CoDeRNet: Selective Cross-Task Routing under Heterogeneous Supervision for Change Detection and Captioning

Eunki Cho **Hyeon Bae Kim** **Seong Tae Kim**

Kyung Hee University, Republic of Korea
{cekkec, hyeonbae.kim, st.kim}@khu.ac.kr

Abstract

Change Detection & Captioning (CDC) aims to jointly localize changed regions and describe their semantic content from bi-temporal remote sensing images. While recent approaches have demonstrated the effectiveness of unified modeling, they primarily rely on shared representations or jointly optimized architectures, leaving task-specific communication under heterogeneous supervision less explored. However, detection and captioning differ fundamentally in their supervision granularity, leading to heterogeneous representations with uneven reliability across tasks and spatial locations. In this paper, we reformulate CDC as a selective information routing problem between heterogeneous task branches, where the key challenge is to determine when and where information should be transferred. To this end, we propose CoDeRNet, a unified framework that preserves task-specific representations while enabling bidirectional exchange of complementary information. CoDeRNet introduces a **Co**nfidence-**De**composed **R**outing (CoDeR) mechanism, which decomposes the routing condition into receiver-side need for support and sender-side reliability cues to guide selective transfer. Experiments on LEVIR-MCI and WHU-CDC show that CoDeRNet achieves strong overall CDC performance, particularly improving caption generation while preserving competitive change localization. Extensive analyses further show that the benefits arise not merely from joint learning or generic cross-task communication, but from explicitly modeling selective cross-task information routing. These results highlight the importance of conditional information transfer for unified CDC under heterogeneous supervision.

1 Introduction

Understanding changes between bi-temporal remote sensing observations is a fundamental problem in remote sensing, with broad applications including urban planning [14], natural disaster assessment [43], and environmental monitoring [16]. Two representative tasks have been widely studied in this context: Change Detection (CD) [1, 7, 9, 11, 13, 18, 19], which localizes changed regions at the pixel level, and Change Captioning (CC) [4, 22–24, 32, 36, 37, 49], which generates natural-language descriptions of semantic changes. Although their output formats differ, both tasks are grounded in the same underlying change event and provide complementary views of the same phenomenon. Motivated by this commonality, recent studies have introduced the unified setting of Change Detection & Captioning (CDC) [25, 35]. This unified setting is particularly appealing because it allows a single model to jointly reason about both where a change occurs and what the change means, offering a more complete form of change interpretation than either task alone.

However, these tasks differ fundamentally in their supervision granularity. Detection is densely supervised at the pixel level, whereas captioning is weakly supervised at the sentence level. This

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

heterogeneity introduces a structural mismatch in how information is learned and represented across the two tasks. Existing CDC approaches primarily rely on shared representations or jointly optimized architectures [25, 35], establishing a unified framework for learning spatial localization and semantic description together. While this paradigm is effective, it mainly emphasizes joint modeling of the two tasks, leaving the role of task-specific communication under heterogeneous supervision less explored. Importantly, uniform interaction or jointly learned interaction mechanisms implicitly assume that cross-task information is equally reliable and equally useful across spatial locations and tasks. However, under heterogeneous supervision, task representations often exhibit asymmetric reliability and support requirements, making unconditional information transfer suboptimal. Therefore, the central challenge is to determine how complementary information should be selectively transferred between heterogeneous task branches, rather than treating CDC merely as a joint optimization problem.

This perspective is particularly important in CDC because the usefulness of cross-task information can vary across both tasks and spatial locations. For example, detection-oriented features may provide strong spatial evidence around changed regions and boundaries, but they have limited capacity to explicitly organize such evidence into sentence-level semantic descriptions. Conversely, captioning-oriented features may capture global context and semantic relations, but sentence-level supervision does not explicitly constrain their spatial precision. This suggests that cross-task communication in CDC should be spatially and task-wise conditional, rather than uniformly applied across branches. Each branch should receive complementary information only when its own representation is insufficient and the other branch provides reliable supporting evidence.

Based on this motivation, we reformulate CDC as a selective information routing problem between heterogeneous task branches. To this end, we propose CoDeRNet, a unified framework that preserves task-specific branch identities while enabling bidirectional exchange of complementary information. CoDeRNet performs selective transfer via Confidence-Decomposed Routing (CoDeR), which decomposes the routing condition into receiver-side need for support and sender-side reliability cues. An overview of the proposed framework is shown in Fig. 1.

We demonstrate that CoDeRNet achieves strong overall CDC performance on representative benchmarks, including LEVIR-MCI and WHU-CDC, with notable gains in caption generation while maintaining competitive change localization. Moreover, through systematic analyses of joint learning, routing directionality, mask decomposition, and alternative interaction operators, we show that the gains arise not merely from joint learning or generic interaction, but from explicitly modeling selective cross-task routing. The code is publicly available at `https://github.com/cekkec/CoDeRNet`.

The main contributions of this paper are summarized as follows.

- We reformulate CDC as a selective cross-task information routing problem under heterogeneous supervision, where complementary information should be transferred in a conditional and task-dependent manner.
- We propose CoDeRNet, a unified CDC framework that preserves the distinct roles of detection and captioning while enabling task-aware bidirectional routing. Its core mechanism, Confidence-Decomposed Routing (CoDeR), guides selective transfer by decomposing the routing condition into receiver-side need for support and sender-side reliability cues.
- We validate CoDeRNet on representative CDC benchmarks, including LEVIR-MCI and WHU-CDC, and demonstrate through systematic analyses of joint learning, routing directionality, mask decomposition, and alternative interaction operators the effectiveness of selective cross-task routing under heterogeneous supervision.

