \frac{1}{2}q_{\text{bnd}}$. Diagonal samples use (τ, τ), interior samples draw τ_1 and τ_2 independently, and boundary samples use $(\tau, 0)$ or $(0, \tau)$ with equal probability. Diagonal and interior samples expose the model to jointly and asymmetrically perturbed feature pairs, while boundary samples emphasize the one-perturbed/one-clean states used in downstream reciprocal generative prediction. Each nonzero flow time is sampled as $\tau = \text{sigmoid}(u)$ with $u \sim \mathcal{N}(0, 1)$. The joint DiT is pretrained with the flow-matching objective (Lipman et al., 2022)

$$\mathcal{L}_{\text{pre}} = \frac{\sum_{n=1}^{M} \sum_{s \in \{T_1, T_2\}} m_n^s \left\| \mathbf{V}_{\theta,n}^s - \mathbf{U}_n^s \right\|_F^2}{NC_z \sum_{n=1}^{M} \sum_{s \in \{T_1, T_2\}} m_n^s}, \tag{3}$$

where M is the minibatch size and $m_n^s \in \{0, 1\}$ indicates whether stream s contributes to the regression loss. Both streams are supervised for diagonal and interior samples. For boundary samples, only the perturbed stream is supervised, while the clean stream serves as temporal context through joint self-attention. This stream-wise supervision directly matches the asymmetric prediction setting used during change captioning.

3.3 RECIPROCAL GENERATIVE PREDICTION

Joint generative pretraining learns a distribution-level prior over paired visual features, whereas change captioning requires instance-specific cues for a given image pair. We instantiate the learned prior through two role-reversed asymmetric predictions. Given clean features $\mathbf{Z}_0^{T_1}$ and $\mathbf{Z}_0^{T_2}$, we alternately perturb one temporal stream while keeping its counterpart clean. At a fixed corruption scale $\sigma \in (0, 1)$, we compute

$$\mathbf{V}^{[T_1]} = G_\theta\left(\mathbf{Z}_\sigma^{T_1}, \mathbf{Z}_0^{T_2}, \sigma, 0\right), \qquad \mathbf{V}^{[T_2]} = G_\theta\left(\mathbf{Z}_0^{T_1}, \mathbf{Z}_\sigma^{T_2}, 0, \sigma\right). \tag{4}$$

The bracketed superscript denotes the perturbed temporal stream. Each reciprocal prediction yields final velocity responses $\mathbf{V}^{[T_q]} = (\mathbf{V}_{T_1}^{[T_q]}, \mathbf{V}_{T_2}^{[T_q]})$ and intermediate DiT representations $\mathbf{H}^{[T_q]} = (\mathbf{H}_{T_1}^{[T_q]}, \mathbf{H}_{T_2}^{[T_q]})$. For example, $\mathbf{H}_{T_2}^{[T_1]}$ denotes the representation of the clean T_2 stream when T_1 is perturbed. Because $\sigma < 1$, the perturbed stream retains instance-specific visual information while the clean counterpart provides cross-temporal context. Each prediction therefore exploits information from both temporal streams rather than inferring one temporal image solely from the other. Reversing the roles provides complementary information for the same image pair. The perturbed-stream final responses provide prediction-derived discrepancy cues, while the intermediate representations encode cross-temporal relations formed during prediction.

3.4 CHANGE REPRESENTATION FROM RECIPROCAL PREDICTION

Reciprocal generative prediction provides complementary cues at the output and representation levels. Generative discrepancy fusion (GDF) converts the perturbed-stream final responses into prediction-derived discrepancy cues and integrates them with the original bitemporal features for visual comparison. Reciprocal relation modeling (RRM) instead uses intermediate DiT representations to characterize local cross-temporal relations and aggregate visual content from the temporal counterpart. Both modules reuse the same two reciprocal DiT predictions.

3.4.1 GENERATIVE DISCREPANCY FUSION

For each perturbed temporal stream, the predicted velocity provides a one-step estimate of its clean feature:

$$\widehat{\mathbf{Z}}_0^s = \mathbf{Z}_\sigma^s - \sigma \mathbf{V}_s^{[s]}, \qquad \mathbf{R}^s = \mathbf{Z}_0^s - \widehat{\mathbf{Z}}_0^s, \quad s \in \{T_1, T_2\}. \tag{5}$$

The prediction residual $\mathbf{R}^s$ measures the deviation between the clean feature and an estimate produced from the retained information in the perturbed stream together with its clean temporal counterpart. It therefore serves as an output-level discrepancy cue complementary to the original visual content. We construct multiscale features $\mathbf{X}^s \in \mathbb{R}^{N\times C}$ by top-down fusion of the third and fourth backbone stages and inject the two prediction residuals through a content-dependent gate:

$$\mathbf{C}_R = P_R\left([\mathbf{R}^{T_1}, \mathbf{R}^{T_2}]\right), \qquad \widetilde{\mathbf{X}}^s = \mathbf{X}^s + \mathrm{sigmoid}\left(P_g([\mathbf{X}^s, \mathbf{C}_R])\right) \odot \mathbf{C}_R. \tag{6}$$

Here, $[\cdot,\cdot]$ denotes channel concatenation and $\odot$ denotes element-wise multiplication. The gate adaptively controls the contribution of prediction-derived discrepancy information to each temporal stream while retaining the original scene content. The resulting features are processed by L_{vis} bitemporal interaction blocks and fused into a base visual feature $\mathbf{M}_0$. For decoder layer l, the corresponding base visual keys and values are $\mathbf{K}_l^0 = \mathcal{K}_l(\mathbf{M}_0)$ and $\mathbf{V}_l^0 = \mathcal{V}_l(\mathbf{M}_0)$, which are further refined by RRM in the next subsection. Implementation details of GDF are provided in Appendix A.2.

3.4.2 RECIPROCAL RELATION MODELING

While GDF exploits the final responses of reciprocal prediction, RRM uses intermediate DiT representations from the same two predictions to capture cross-temporal relations. These intermediate states provide features for estimating how local content is related across the two temporal streams, while the original multiscale features $\mathbf{X}^{T_1}$ and $\mathbf{X}^{T_2}$ provide the visual content to be aggregated.

Reciprocal relation features. Each temporal stream appears once in the clean role and once in the perturbed role. For $q \in \{1,2\}$ and $p = 3-q$, we define

$$\mathbf{A}^{T_q} = \mathcal{A}\left(\mathbf{H}_{T_q}^{[T_p]}\right), \qquad \mathbf{D}^{T_q} = \mathcal{B}\left(\mathbf{H}_{T_q}^{[T_q]} - \mathbf{H}_{T_q}^{[T_p]}\right), \tag{7}$$

where $\mathcal{A}$ and $\mathcal{B}$ are shared adapters that map the intermediate representations to $\mathbb{R}^{N\times C_r}$. $\mathbf{A}^{T_q}$ represents stream T_q in its clean role and is used to estimate cross-temporal relations, whereas $\mathbf{D}^{T_q}$ characterizes the change in its intermediate representation when the stream switches from the clean role to the perturbed role.

Local cross-temporal relations. For each location i in T_q, we estimate local relation weights over a neighborhood $\mathcal{N}(i)$ in the temporal counterpart:

$$\begin{aligned} w_{ij}^{T_p\to T_q} &= \frac{\exp\left(\cos(\mathbf{A}_i^{T_q}, \mathbf{A}_j^{T_p})/\kappa\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\cos(\mathbf{A}_i^{T_q}, \mathbf{A}_k^{T_p})/\kappa\right)}, \\ \overline{\mathbf{A}}_i^{T_p\to T_q} &= \sum_{j\in\mathcal{N}(i)} w_{ij}^{T_p\to T_q} \mathbf{A}_j^{T_p}, \qquad \overline{\mathbf{X}}_i^{T_p\to T_q} = \sum_{j\in\mathcal{N}(i)} w_{ij}^{T_p\to T_q} \mathbf{X}_j^{T_p}. \end{aligned} \tag{8}$$

Here, κ is a learnable temperature, and $\mathcal{N}(i)$ is centered at the same grid location in the temporal counterpart. Since each feature token already represents contextualized local content, the neighborhood allows nearby semantic information to participate in relation estimation. Importantly, the relation weights are computed solely from the clean-role features $\mathbf{A}$ and are then reused to aggregate

both relation features and original visual content. This separates relation estimation from content aggregation.

Relation and content features. We first combine the clean-role features with their locally aggregated counterpart context:

$$\begin{aligned}\mathbf{u}_i &= [\mathbf{A}_i^{T_1} + \mathbf{A}_i^{T_2},\, \mathbf{A}_i^{T_1} - \overline{\mathbf{A}}_i^{T_2\to T_1},\, \mathbf{A}_i^{T_2} - \overline{\mathbf{A}}_i^{T_1\to T_2}],\\ \mathbf{k}_i &= \mathrm{RMSNorm}\left(f_a(\mathbf{u}_i) + f_d\left([\mathbf{D}_i^{T_1} + \mathbf{D}_i^{T_2},\, \mathbf{D}_i^{T_2} - \mathbf{D}_i^{T_1}]\right)\right).\end{aligned} \tag{9a}$$

The content branch uses the same relation weights to aggregate the original counterpart features:

$$\begin{aligned}\mathbf{c}_i^{T_q} &= f_c\left([\mathbf{X}_i^{T_q},\, \mathbf{X}_i^{T_q} - \overline{\mathbf{X}}_i^{T_p\to T_q}]\right), \qquad q\in\{1,2\},\\ \mathbf{v}_i &= \mathrm{RMSNorm}\left(f_v\left([\mathbf{c}_i^{T_1} + \mathbf{c}_i^{T_2},\, \mathbf{c}_i^{T_2} - \mathbf{c}_i^{T_1}]\right)\right).\end{aligned} \tag{9b}$$

The symmetric terms combine information from both temporal streams, while the ordered differences preserve temporal direction. Stacking the resulting token features gives $\mathbf{k}, \mathbf{v} \in \mathbb{R}^{N\times C}$. We use $\mathbf{k}$ to refine the decoder keys with cross-temporal relation information and $\mathbf{v}$ to enrich the visual content carried by the values:

$$\mathbf{K}_l = \mathbf{K}_l^0 + \alpha_K \tanh\left(\Delta\mathcal{K}_l(\mathbf{k})\right), \qquad \mathbf{V}_l = \mathbf{V}_l^0 + \alpha_V \tanh\left(\Delta\mathcal{V}_l(\mathbf{v})\right). \tag{10}$$

Thus, the decoder keys incorporate cross-temporal relation cues for visual selection, while the values provide relation-enriched scene content. Adapter structures, neighborhood settings, and projection details are provided in Appendix A.3.

3.5 Caption Generation and Learning Objectives

The autoregressive Transformer decoder attends to the layer-wise visual keys and values $\{(\mathbf{K}_l, \mathbf{V}_l)\}_{l=1}^{L_{\mathrm{dec}}}$ through cross-attention to generate the change caption. During cross-entropy (CE) training, we retain the flow-matching objective on the perturbed stream of each reciprocal prediction as an auxiliary training signal for the generative predictor. The visual encoder, joint DiT, GDF, RRM, and caption decoder are jointly optimized with

$$\mathcal{L}_1 = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{FM}}\mathcal{L}_{\mathrm{FM}}^{\mathrm{cap}}. \tag{11}$$

where $\mathcal{L}_{\mathrm{CE}}$ denotes the caption cross-entropy loss, $\mathcal{L}_{\mathrm{FM}}^{\mathrm{cap}}$ is the velocity-regression loss over the perturbed streams of the two reciprocal predictions, and λ_{FM} controls its contribution. We then freeze the visual encoder, joint DiT, GDF, and RRM and refine only the caption decoder using self-critical sequence training (SCST) (Rennie et al., 2017): $\mathcal{L}_2 = 0.2\mathcal{L}_{\mathrm{CE}} + 0.8\mathcal{L}_{\mathrm{SCST}}$. The SCST reward is the mean of BLEU-4, METEOR, ROUGE-L, and CIDEr-D. Detailed formulations of the captioning and auxiliary objectives are provided in Appendix A.4.