2 Related Work

Remote Sensing Change Detection. Remote sensing change detection (RSCD) aims to identify meaningful changes from bi-temporal images of the same geographic area. Deep RSCD methods have evolved from CNN-based fully convolutional and Siamese encoder-decoder architectures [5, 8, 9, 11, 13, 18, 19, 47] to Transformer-based models that capture long-range dependencies and global context [1, 6, 12, 30, 34, 45, 48]. These studies have improved change localization by enhancing multi-scale representation learning, temporal feature comparison, and contextual reasoning over bi-temporal observations. More recently, Mamba-based architectures have been explored for efficient

spatiotemporal modeling [7, 46], further expanding RSCD toward scalable and globally aware change representation.

Remote Sensing Image Change Captioning. Remote sensing image change captioning (RSICC) aims to generate natural-language descriptions of semantic changes from bi-temporal images. Existing methods have improved change captioning through encoder-decoder architectures [32], Transformer-based cross-temporal modeling and multi-scale change representation [22, 23], pretrained language models and decoupled reasoning [24], attentive or lightweight architectures [4, 36], bi-temporal pretraining [49], and diffusion-based change modeling [37]. These approaches have progressively strengthened the ability to localize visual change cues, model temporal differences, and translate them into semantically coherent descriptions.

Remote Sensing Change Detection & Captioning. Remote sensing change detection and captioning has recently emerged as a unified task for comprehensive change interpretation, combining pixel-level localization with semantic-level change description. MCINet [25] introduced LEVIR-MCI and proposed a dual-branch framework that jointly performs change detection and change captioning. CTMTNet [35] further formalized this direction as the RS-CDC task and developed a shared multitask architecture with task-specific decoders. These studies established unified CDC as a promising direction for jointly modeling spatial and semantic change information.

3 Method

3.1 Overview

CoDeRNet addresses remote sensing change detection and change captioning within a unified yet selective information routing framework. Given a bi-temporal image pair (I_{t_0}, I_{t_1}), the model first extracts multi-scale visual features from the two images and constructs shared change representations. These shared change features are then passed through branch-specific interfaces that provide separate feature spaces for detection and captioning. On top of these branch-specific features, Confidence-Decomposed Routing (CoDeR) selectively routes complementary information between the two branches. Finally, task-specific decoders produce the change mask and the change description, respectively. As illustrated in Fig. 1, the entire framework is optimized end-to-end under a joint training objective.

3.2 Shared Change Representation

Given a bi-temporal image pair (I_{t_0}, I_{t_1}), where I_{t_0} and I_{t_1} denote the earlier and later observations, respectively, we first extract multi-scale visual features from the two images using a shared encoder. Let $F_{t_0}^{(s)}, F_{t_1}^{(s)} \in \mathbb{R}^{B\times C_s\times H_s\times W_s}$ denote the features of the two observations at scale s. To construct a shared representation of temporal change, we combine the two temporal features with two complementary change cues: a magnitude change cue, which captures the strength of change, and a signed change cue, which preserves directional temporal differences. The resulting change representation at each scale is defined as

$$F^{(s)} = \mathrm{Conv}_{1\times1}\big([F_{t_0}^{(s)}, F_{t_1}^{(s)}, F_{|\Delta|}^{(s)}, F_{\Delta}^{(s)}]\big),$$

where $F_{|\Delta|}^{(s)} = |F_{t_1}^{(s)} - F_{t_0}^{(s)}|$ and $F_{\Delta}^{(s)} = F_{t_1}^{(s)} - F_{t_0}^{(s)}$ denote the magnitude and signed change cues.

3.3 Branch-Specific Feature Interface

Before cross-task routing, we instantiate branch-specific feature interfaces from the shared change representation. This step provides separate feature spaces in which CoDeR can estimate branch-dependent routing cues and apply residual information transfer:

$$F_{\mathrm{seg}}^{(s)} = \phi_{\mathrm{seg}}^{(s)}(F^{(s)}), \qquad F_{\mathrm{cap}}^{(s)} = \phi_{\mathrm{cap}}^{(s)}(F^{(s)}),$$

where $\phi_{\mathrm{seg}}^{(s)}$ and $\phi_{\mathrm{cap}}^{(s)}$ are lightweight projection layers. These interfaces provide the feature spaces on which selective information routing is performed, while the actual cross-task transfer is handled by CoDeR. The same interfaces are used for all compared variants, so that subsequent comparisons focus on the routing mechanism rather than differences in branch construction.

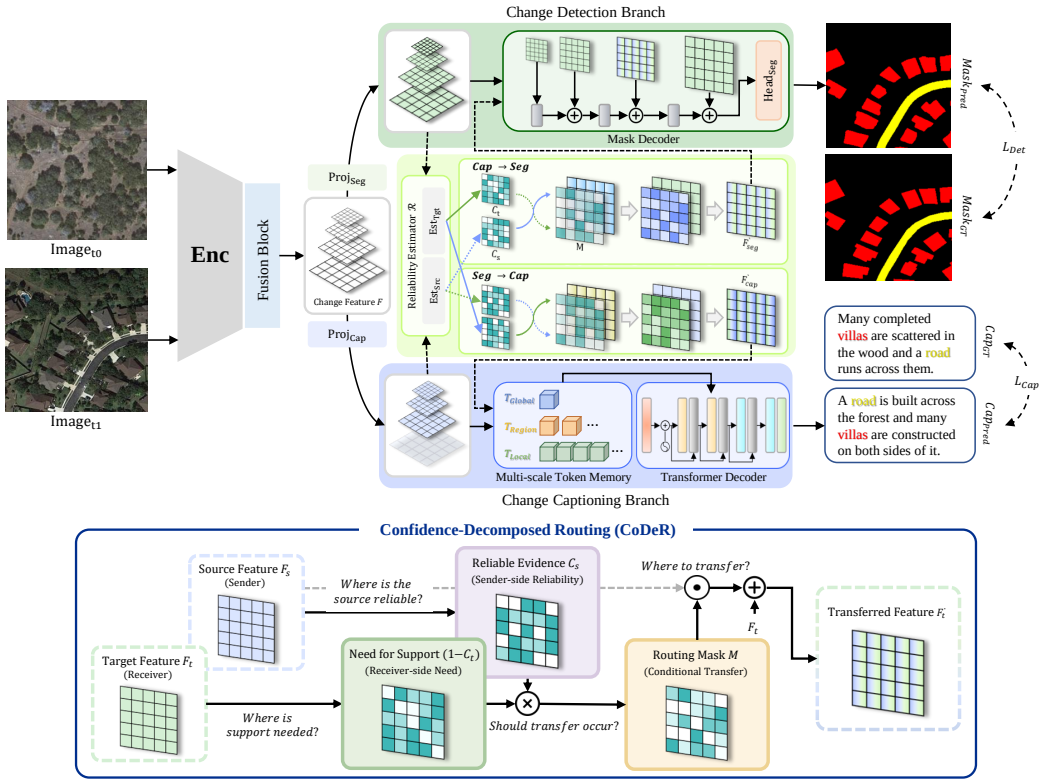


Figure 1: Overview of CoDeRNet. Shared change features are transformed into task-specific detection and captioning features, followed by Confidence-Decomposed Routing (CoDeR) for selective bidirectional transfer before task-specific decoding.

3.4 Confidence-Decomposed Routing

After branch-specific interfaces are constructed, the two branches encode different yet complementary aspects of the same change. The detection branch tends to provide stronger cues for modeling where a change occurs, whereas the captioning branch tends to provide richer cues for reasoning about what the change means. Rather than directly fusing the two branches, we explicitly model where cross-task transfer is likely to be beneficial. Consider a target feature map $F_t \in \mathbb{R}^{B\times C\times H\times W}$ from the receiving branch and a source feature map $F_s \in \mathbb{R}^{B\times C\times H\times W}$ from the sending branch. CoDeR estimates target-side sufficiency and source-side reliability cues using lightweight routing estimators and derives a routing mask as

$$C_t = \mathcal{R}_t(F_t), \qquad C_s = \mathcal{R}_s(F_s), \qquad M = (1 - C_t) \cdot C_s,$$

where $\mathcal{R}_t(\cdot) = \sigma(h_t(\cdot))$ and $\mathcal{R}_s(\cdot) = \sigma(h_s(\cdot))$ are lightweight routing estimators implemented by 1×1 convolution heads with sigmoid activation, and $C_t, C_s, M \in [0, 1]^{B\times 1\times H\times W}$. Here, C_t and C_s are not calibrated predictive probabilities, but task-dependent routing cues learned to estimate where cross-task transfer is expected to be complementary. The term $(1 - C_t)$ represents the receiver-side need for support, while C_s represents the sender-side reliability cue. Their multiplicative combination acts as a soft conjunction, making the routing mask large only when the receiving branch requires support and the sending branch provides reliable evidence. In this way, CoDeR routes information conditionally rather than applying cross-task transfer uniformly across all spatial locations.

The source feature is then mapped to the target feature space through a lightweight adapter g and selectively injected via residual transfer:

$$F_t' = F_t + M \odot g(F_s),$$

where $\odot$ denotes element-wise multiplication with spatial broadcasting. When M is small, the original target representation is largely preserved; when M is large, the target branch receives

Table 1: Comparison on LEVIR-MCI [25]. "Change Detection" and "Change Captioning" denote single-task models trained for only one task, whereas "Change Detection & Captioning" denotes unified models that perform both tasks jointly within a single framework. **Bold** and underlined values indicate the best and second-best results, respectively.

Task	Method	Metrics							
		mIoU	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Change Detection	SNUNet [11]	82.53	-	-	-	-	-	-	-
	ChangeFormer [1]	84.35	-	-	-	-	-	-	-
	TFI-GR [18]	79.34	-	-	-	-	-	-	-
	A2Net [19]	80.82	-	-	-	-	-	-	-
	HANet [13]	71.23	-	-	-	-	-	-	-
	ChangeMamba [7]	87.61	-	-	-	-	-	-	-
Change Captioning	DUDA [32]	-	81.44	72.22	64.24	57.79	37.15	71.04	124.32
	MCCFormers [33]	-	82.16	72.95	65.42	59.41	38.26	72.10	128.34
	RSICCformer [22]	-	84.11	75.40	68.01	61.93	38.79	73.02	131.40
	Prompt-CC [24]	-	83.66	75.73	69.10	63.54	38.82	73.72	136.44
	Chg2Cap [4]	-	85.14	76.91	69.86	64.09	39.83	74.62	135.41
	SparseFocus [36]	-	84.56	75.87	68.64	62.87	39.93	74.69	137.05
	SEN [49]	-	85.10	77.05	70.01	64.09	39.59	74.57	136.02
	MaskApproxNet [37]	-	85.90	77.12	70.72	64.32	39.91	75.67	137.71
Change Detection & Captioning	CTMTNet [35]	85.37	85.16	77.24	70.13	64.16	39.34	74.03	134.50
	MCINet [25]	86.43	86.68	78.75	71.74	65.95	40.80	75.96	140.29
	CoDeRNet (Ours)	**87.79**	**87.25**	**79.42**	**72.26**	**66.27**	**41.08**	**76.36**	**141.79**

complementary evidence from the source branch. This selective residual update allows CoDeRNet to preserve branch-specific representations while refining regions that may benefit from cross-task support.