4 Experiments

4.1 Experimental Setup

We evaluate BiGenCap on two widely used RSICC benchmarks, LEVIR-CC (Liu et al., 2022a) and WHU-CDC (Shi et al., 2024), containing 10,077 and 7,434 bitemporal image pairs, respectively, with five reference captions per pair. We report BLEU-1/2/3/4 (Papineni et al., 2002), METEOR (Lavie & Agarwal, 2007), ROUGE-L (Lin, 2004), CIDEr-D (Vedantam et al., 2015), and the aggregate score S_{avg}, defined as the mean of BLEU-4, METEOR, ROUGE-L, and CIDEr-D. We use SegFormer MiT-B1 (Xie et al., 2021) as the visual encoder and initialize the joint DiT from the generative pretraining. Reciprocal prediction uses $(\sigma_{\mathrm{tar}}, \sigma_{\mathrm{ctx}}) = (0.3, 0)$ by default, where σ_{tar} and σ_{ctx} denote the corruption levels of the perturbed stream and its temporal counterpart, respectively. Caption training consists of CE optimization followed by decoder-only SCST, with $\lambda_{\mathrm{FM}} = 0.01$ on LEVIR-CC and 0.001 on WHU-CDC. Final SCST results are used for comparison with previous methods. Since SCST freezes the visual model and updates only the caption decoder, visual-component ablations are evaluated after CE training under a common setting, with caption-stage flow regression disabled to isolate the effects of individual components. Additional architectural and implementation details are provided in Appendix B.

Table 1: Comparison with previous methods on the LEVIR-CC and WHU-CDC benchmarks. Higher values indicate better performance.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D	S_{avg}
LEVIR-CC								
RSICCformer (Liu et al., 2022a)	84.72	76.27	68.87	62.77	39.61	74.12	134.12	77.66
Chg2Cap (Chang & Ghamisi, 2023)	86.14	78.08	70.66	64.39	40.03	75.12	136.61	79.04
RSCaMa (Liu et al., 2024a)	85.79	77.99	71.04	65.24	39.91	75.24	136.56	79.24
SEN (Zhou et al., 2024)	85.10	77.05	70.01	64.09	39.59	74.57	136.02	78.57
Mask Approx Net (Sun et al., 2025)	85.90	77.12	70.72	64.32	39.91	75.67	137.71	79.40
KCFI (Yang et al., 2025)	86.34	77.31	70.89	65.30	39.42	75.47	138.25	79.61
HiSem (Wang et al., 2026a)	86.60	78.74	71.73	65.82	40.39	75.77	138.86	80.21
FST-Net (Zou et al., 2025)	86.76	78.82	71.71	65.67	40.51	76.15	140.04	80.59
SCNet (Sun et al., 2026)	86.14	78.19	71.44	65.82	40.51	75.37	140.23	80.48
BiGenCap (ours)	**87.42**	**79.83**	**73.11**	**67.47**	**41.72**	**77.03**	**143.97**	**82.55**
WHU-CDC								
RSICCformer (Liu et al., 2022a)	80.05	74.24	69.61	66.54	42.65	73.91	133.44	79.14
Chg2Cap (Chang & Ghamisi, 2023)	78.93	72.64	67.20	62.71	41.46	77.95	144.18	81.58
SEN (Zhou et al., 2024)	80.60	74.64	67.69	61.97	36.76	71.70	133.57	76.00
Mask Approx Net (Sun et al., 2025)	81.34	75.68	71.16	67.73	43.89	75.41	135.31	80.59
CTMTNet (Shi et al., 2024)	83.56	77.66	72.76	69.00	45.39	79.23	149.40	85.76
HiSem (Wang et al., 2026a)	87.49	82.89	79.17	76.52	48.77	82.00	158.35	91.41
FST-Net (Zou et al., 2025)	88.15	83.46	79.55	76.78	48.78	82.91	160.01	92.12
SCNet (Sun et al., 2026)	84.47	78.72	73.45	69.01	44.44	80.48	153.54	86.87
BiGenCap (ours)	**88.83**	**84.11**	**80.12**	**76.99**	**50.27**	**84.11**	**165.31**	**94.17**

4.2 Comparison with Previous Methods

Table 1 compares BiGenCap with previous RSICC methods on LEVIR-CC and WHU-CDC. The compared approaches span different strategies for change representation learning, including bitemporal feature interaction, auxiliary change-detection supervision, and generative modeling. For example, Mask Approximation Net (Sun et al., 2025) employs diffusion-based change-mask generation, while KCFI (Yang et al., 2025), CTMTNet (Shi et al., 2024), and FST-Net (Zou et al., 2025) incorporate auxiliary change-detection supervision. In contrast, BiGenCap learns a generative prior directly from large-scale paired imagery without semantic annotations during generative pretraining and instantiates it through reciprocal generative prediction.

BiGenCap achieves the highest scores across all seven captioning metrics on both benchmarks, reaching an S_{avg} of 82.55 on LEVIR-CC and 94.17 on WHU-CDC, improving the previous best results by 1.96 and 2.05 points, respectively. The consistent gains across all metrics indicate that the improvement is not confined to a particular evaluation criterion and support large-scale paired generative pretraining as a competitive source of information for change captioning beyond direct bitemporal interaction and auxiliary dense supervision.

Figure 2 compares representative predictions from Chg2Cap, HiSem, and BiGenCap. The examples highlight differences in changed-object recognition, quantity estimation, and spatial description. For instance, BiGenCap correctly identifies a newly constructed house on the right side of the scene, whereas Chg2Cap overestimates the number of changed objects and HiSem omits the spatial location. These cases illustrate BiGenCap's ability to capture both change semantics and spatial details.



Figure 2: Qualitative comparison on representative bitemporal image pairs. (a) Reference captions; (b)–(d) predictions from Chg2Cap, HiSem, and BiGenCap, respectively. Green and red text highlight correct and erroneous descriptions, respectively.

Table 2: Ablation of GDF and RRM under a common CE training protocol.

		LEVIR-CC					WHU-CDC				
GDF	RRM	BLEU-4	METEOR	ROUGE-L	CIDEr-D	S_{avg}	BLEU-4	METEOR	ROUGE-L	CIDEr-D	S_{avg}
×	×	65.75	40.48	75.56	139.40	80.30	76.88	49.39	82.94	161.53	92.69
✓	×	66.14	40.97	76.03	141.66	81.20	77.36	49.83	83.45	162.67	93.33
×	✓	66.15	40.84	75.91	140.67	80.89	77.55	49.68	83.47	162.47	93.29
✓	✓	**66.58**	**41.23**	**76.35**	**141.83**	**81.50**	**77.78**	**49.99**	**83.50**	**162.82**	**93.52**

Table 3: Effect of generative pretraining and bitemporal pairing under the same architecture.

	LEVIR-CC					WHU-CDC				
DiT Initialization	BLEU-4	METEOR	ROUGE-L	CIDEr-D	S_{avg}	BLEU-4	METEOR	ROUGE-L	CIDEr-D	S_{avg}
Random Init.	66.15	40.94	75.98	140.31	80.84	77.47	49.73	83.35	161.17	92.93
Shuffled-pair Pretraining	65.89	41.05	76.00	141.24	81.04	76.88	49.49	83.48	162.75	93.15
Genuine Paired Pretraining	**66.58**	**41.23**	**76.35**	**141.83**	**81.50**	**77.78**	**49.99**	**83.50**	**162.82**	**93.52**

4.3 Ablation Studies and Analysis

Generative discrepancy and relation modeling. We evaluate GDF and RRM under a common CE training protocol. All variants involving the DiT use the same pretrained initialization. GDF exploits prediction-derived discrepancy cues from the final DiT responses, whereas RRM uses intermediate DiT representations to model local cross-temporal relations. As shown in Table 2, either component consistently improves the baseline on both benchmarks, while their combination achieves the best results across all reported metrics. Specifically, S_{avg} increases from 80.30 to 81.50 on LEVIR-CC and from 92.69 to 93.52 on WHU-CDC. The additional gains obtained when combining GDF and RRM indicate that output-level prediction discrepancies and representation-level cross-temporal relations provide complementary information for change representation learning.

Joint generative pretraining. We further examine whether the benefit of generative pretraining mainly arises from exposure to additional remote sensing imagery or from preserving genuine bitemporal pairing. Under the same GDF+RRM architecture, we compare a randomly initialized DiT, a DiT pretrained on shuffled image pairs, and a DiT pretrained on genuine bitemporal pairs. Shuffling preserves the same individual images, data scale, and temporal marginals while removing the empirical dependency between paired observations, providing a controlled test of whether genuine pairing offers benefits beyond exposure to the same imagery. As shown in Table 3, shuffled-pair pretraining yields only limited and partly inconsistent gains over random initialization, whereas genuine-pair pretraining improves every reported metric over the shuffled setting on both benchmarks. S_{avg} increases from 81.04 to 81.50 on LEVIR-CC and from 93.15 to 93.52 on WHU-CDC. This distinction is particularly relevant to reciprocal prediction: with genuine pairs, the clean temporal counterpart provides informative context for predicting the perturbed stream, enabling the DiT to exploit cross-temporal dependencies in addition to within-stream semantics. These results suggest that the benefit of BiGenCap stems not only from large-scale remote sensing pretraining, but also from preserving genuine bitemporal pairing during generative modeling.

Feature prediction across corruption states. To probe how genuine pairing affects the learned joint predictor, we compare genuine-pair and shuffled-pair pretrained DiTs. We evaluate one-step clean-feature prediction over an 11×11 corruption grid, reusing identical Gaussian noise realizations for both models at each corruption state. Figure 3(a,b) reports the prediction-error difference

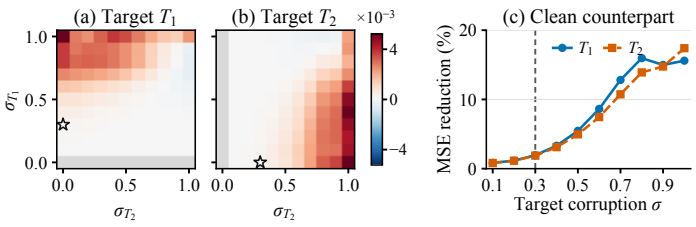


Figure 3: Feature prediction under genuine-pair and shuffled-pair pretraining. (a,b) Prediction-MSE differences across the corruption space; (c) relative MSE reduction with a clean counterpart.

Table 4: Sensitivity to the corruption configuration used for reciprocal generative prediction.

	LEVIR-CC					WHU-CDC				
$(\sigma_{\mathrm{tar}}, \sigma_{\mathrm{ctx}})$	BLEU-4	METEOR	ROUGE-L	CIDEr-D	S_{avg}	BLEU-4	METEOR	ROUGE-L	CIDEr-D	S_{avg}
$(0.3, 0)$	**66.58**	**41.23**	**76.35**	141.83	**81.50**	**77.78**	**49.99**	**83.50**	**162.82**	**93.52**
$(0.7, 0)$	66.26	41.13	76.25	**141.94**	81.39	77.59	49.81	83.38	162.39	93.29
$(0.7, 0.3)$	66.56	41.07	76.17	141.56	81.34	77.46	49.86	83.33	162.64	93.32

$E_{\mathrm{shuffled}} - E_{\mathrm{genuine}}$ across the corruption space, where positive values indicate lower error with genuine-pair pretraining. The stars and dashed line mark the default corruption setting used for captioning. Panel (c) further reports the relative MSE reduction along the boundary where the temporal counterpart remains clean. Genuine-pair pretraining reduces prediction error at 93 of the 110 nontrivial states (84.5%) in each prediction direction and at every evaluated target corruption level with a clean counterpart. Notably, the advantage becomes more pronounced as the target stream is increasingly corrupted while the counterpart remains clean. This behavior indicates that genuine-pair pretraining enables the joint DiT to exploit cross-temporal context when within-stream evidence becomes less reliable, providing further evidence for the cross-temporal dependencies learned from genuine bitemporal pairs.