CoDeR is applied bidirectionally, i.e., from captioning to detection and from detection to captioning, with independent routing estimators and adapters for the two directions:

$$F_{\mathrm{seg}}^{\prime(s_g)} = \mathrm{CoDeR}_{\mathrm{cap}\to\mathrm{seg}}\big(F_{\mathrm{seg}}^{(s_g)}, F_{\mathrm{cap}}^{(s_g)}\big), \quad F_{\mathrm{cap}}^{\prime(s_c)} = \mathrm{CoDeR}_{\mathrm{seg}\to\mathrm{cap}}\big(F_{\mathrm{cap}}^{(s_c)}, F_{\mathrm{seg}}^{(s_c)}\big).$$

Within each direction, the source and target features are taken from the same spatial scale, which allows the routing mask to be applied directly without additional resizing. The design is asymmetric across the two routing directions: the detection branch receives routed information at a relatively higher-resolution scale, whereas the captioning branch receives routed information at a lower-resolution scale. This design reflects the different spatial granularity requirements of the two tasks.

3.5 Task-Specific Decoders

After branch-specific interfaces and CoDeR-based routing, each branch maintains a multi-scale feature hierarchy. CoDeR is applied only at a designated receiving scale for each branch, while the remaining scales are preserved as branch-specific features. The resulting hierarchies are then decoded by task-specific heads that reflect the distinct granularity and inductive bias required by change detection and change captioning. For notational simplicity, we use the primed feature notation to denote the branch features after CoDeR. At scales without CoDeR, the primed features are identical to the original branch features.

Segmentation Decoder. Change detection requires spatially precise prediction with progressive recovery of spatial detail. We therefore use an FPN-style top-down decoder [21] to aggregate the multi-scale segmentation branch features. The decoder progressively combines coarse semantic features with higher-resolution features and produces pixel-wise class logits:

$$\hat{M} = \mathcal{D}_{\mathrm{seg}}\big(\{F_{\mathrm{seg}}^{\prime(s)}\}_{s\in\mathcal{S}}\big) \in \mathbb{R}^{B\times K\times H\times W},$$

where $\mathcal{S}$ denotes the set of feature scales and K denotes the number of change classes.

Captioning Decoder. Unlike segmentation, caption generation requires organizing change evidence into semantic units. We therefore construct a multi-granularity memory from the caption-oriented branch features:

$$\mathcal{M} = [\, t_{\mathrm{glob}};\, T_{\mathrm{reg}};\, T_{\mathrm{loc}}\,], \qquad \hat{p}_{t+1} = \mathcal{D}_{\mathrm{cap}}(y^*_{1:t}, \mathcal{M}).$$

Table 2: Comparison on WHU-CDC [35]. "Change Detection" and "Change Captioning" denote single-task models trained for only one task, whereas "Change Detection & Captioning" denotes unified models that perform both tasks jointly within a single framework. **Bold** and underlined values indicate the best and second-best results, respectively.

Task	Method	Metrics							
		IoU	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Change Detection	SNUNet [11]	74.81	-	-	-	-	-	-	-
	ChangeFormer [1]	77.88	-	-	-	-	-	-	-
	TFI-GR [18]	77.60	-	-	-	-	-	-	-
	A2Net [19]	80.43	-	-	-	-	-	-	-
	HANet [13]	79.95	-	-	-	-	-	-	-
	ChangeMamba [7]	89.69	-	-	-	-	-	-	-
Change Captioning	DUDA [32]	-	79.04	69.53	61.57	55.64	34.29	68.98	121.85
	MCCFormers [33]	-	81.12	75.04	69.95	65.34	42.11	78.52	147.09
	RSICCformer [22]	-	80.05	74.24	69.61	66.54	42.65	73.91	133.44
	Prompt-CC [24]	-	81.12	73.96	67.22	61.45	36.99	71.88	134.50
	Chg2Cap [4]	-	77.27	71.33	65.10	58.44	39.95	78.67	160.64
	SparseFocus [36]	-	81.17	72.90	66.06	60.27	37.34	72.63	134.64
	SEN [49]	-	80.60	74.64	67.69	61.97	36.76	71.70	133.57
	MaskApproxNet [37]	-	81.34	75.68	71.16	67.73	43.89	75.41	135.31
Change Detection & Captioning	CTMTNet [35]	81.54	83.56	77.66	72.76	69.00	45.39	79.23	149.40
	MCINet [25]	88.23	**87.82**	**82.84**	**78.66**	**75.49**	49.03	82.54	159.57
	CoDeRNet (Ours)	**90.25**	87.44	82.38	77.85	74.17	**49.04**	**83.58**	**163.37**

Here, t_{glob}, T_{reg}, and T_{loc} denote global, region-level, and local tokens that capture overall context, intermediate semantic units, and fine spatial evidence, respectively. The Transformer decoder $\mathcal{D}_{\text{cap}}$ predicts the next-token distribution using the caption prefix and the multi-granularity memory.