Corruption configuration. We next examine the corruption configuration used during reciprocal prediction. We denote each setting as $(\sigma_{\mathrm{tar}}, \sigma_{\mathrm{ctx}})$ and compare the default $(0.3, 0)$ with stronger target corruption $(0.7, 0)$ and additional context corruption $(0.7, 0.3)$. As shown in Table 4, $(0.3, 0)$ achieves the highest S_{avg} on both benchmarks. Increasing the target corruption to 0.7 slightly reduces captioning performance, suggesting that excessive perturbation may remove useful instance-specific information from the target stream. At $\sigma_{\mathrm{tar}} = 0.7$, additionally perturbing the temporal counterpart provides no consistent improvement and remains below the default setting. These results favor a moderate target corruption while keeping the counterpart clean, and we therefore use $(0.3, 0)$ for reciprocal prediction.

Caption-stage flow regression. We examine the effect of retaining flow-matching supervision during CE adaptation. Since this objective is active only while the visual model remains trainable, Table 5 reports both CE-stage and final SCST performance. On LEVIR-CC, $\lambda_{\mathrm{FM}} = 0.01$ gives the best final S_{avg} of 82.55, compared with 82.11 without flow regression. On WHU-CDC, $\lambda_{\mathrm{FM}} = 0.001$ improves the final score from 93.86 to 94.17, although $\lambda_{\mathrm{FM}} = 0$ gives the highest CE-stage score. A larger weight of 0.05 is suboptimal on both benchmarks. These results indicate that flow regression is effective as a weak auxiliary objective during visual adaptation, while caption supervision remains the primary training signal.

Table 5: Effect of the flow-regression weight λ_{FM}. Each entry reports CE-stage/final SCST scores. Best and second-best results within each stage are shown in bold and underlined, respectively.

	LEVIR-CC					WHU-CDC				
λ_{FM}	BLEU-4	METEOR	ROUGE-L	CIDEr-D	S_{avg}	BLEU-4	METEOR	ROUGE-L	CIDEr-D	S_{avg}
0	66.58/66.99	**41.23**/41.55	76.35/76.70	141.83/143.21	81.50/82.11	77.78/77.15	**49.99**/50.12	83.50/83.91	162.82/164.26	**93.52**/93.86
0.001	**66.98**/67.24	41.03/41.47	76.20/76.68	141.52/142.88	81.43/82.07	76.95/76.99	49.88/**50.27**	83.54/**84.11**	**163.41/165.31**	93.45/**94.17**
0.01	66.70/**67.47**	41.19/**41.72**	**76.54/77.03**	**142.37/143.97**	**81.70/82.55**	77.50/77.57	49.75/50.11	**83.58**/84.08	162.56/164.23	93.35/94.00
0.05	66.32/66.77	40.96/41.37	76.00/76.49	141.12/142.60	81.10/81.81	**77.95/77.72**	49.88/50.14	83.50/84.00	162.37/163.99	93.42/93.96

5 CONCLUSION

We presented BiGenCap, which introduces joint generative modeling of bitemporal images into RSICC and instantiates the learned prior through reciprocal generative prediction. Experiments on two benchmarks demonstrate strong performance against previous RSICC methods, while ablations and feature-prediction analyses highlight the importance of genuine bitemporal pairing and the complementary contributions of prediction-derived discrepancies and intermediate cross-temporal relations. These findings suggest that a generative prior learned from paired imagery can provide complementary cross-temporal information for change representation beyond direct bitemporal comparison. Future work will explore more efficient integration of generative priors into captioning models.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript organization and language polishing, as well as literature retrieval and discovery. All AI-assisted content was carefully reviewed, verified, and revised by the authors, who take full responsibility for the final content of this work.

REFERENCES

Shizhen Chang and Pedram Ghamisi. Changes to captions: An attentive network for remote sensing change captioning. *IEEE Transactions on Image Processing*, 32:6047–6060, 2023.

Shimaa Holail, Tamer Saleh, Xiongwu Xiao, and Deren Li. Afde-net: Building change detection using attention-based feature differential enhancement for satellite imagery. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023.

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.

Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, pp. 228–231, 2007.

Xiliang Li, Bin Sun, Zhenhua Wu, Shutao Li, and Hu Guo. Cd4c: Change detection for remote sensing image change captioning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:9181–9194, 2025.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Haihan Lin, Xiaoqin Wang, Mengmeng Li, Dehua Huang, and Ruijiao Wu. A multi-task consistency enhancement network for semantic change detection in hr remote sensing images and application of non-agriculturalization. *Remote Sensing*, 15(21):5106, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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, 2022a.

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, 2023.

Chenyang Liu, Keyan Chen, Bowen Chen, Haotian Zhang, Zhengxia Zou, and Zhenwei Shi. Rscama: Remote sensing image change captioning with state space model. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024a.

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

Mengxi Liu, Zhuoqun Chai, Haojun Deng, and Rong Liu. A cnn-transformer network with multi-scale context aggregation for fine-grained cropland change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4297–4306, 2022b.

Mengxi Liu, Simin Lin, Yutong Zhong, Qian Shi, and Jiaqi Li. A memory-guided network and a novel dataset for cropland semantic change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024c.

Shenbo Liu, Dongxue Zhao, Yuheng Zhou, Ying Tan, Huang He, Zhao Zhang, and Lijun Tang. Network and dataset for multiscale remote sensing image change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:2851–2866, 2024d.

Yunlong Liu, Kai Zhang, Chunan Guan, Shanxin Zhang, Hong Li, Wenbo Wan, and Jiande Sun. Building change detection in earthquake: A multiscale interaction network with offset calibration and a dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024e.

Ziyuan Liu, Ruifei Zhu, Long Gao, Yuanxiu Zhou, Jingyu Ma, and Yuantao Gu. Jl1-cd: A new benchmark and a robust multi-teacher knowledge distillation framework for remote sensing change detection. *IEEE Transactions on Instrumentation and Measurement*, 2026.

Mubashir Noman, Noor Ahsan, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Cdchat: A large multimodal model for remote sensing change description. In *IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium*, pp. 7033–7037. IEEE, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1179–1195. IEEE, 2017.

Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. S2looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24):5094, 2021.

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

Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE transactions on geoscience and remote sensing*, 60:1–16, 2021.

Dongwei Sun, Jing Yao, Wu Xue, Changsheng Zhou, Pedram Ghamisi, and Xiangyong Cao. Mask approximation net: A novel diffusion model approach for remote sensing change captioning. *IEEE transactions on geoscience and remote sensing*, 63:1–11, 2025.

Dongwei Sun, Yuduo Wang, Jing Yao, Weikang Yu, Xiangyong Cao, and Pedram Ghamisi. Scnet: Lightweight spatial-channel attention network for remote sensing change captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

Xiaoliang Tan, Guanzhou Chen, Xiaodong Zhang, Tong Wang, Jiaqi Wang, Kui Wang, and Tingxuan Miao. Triples: Mitigating multi-task learning conflicts for semantic change detection in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 230:374–401, 2025.

Shiqi Tian, Ailong Ma, Zhuo Zheng, and Yanfei Zhong. Hi-ucd: A large-scale dataset for urban semantic change detection in remote sensing imagery. *arXiv preprint arXiv:2011.03247*, 2020.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Man Wang, Chenyang Liu, Wenjun Li, Feng Ni, Bing Jia, Baoqi Huang, Riting Xia, and Zhenwei Shi. Hisem: Hierarchical semantic disentangling for remote sensing image change captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 64:5628117–5628117, 2026a.

Qi Wang, Wei Jing, Kaichen Chi, and Yuan Yuan. Cross-difference semantic consistency network for semantic change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024a.

Yelin Wang, Zijia Song, Shuo Ye, Chuanguang Yang, Miaoyu Wang, Yong Xu, Zhulin An, Yongjun Xu, and Zitong Yu. Rsicclm: A multimodal large language model for remote sensing image change captioning. In *European Conference on Computer Vision*, pp. 253–272. Springer, 2026b.

Yuduo Wang, Weikang Yu, Michael Kopp, and Pedram Ghamisi. Changeminds: Multi-task framework for detecting and describing changes in remote sensing. *arXiv preprint arXiv:2410.10047*, 2024b.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Cong Yang, Zuchao Li, Hongzan Jiao, Zhi Gao, and Lefei Zhang. Enhancing perception of key changes in remote sensing image change captioning. *IEEE Transactions on Image Processing*, 2025.

Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2021.

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.

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

Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:183–200, 2020.

Haotian Zhang, Han Guo, Keyan Chen, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Foba: A foreground–background co-guided method and new benchmark for remote sensing semantic change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–19, 2025.

Jindou Zhang, Zhenfeng Shao, Qing Ding, Xiao Huang, Yu Wang, Xuechao Zhou, and Deren Li. Aernet: An attention-guided edge refinement network and a dataset for remote sensing building change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

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.

Yanpeng Zhou, Jinjie Wang, Jianli Ding, Bohua Liu, Nan Weng, and Hongzhi Xiao. Signet: A siamese graph convolutional network for multi-class urban change detection. *Remote Sensing*, 15 (9):2464, 2023.

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

Shiwei Zou, Yingmei Wei, Yuxiang Xie, and Xidao Luan. Frequency–spatial–temporal domain fusion network for remote sensing image change captioning. *Remote Sensing*, 17(8):1463, 2025.

A Additional Method Details

A.1 Joint DiT Architecture and Conditioning

The joint generative model is implemented with a ten-block Diffusion Transformer (DiT) (Peebles & Xie, 2023) with hidden width $D = 512$ and eight attention heads. The DiT module contains 42.5M parameters and is shared across both reciprocal prediction directions. A shared affine projection P maps the input visual features to the DiT hidden space. For temporal stream T_q, $q \in \{1, 2\}$, we add a corruption embedding $\mathbf{e}_q(\tau_q)$ and a learned temporal-role embedding $\mathbf{r}_q$ to each spatial token:

$$\begin{aligned} \mathbf{H}_0^{T_q} &= P(\mathbf{Z}_{\tau_q}^{T_q}) + \mathbf{e}_q(\tau_q) + \mathbf{r}_q, \\ \mathbf{H}_0 &= \mathrm{Concat}_{\mathrm{tokens}} \left(\mathbf{H}_0^{T_1}, \mathbf{H}_0^{T_2} \right), \\ \mathbf{c} &= g([\boldsymbol{\gamma}_1(\tau_1), \boldsymbol{\gamma}_2(\tau_2)]) \,. \end{aligned} \tag{12}$$

Here, $\mathbf{e}_q(\tau_q)$ provides token-level conditioning on the corruption state of each temporal stream, while $\boldsymbol{\gamma}_q(\tau_q)$ denotes an independently learned global corruption embedding. The ordered global embeddings are mapped by g to the conditioning vector $\mathbf{c}$ used for adaptive normalization in the DiT blocks. Rotary positional encodings incorporate spatial coordinates and temporal-stream identity. Joint self-attention operates over the concatenated temporal token groups, and a shared output projection maps the final hidden states to C_z velocity channels for each stream.

A.2 Generative Discrepancy Fusion Details

The residual projection P_R in Equation (6) is implemented as a 1×1 convolution followed by batch normalization and ReLU, while P_g is a 1×1 convolution. Both projections are shared across the two temporal streams. The interaction stack starts from the gated features in Equation (6), with $\mathbf{P}_0 = \widetilde{\mathbf{X}}^{T_1}$ and $\mathbf{Q}_0 = \widetilde{\mathbf{X}}^{T_2}$. For a directional update, let $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{N \times C}$ denote the receiving and counterpart streams, respectively. We first introduce local spatial context through lightweight depthwise spatial operators:

$$\overline{\mathbf{P}} = \mathbf{P} + \mathcal{D}_q(\mathbf{P}), \qquad \overline{\mathbf{Q}} = \mathbf{Q} + \mathcal{D}_k(\mathbf{Q}),$$

where $\mathcal{D}_q$ and $\mathcal{D}_k$ consist of depthwise spatial filtering followed by normalization, nonlinearity, and channel projection. The directional interaction is then computed as

$$\begin{aligned} \mathbf{T}_{PQ} &= \psi \left(\overline{\mathbf{P}} \odot (\overline{\mathbf{Q}} - \overline{\mathbf{P}}) \right), \\ \mathbf{U}_{PQ} &= \mathrm{LN} \left(\overline{\mathbf{P}} + \mathrm{MHA}(\overline{\mathbf{P}}, \mathbf{T}_{PQ}, \mathbf{T}_{PQ}) \right), \\ \mathcal{T}(\mathbf{P}, \mathbf{Q}) &= \mathrm{LN} \left(\mathbf{U}_{PQ} + \mathrm{FFN}(\mathbf{U}_{PQ}) \right). \end{aligned} \tag{13}$$