3.6 Training Objective

CoDeRNet is trained in a multi-task setting for change detection and change captioning. Following prior CDC practice [25], we use task-specific losses for the two branches and optimize them jointly:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{CE}}(\hat{M}, M^*) + \mathcal{L}_{\text{Dice}}(\hat{M}, M^*), \quad \mathcal{L}_{\text{cap}} = -\sum_{t=1}^{T} \log p(y_t^* \mid y_{<t}^*, \mathcal{M}) .$$

Here, M^* denotes the ground-truth change mask. The segmentation branch is supervised by pixel-wise cross-entropy and Dice loss, while the captioning branch is trained with token-level cross-entropy under teacher forcing over valid non-padded positions.

For stable joint optimization, we normalize each loss by its detached current magnitude:

$$\mathcal{L}_{\text{total}} = \frac{\mathcal{L}_{\text{seg}}}{\text{sg}(\mathcal{L}_{\text{seg}}) + \epsilon} + \frac{\mathcal{L}_{\text{cap}}}{\text{sg}(\mathcal{L}_{\text{cap}}) + \epsilon},$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator and ϵ is a small constant for numerical stability. This normalization is used only as a training stabilization strategy to balance the scales of the two task losses.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate CoDeRNet on two representative CDC benchmarks, LEVIR-MCI [25] and WHU-CDC [35]. LEVIR-MCI, extended from LEVIR-CC [22], contains 10,077 bi-temporal image pairs of size 256×256 at 0.5 m resolution, each annotated with a multi-class change mask for buildings and roads and five human-written captions. Following the standard split, it is divided into 6,815 training, 1,333 validation, and 1,929 test samples.

WHU-CDC, derived mainly from WHU-CD [15], contains 7,434 bi-temporal image pairs of size 256×256 at 0.075 m resolution, focusing on post-earthquake urban reconstruction in Christchurch, New Zealand. Each pair has a binary building-change mask and five captions describing multiple



Figure 2: Qualitative comparison on LEVIR-MCI. In the captions, building-related words (e.g., *house*, *building*) and road-related words (e.g., *road*) are highlighted using colors consistent with the corresponding mask classes.

change categories, including buildings, parking lots, roads, vegetation, and water. The training, validation, and test sets contain 5,947, 743, and 744 samples, respectively. These datasets allow us to evaluate CoDeRNet under different levels of mask granularity and caption semantics.

Both datasets can contain multiple change events per image pair and cover multiple change categories in their annotations, although WHU-CDC masks are restricted to building changes.

Evaluation Metrics. For change detection, we follow prior CDC benchmarks and report mIoU on LEVIR-MCI [25], which provides multi-class change masks for buildings, roads, and no-change regions. On WHU-CDC [35], which provides binary building-change masks, we report the IoU of the change class. For change captioning, we adopt standard language generation metrics, including BLEU-N (N = 1,2,3,4) [31], METEOR [2], ROUGE-L [20], and CIDEr-D [39], covering n-gram overlap, semantic matching, sequence similarity, and consensus-based caption quality.

4.2 Implementation Details

All experiments are implemented in PyTorch and conducted on a single NVIDIA RTX 4090 GPU. We use a ConvNeXt-Base backbone [27] with ImageNet-1K pretrained weights and apply it to the two temporal images in a weight-shared manner. The maximum caption length is set to 41 for LEVIR-MCI [25] and 26 for WHU-CDC [35], and greedy autoregressive decoding is used during inference. For training, we use Adam [17] with a learning rate of 1×10^{-4} and an effective batch size of 64, achieved through a micro-batch size of 16 and four gradient accumulation steps. The learning rate is decayed by a factor of 0.5 every 50 epochs, and the model is trained for 100 epochs on LEVIR-MCI and 150 epochs on WHU-CDC. We apply random horizontal and vertical flips together with 90° rotations during training. When geometric augmentation is applied, direction-sensitive caption words are remapped accordingly. Detailed remapping rules are provided in Sec. 7.3. For both LEVIR-MCI and WHU-CDC, the input size is fixed to 256×256. For this setup, a single training run takes approximately 12 hours.

4.3 Comparison with State-of-the-Art Methods

We compare CoDeRNet with recent state-of-the-art methods on LEVIR-MCI [25] and WHU-CDC [35], including single-task change detection models, single-task change captioning models, and unified change detection-captioning models. This comparison evaluates whether a unified framework can improve spatial localization and semantic description of changes at the same time. As shown in

Table 3: Ablation studies on LEVIR-MCI [25]. (a) Effect of joint training and explicit interaction, comparing single-task training, joint training without CoDeR, and joint training with CoDeR. (b) Effect of CoDeR directionality, comparing two single-direction settings with the bidirectional design. **Bold** indicates the best result in each column.

(a) Joint training and CoDeR

Seg	Cap	CoDeR	mIoU	BLEU-4	METEOR	ROUGE-L	CIDEr-D
✓	-	-	87.03	-	-	-	-
-	✓	-	-	63.18	40.21	75.27	137.41
✓	✓	-	87.67	64.41	39.93	74.98	136.18
✓	✓	✓	**87.79**	**66.27**	**41.08**	**76.36**	**141.79**

(b) CoDeR directionality

Direction	mIoU	BLEU-4	METEOR	ROUGE-L	CIDEr-D
cap → seg	87.53	64.79	40.27	75.25	138.62
seg → cap	87.60	64.96	40.95	76.33	**141.95**
seg ↔ cap	**87.79**	**66.27**	**41.08**	**76.36**	141.79

Table 1, CoDeRNet achieves the best overall performance on LEVIR-MCI across both detection and captioning metrics. For example, CoDeRNet achieves BLEU-4 66.27, CIDEr-D 141.79, and mIoU 87.79 on LEVIR-MCI, showing stronger overall performance than the strongest prior unified model.