Here, ψ denotes a channel projection with normalization and nonlinearity; MHA, LN, and FFN denote multi-head attention, layer normalization, and a token-wise feed-forward network, respectively. In the MHA operation, the receiving stream provides the query, while the difference-conditioned feature $\mathbf{T}_{PQ}$ provides the keys and values. We use $L_{\mathrm{vis}} = 3$ bitemporal interaction blocks, each performing sequential updates in the two temporal directions:

$$\mathbf{P}_b = \mathcal{T}_{b,1}(\mathbf{P}_{b-1}, \mathbf{Q}_{b-1}), \qquad \mathbf{Q}_b = \mathcal{T}_{b,2}(\mathbf{Q}_{b-1}, \mathbf{P}_b).$$

Thus, the second directional update uses the newly updated counterpart stream. After the interaction stack, the two streams are concatenated and processed by channel projection and a convolutional residual block to produce the base visual feature $\mathbf{M}_0$, from which $\mathbf{K}_l^0$ and $\mathbf{V}_l^0$ are constructed.

A.3 Reciprocal Relation Modeling Details

Intermediate representations and adapters. RRM extracts intermediate representations after the fifth DiT block. The adapters $\mathcal{A}$ and $\mathcal{B}$ in Equation (7) are shared across the two temporal streams and output $C_r = 256$ dimensional features. Adapter $\mathcal{A}$ applies RMS normalization, an affine projection, and output RMS normalization. Adapter $\mathcal{B}$ encodes both the normalized direction and magnitude of the reciprocal hidden-state difference, using a projected normalized difference together with a nonlinear embedding of its RMS magnitude. This allows $\mathbf{D}^{T_q}$ to characterize how the intermediate representation changes between the clean and perturbed roles.

Local relation estimation. The neighborhood $\mathcal{N}(i)$ in Equation (8) is a 5×5 local window centered at the same grid location in the temporal counterpart, excluding positions outside the feature map. The temperature κ is learned jointly with the model, and relation weights are normalized independently in the two temporal directions. Importantly, the weights are determined solely from the clean-role features $\mathbf{A}$; the role-change features $\mathbf{D}$ are incorporated later when constructing the relation feature $\mathbf{k}$. The same relation weights aggregate both the intermediate relation features and the original visual content.

Relation and content projections. The mappings f_a and f_d in Equation (9a) and f_c and f_v in Equation (9b) are two-layer MLPs with GELU activations and output width C. The content mapping f_c is shared across the two temporal streams, while the relation and content branches use separate RMS normalizations. The layer-specific affine maps $\Delta\mathcal{K}_l$ and $\Delta\mathcal{V}_l$ in Equation (10) project the relation and content features to the Transformer-decoder width, where α_K and α_V are learnable scales. These residual projections refine the GDF-derived base keys and values with cross-temporal relation and content information.

A.4 Caption Decoder and Learning Objectives

Autoregressive caption generation. The caption generator is an autoregressive Transformer decoder that attends to the layer-wise visual context $\mathcal{C} = \{(\mathbf{K}_l, \mathbf{V}_l)\}_{l=1}^{L_{\text{dec}}}$. For a caption $\mathbf{Y} = (y_1, \ldots, y_T)$, its conditional probability is factorized as

$$p\left(\mathbf{Y} \mid \mathbf{I}^{T_1}, \mathbf{I}^{T_2}\right) = \prod_{k=1}^{T} p_\phi\left(y_k \mid y_{<k}, \mathcal{C}\right). \tag{14}$$

Causal self-attention models the caption prefix, while cross-attention retrieves visual information from $\mathcal{C}$. For attention head a of decoder layer l, the visual response is

$$\mathbf{O}_l^{(a)} = \text{softmax}\left(\frac{\mathbf{Q}_l^{(a)}(\mathbf{K}_l^{(a)})^\intercal}{\sqrt{d_h}}\right)\mathbf{V}_l^{(a)}, \tag{15}$$

where d_h is the head dimension, $\mathbf{Q}_l^{(a)}$ is projected from the caption states after causal self-attention, and $\mathbf{K}_l^{(a)}$ and $\mathbf{V}_l^{(a)}$ are the corresponding head-specific visual keys and values. Relation information injected into the keys therefore influences visual selection through the attention weights, while the values provide the corresponding relation-enriched visual content.

Caption training objectives. For a minibatch of M image pairs, let T_n denote the number of valid target tokens for example n, including the end token. The caption cross-entropy objective is

$$\mathcal{L}_{\text{CE}} = -\frac{1}{M}\sum_{n=1}^{M}\frac{1}{T_n}\sum_{k=1}^{T_n}\log p_\phi\left(y_{n,k} \mid y_{n,<k}, \mathcal{C}_n\right). \tag{16}$$

During CE training, each reciprocal prediction also retains velocity supervision on its perturbed stream. Averaging over the two directions gives

$$\mathcal{L}_{\text{FM}}^{\text{cap}} = \frac{1}{2MNC_z}\sum_{n=1}^{M}\sum_{s\in\{T_1,T_2\}}\left\|\mathbf{V}_{s,n}^{[s]} - \left(\boldsymbol{\epsilon}_n^s - \mathbf{Z}_{0,n}^s\right)\right\|_F^2. \tag{17}$$

Decoder refinement. After CE training, the visual encoder, joint DiT, GDF, and RRM are frozen, and only the Transformer decoder is refined using self-critical sequence training (SCST) (Rennie et al., 2017). For each example, a sampled caption $\mathbf{Y}_n^{\text{sam}}$ and a greedy caption $\mathbf{Y}_n^{\text{greedy}}$ share the same fixed visual context. Let

$$\begin{aligned} a_n &= r_n(\mathbf{Y}_n^{\text{sam}}) - r_n(\mathbf{Y}_n^{\text{greedy}}), \\ \mathcal{L}_{\text{SCST}} &= -\frac{1}{M}\sum_{n=1}^{M}\text{sg}(a_n)\sum_{k=1}^{T_n^{\text{sam}}}\log p_\phi\left(y_{n,k}^{\text{sam}} \mid y_{n,<k}^{\text{sam}}, \mathcal{C}_n\right), \\ \mathcal{L}_2 &= 0.2\mathcal{L}_{\text{CE}} + 0.8\mathcal{L}_{\text{SCST}}. \end{aligned} \tag{18}$$

Here, $r_n(\mathbf{Y})$ is the mean of BLEU-4, METEOR, ROUGE-L, and CIDEr-D against the reference captions, sg denotes stop-gradient, and T_n^{sam} is the sampled caption length.

Table 6: Composition of the 229,642-pair generative pretraining corpus.

Data source	Image pairs	Resolution (m)	Geographic coverage
JL1-CD (Liu et al., 2026)	5,000	0.5–0.75	Multiple regions (China)
TUE-CD (Liu et al., 2024e)	1,656	–	Southern Turkey
MSRSCD (Liu et al., 2024d)	3,364	0.5	Southern cities (China)
EGY-BCD (Holail et al., 2023)	6,091	–	Four cities (Egypt)
HRCUS-CD (Zhang et al., 2023)	11,388	–	–
DSIFN (Zhang et al., 2020)	3,988	–	–
CLCD (Liu et al., 2022b)	600	0.5–2	Guangdong (China)
SYSU-CD (Shi et al., 2021)	20,000	0.5	Hong Kong (China)
WHU-CD (Ji et al., 2018)	6,690	–	–
LEVIR-MCI (Liu et al., 2024b)	8,148	0.5	Texas (USA)
FZ-SCD (Lin et al., 2023)	8,719	–	Fuzhou (China)
CCSCD5 (Tan et al., 2025)	953	–	Central China
SCSCD7 (Tan et al., 2025)	1,722	–	Southern China
CropSCD (Liu et al., 2024c)	4,141	0.5–2	Guangdong (China)
CNAM-CD (V1) (Zhou et al., 2023)	2,504	–	National new areas (China)
Hi-UCD (Tian et al., 2020)	40,800	0.1	Tallinn (Estonia)
S2Looking (Shen et al., 2021)	80,000	–	–
WHU-GCD (Zan et al., 2025)	7,396	–	–
JL1-SCD (Wang et al., 2024a)	7,950	0.75	–
LevirSCD (Zhang et al., 2025)	3,870	1–2	Beijing (China)
SECOND (Yang et al., 2021)	4,662	–	Hangzhou, Chengdu, Shanghai (China)
Total	229,642		

Dashes indicate unavailable or unverified metadata.

B Additional Experimental Details

Pretraining corpus. We collect 229,642 image pairs from 21 public remote sensing data sources covering building changes and land-cover transitions. For sources overlapping with the downstream benchmarks, the corresponding test pairs are excluded before pretraining. Table 6 summarizes the retained sample counts, and Figure 4 shows one illustrative bitemporal pair from each source. The examples span built-up areas, transportation infrastructure, cropland, water bodies, and vegetation, illustrating the visual diversity of the corpus.

Generative pretraining optimization. We pretrain the joint DiT for 400,000 optimization steps using AdamW with a constant learning rate of 10^{-4}, weight decay 10^{-2}, and a global batch size of 512 on four H800 GPUs. The visual encoder remains fixed throughout pretraining, while all DiT parameters are optimized. Genuine-pair and shuffled-pair pretraining use the same optimization settings. Corruption-space sampling and stream-wise loss masking follow Section 3.2.2.

Captioning model configuration and optimization. The captioning model uses a MiT-B1 backbone. The joint DiT receives 16×16 third-stage features of width $C_z = 320$, while multiscale content features are projected to width $C = 512$. The caption decoder has one layer and eight attention heads, and reciprocal prediction uses $\sigma = 0.3$ by default; the remaining GDF and RRM settings are given in Appendix A. Caption CE training uses AdamW with batch size 32, weight decay 10^{-4}, and constant learning rates of 5×10^{-5} for the backbone and 10^{-4} for the joint DiT, GDF, RRM, and caption decoder.

Feature-prediction diagnostics. The feature-prediction analysis in Figure 3 uses the genuine-pair and shuffled-pair pretrained DiTs at step 400,000, before caption training, with the frozen visual encoder and identical preprocessing. For the 1,929 LEVIR-CC test pairs, identical Gaussian noise realizations are reused across the two models and all $(\sigma_{T_1}, \sigma_{T_2}) \in \{0, 0.1, \ldots, 1\}^2$ corruption states. This analysis is used only as a post-hoc diagnostic and does not affect model selection.

For stream s, the one-step clean-feature estimate is $\widehat{\mathbf{Z}}_0^s = \mathbf{Z}_{\sigma_s}^s - \sigma_s \mathbf{V}_\theta^s$. Prediction MSE is computed as $\|\widehat{\mathbf{Z}}_0^s - \mathbf{Z}_0^s\|_F^2/(NC_z)$ and averaged over image pairs. States with $\sigma_s = 0$ are excluded because



Figure 4: Illustrative bitemporal image pairs from the 21 sources. Within each source, the upper and lower images correspond to T_1 and T_2, respectively.

the estimate then equals the clean input independently of the predictor, leaving 110 nontrivial states per stream. Figure 3(c) reports the relative MSE reduction $100(E_{\text{shuffled}} - E_{\text{genuine}})/E_{\text{shuffled}}$ along the clean-counterpart boundary. Under the linear probability path, prediction MSE is proportional to velocity MSE by σ_s^2 at a fixed corruption state, so both measures yield the same relative reduction.