Table 2 shows a similar trend on WHU-CDC. CoDeRNet achieves the best IoU of 90.25 and improves consensus-oriented captioning metrics, obtaining the best CIDEr-D of 163.37 and ROUGE-L of 83.58, while remaining competitive on BLEU-based metrics.

Compared with prior unified CDC frameworks, CoDeRNet preserves strong localization accuracy while improving semantic description quality across both datasets, leading to a better overall balance between the two tasks. Overall, these results show that CoDeRNet effectively supports both spatial and semantic change understanding within a unified framework, while remaining robust across datasets with different change characteristics and annotation settings.

Qualitative comparisons in Fig. 2 further show that the visual improvements are consistent with the quantitative results on LEVIR-MCI. Additional qualitative examples, including further results on LEVIR-MCI (Fig. 4) and WHU-CDC (Fig. 5), are provided in Appendix.

4.4 Ablation Study

Effect of Joint Training and CoDeR. We first examine whether selective routing provides additional benefits beyond simple joint optimization, as shown in Table 3(a). Compared with single-task training and joint training without routing, the joint model with CoDeR achieves the best overall performance across both detection and captioning metrics. This result suggests that joint learning alone is helpful but not sufficient, and that explicit cross-task routing further improves unified change understanding.

Effect of Routing Directionality. To examine how CoDeR should be applied between the two branches, we compare two single-direction settings with the bidirectional design. As shown in Table 3(b), the bidirectional design achieves the best overall balance across metrics among the three compared settings. This suggests that, once cross-task routing is introduced, mutual information exchange between the two branches is more effective than one-way transfer alone.

Effect of Routing Mask Decomposition. To validate the core design of CoDeR, we ablate the decomposed routing mask and compare it with simplified variants. As shown in Table 4(a), the full formulation consistently performs best, while removing either the target-dependent or source-dependent term degrades performance. Replacing the decomposition with a single gate also results in weaker performance. These results show that the benefit of CoDeR does not arise merely from gating itself, but from explicitly decomposing the routing condition into receiver-side need for support and sender-side reliability cues.

Effect of Interaction Operators. To verify that the gains of CoDeRNet come from the proposed routing design rather than simply adding cross-task interaction, we compare CoDeR with representative interaction operators in Table 4(b), including co-attention [28, 29], cross-attention [41, 44], channel exchange [40], and gated modulation. Since these methods were originally designed for different tasks and architectures, we implement their core interaction principles within the same CoDeRNet pipeline while keeping all other components fixed. Detailed formulations are provided in Appendix 7.4. As shown in the table, generic interaction operators can improve cross-branch communication, but they do not explicitly determine whether the receiving branch needs support or whether the sending branch provides reliable evidence. In contrast, CoDeR structures cross-task communication as conditional routing under heterogeneous supervision, leading to stronger captioning performance while preserving competitive detection accuracy. These results show that the

Table 4: Ablation studies on LEVIR-MCI [25] for analyzing the effectiveness of CoDeR. (a) Routing mask decomposition ablation compares CoDeR with reduced variants and a single-gate design. (b) Interaction operator ablation compares CoDeR with representative cross-task communication operators under the same framework. **Bold** and underline indicate the best and second-best results in each column, respectively.

(a) CoDeR mask decomposition

Mask	mIoU	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Sender reliability (C_s)	87.60	65.12	40.49	75.73	139.57
Receiver need ($1 - C_t$)	87.65	64.65	40.81	75.94	141.02
Single gate	87.59	65.46	40.27	75.54	138.19
Decomposed routing ($(1 - C_t) \cdot C_s$)	**87.79**	**66.27**	**41.08**	**76.36**	**141.79**

(b) Interaction operators

Interaction	mIoU	BLEU-4	METEOR	ROUGE-L	CIDEr-D	Params (M)	FLOPs (G)
Co-Attn	**87.84**	64.15	40.64	75.44	139.33	102.8	67.46
Cross-Attn	87.70	65.01	40.60	75.53	138.87	103.2	67.71
Exchange	87.53	65.15	40.61	75.47	139.12	102.9	67.46
Gating	87.75	<u>65.66</u>	<u>40.73</u>	<u>76.09</u>	<u>140.86</u>	103.1	67.46
CoDeR (Ours)	<u>87.79</u>	**66.27**	**41.08**	**76.36**	**141.79**	102.8	67.46

effectiveness of cross-task communication in CDC depends not only on exchanging information, but also on how the transfer condition is defined.

Further analyses on CoDeR and other design choices are provided in Appendix Secs. 7.1–7.3, where we examine the effect of routing scale, backbone, and additional architectural and training factors.

4.5 Human Evaluation of Caption Grounding and Cross-Output Consistency

Standard captioning metrics measure agreement with reference captions but do not directly establish whether generated descriptions faithfully reflect observable changes or agree with the model's spatial predictions. We therefore conduct a blind human evaluation using two complementary criteria: factual grounding, which assesses the correctness of a caption with respect to the annotated changes, and cross-output consistency, which assesses agreement between a caption and the change mask predicted by the same model.

We evaluate MCINet, our baseline without CoDeR, and CoDeRNet on 660 change-containing image pairs from the LEVIR-MCI test set. The samples are stratified by change composition (building only, road only, or both) and change area. Seven researchers participated in the blind evaluation, with model identities concealed. Each evaluator assesses all three systems on the same assigned set of 120 image pairs in two separate rounds. In the factual-grounding round, captions are evaluated against the ground-truth change masks, without access to predicted masks. In the cross-output-consistency round, captions are evaluated against the corresponding predicted change masks, while ground-truth masks are withheld. Both criteria are rated on a five-point scale, with higher scores indicating closer agreement between the caption and the mask provided in the corresponding evaluation round. Both criteria are restricted to building- and road-related changes covered by the annotations; mentions of categories outside this scope are treated as neutral. Thirty image pairs are shared among all seven evaluators to assess inter-rater agreement. Each round yields 840 ratings per system and 2,520 ratings across the three systems, resulting in 5,040 criterion scores across both rounds.

Table 5 reports the mean and standard deviation of the ratings. CoDeRNet achieves the highest mean scores on both criteria, improving factual grounding and cross-output consistency over MCINet by 0.138 and 0.181, respectively. Both improvements are statistically significant under the sign test, with $p = 0.045$ for factual grounding and $p = 0.0004$ for cross-output consistency. CoDeRNet also achieves slightly higher mean scores than the baseline without CoDeR on both criteria. On the shared evaluation subset, factual-grounding ratings yield Krippendorff's $\alpha = 0.739$, while the reliability of the average ratings across the seven evaluators, measured by $\mathrm{ICC}(A, 7)$, reaches 0.974 for factual grounding and 0.892 for cross-output consistency. Together, these results provide evidence beyond reference-based captioning metrics that CoDeRNet improves factual grounding and cross-output consistency over MCINet.

5 Discussion

Our findings suggest that CDC highlights a broader challenge in multi-task learning under heterogeneous supervision, where dense prediction and language generation require different forms of representation. Rather than assuming uniform feature sharing, CoDeRNet shows that conditional information routing can provide a more suitable way to coordinate such heterogeneous task branches.

Limitation. Although CoDeRNet improves both tasks within a unified CDC framework, the performance gains are asymmetric. While captioning benefits significantly from detection cues, the reverse

Table 5: Blind human evaluation on the change-containing subset of LEVIR-MCI. Values are mean $\pm$ standard deviation on a five-point scale, with 840 ratings per system for each criterion across 660 unique image pairs. Higher scores indicate better factual grounding or cross-output consistency. **Bold** indicates the highest mean in each column.

Model	Factual grounding	Cross-output consistency
MCINet [25]	3.520 ± 1.38	3.632 ± 1.25
Ours (w/o CoDeR)	3.645 ± 1.22	3.782 ± 1.09
Ours (w/ CoDeR)	$\mathbf{3.658} \pm 1.27$	$\mathbf{3.813} \pm 1.13$

impact is more constrained by the semantic gap. This suggests that spatially grounded cues from the detection branch are particularly effective for improving semantic change description, whereas the reverse direction provides a more moderate benefit. In addition, our evaluation is restricted to two bi-temporal remote-sensing CDC benchmarks, and the effectiveness of CoDeRNet in multitemporal scenarios and broader dense-prediction-and-language settings remains to be verified. We also do not explicitly supervise the routing cues with region-language correspondence, which may limit semantic faithfulness in cases where visual changes are spatially subtle or linguistically ambiguous.

Broader Impact. CoDeRNet may support applications such as urban monitoring, disaster assessment, and environmental analysis. However, incorrect change localization or captioning could mislead downstream decisions, so human verification is recommended in high-stakes applications.

Future Work. A promising direction for future work is to strengthen semantic correspondence between the two branches more explicitly. For example, region-language alignment or class-aware semantic transfer may help cross-task routing reflect not only spatial relevance but also the semantic structure shared by the two tasks. Such designs may provide a more effective way to translate semantic cues into region-level predictions within CDC.

6 Conclusion

In this paper, we revisit change detection and captioning not simply as a conventional joint learning problem, but as a conditional information routing problem under heterogeneous supervision. We proposed CoDeRNet, a unified framework that preserves task-specific representations while enabling selective bidirectional exchange of complementary information. At its core, Confidence-Decomposed Routing decomposes the routing condition into receiver-side need for support and sender-side reliability cues, allowing cross-task information to be transferred only when it is expected to be beneficial. Experiments on LEVIR-MCI and WHU-CDC show that CoDeRNet improves semantic description quality while preserving strong localization accuracy. Through systematic analyses, we further show that the gains arise not merely from generic feature interaction, but from explicitly modeling selective cross-task communication. The key contribution of CoDeR is therefore not the introduction of another interaction operator, but the reformulation of CDC communication as a conditional routing problem under heterogeneous supervision. This emphasizes the need to model task-specific reliability and support requirements when coordinating branches with different supervision signals. We believe this perspective extends beyond remote sensing CDC to broader multi-task and multimodal learning settings involving heterogeneous supervision and task coordination.

Acknowledgments

This work was supported in part by InSpace Company Ltd. (Inspace-Kyung Hee Artificial Intelligence Research Institute), by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00334321), by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (No. IITP-2026-RS-2023-00258649) supervised by the IITP (Institute for Information and Communications Technology Planning and Evaluation), and by Global - Learning & Academic research institution for Master's·PhD students, and Postdocs (G-LAMP) Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education(No. RS-2025-25442355). The authors declare no competing interests.

References

[1] Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210. IEEE, 2022.

[2] Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, 2005.

[3] Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, et al. Tipsv2: Advancing vision-language pretraining with enhanced patch-text alignment. *arXiv preprint arXiv:2604.12012*, 2026.

[4] 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.

[5] Hao Chen, Wenyuan Li, and Zhenwei Shi. Adversarial instance augmentation for building change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021.

[6] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

[7] Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. Changemamba: Remote sensing change detection with spatiotemporal state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024.

[8] Andrea Codegoni, Gabriele Lombardi, and Alessandro Ferrari. Tinycd: A (not so) deep learning model for change detection. *Neural Computing and Applications*, 35(11):8471–8486, 2023.

[9] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 4063–4067. IEEE, 2018.

[10] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *European conference on computer vision*, pages 74–92. Springer, 2022.

[11] Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li. Snunet-cd: A densely connected siamese network for change detection of vhr images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[12] Sheng Fang, Kaiyu Li, and Zhe Li. Changer: Feature interaction is what you need for change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–11, 2023.

[13] Chengxi Han, Chen Wu, Haonan Guo, Meiqi Hu, and Hongruixuan Chen. Hanet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:3867–3878, 2023.

[14] Xin Huang, Liangpei Zhang, and Tingting Zhu. Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(1):105–115, 2013.

[15] 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.

[16] Robert E Kennedy, Philip A Townsend, John E Gross, Warren B Cohen, Paul Bolstad, YQ Wang, and Phyllis Adams. Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects. *Remote sensing of environment*, 113 (7):1382–1396, 2009.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Zhenglai Li, Chang Tang, Lizhe Wang, and Albert Y Zomaya. Remote sensing change detection via temporal feature interaction and guided refinement. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022.

[19] Zhenglai Li, Chang Tang, Xinwang Liu, Wei Zhang, Jie Dou, Lizhe Wang, and Albert Y Zomaya. Lightweight remote sensing change detection with progressive feature aggregation and supervised attention. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023.

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

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

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

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

[24] 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.

[25] 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, 2024.

[26] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.

[27] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[28] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016.

[29] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6087–6096, 2018.

[30] Mubashir Noman, Mustansar Fiaz, Hisham Cholakkal, Sanath Narayan, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Remote sensing change detection with transformers trained from scratch. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

[31] 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*, pages 311–318, 2002.

[32] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4624–4633, 2019.

[33] Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. Describing and localizing multiple changes with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1971–1980, 2021.

[34] Blaž Rolih, Matic Fučka, Filip Wolf, and Luka Čehovin Zajc. Be the change you want to see: Revisiting remote sensing change detection practices. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[35] 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.

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

[37] 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*, 2025.

[38] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[39] 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*, pages 4566–4575, 2015.

[40] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *Advances in neural information processing systems*, 33:4835–4845, 2020.

[41] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020.

[42] 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.

[43] Joseph Z Xu, Wenhan Lu, Zebo Li, Pranav Khaitan, and Valeriya Zaytseva. Building damage detection in satellite imagery using convolutional neural networks. *arXiv preprint arXiv:1910.06444*, 2019.

[44] Yuechen Yu, Yilei Xiong, Weilin Huang, and Matthew R Scott. Deformable siamese attention networks for visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6728–6737, 2020.

[45] Haotian Zhang, Hao Chen, Chenyao Zhou, Keyan Chen, Chenyang Liu, Zhengxia Zou, and Zhenwei Shi. Bifa: Remote sensing image change detection with bitemporal feature alignment. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024.

[46] Haotian Zhang, Keyan Chen, Chenyang Liu, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Cdmamba: Incorporating local clues into mamba for remote sensing image binary change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[47] Zhuo Zheng, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15193–15202, 2021.

[48] Zhuo Zheng, Yanfei Zhong, Ailong Ma, and Liangpei Zhang. Single-temporal supervised learning for universal remote sensing change detection. *International Journal of Computer Vision*, 132(12):5582–5602, 2024.

[49] 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.

7 Appendix

7.1 Scale Selection for CoDeR

We further examine how the receiving scale affects CoDeR. We first vary the routing scale in each direction while fixing the opposite direction to its default setting, as shown in Table 6. We then compare the default asymmetric single-scale design with multi-scale CoDeR variants in Table 8(a) to test whether broader cross-scale routing provides additional benefit.

Table 6(a) shows that, for *cap* $\rightarrow$ *seg*, applying CoDeR at $1/8$ yields the strongest overall performance, suggesting that the detection branch benefits most when caption-derived information is injected at a relatively higher-resolution stage. Table 6(b) shows that, for *seg* $\rightarrow$ *cap*, the best overall performance is obtained at $1/16$, indicating that the captioning branch benefits more from interaction at a relatively lower-resolution stage aligned with semantic abstraction.

Table 8(a) further shows that the default asymmetric single-scale design remains the most consistent configuration overall. The fully multi-scale variant degrades all caption metrics, while the direction-specific multi-scale variants do not establish a clear advantage over the default. These results indicate that the effectiveness of CoDeR comes not from broader cross-scale transfer, but from selecting an appropriate routing scale for each direction.

Overall, the results support the asymmetric scale design of CoDeRNet: *cap* $\rightarrow$ *seg* is most effective at a higher-resolution scale, whereas *seg* $\rightarrow$ *cap* is most effective at a lower-resolution scale. More importantly, CoDeR should be applied selectively at an appropriate granularity rather than uniformly across all scales.

7.2 Backbone-Level Generalizability of CoDeR

We further examine whether the effect of CoDeR is consistent across different encoder choices, rather than being tied to a specific backbone architecture. To this end, Table 7 compares the baseline without CoDeR and CoDeRNet under the same backbone setting, across diverse backbone families with image-only pretraining (e.g., ConvNeXt [27]) or vision-language pretraining (e.g., TIPSv2 [3]), as well as different model sizes within the ConvNeXt series.

As shown in Table 7(a), enabling CoDeR generally improves captioning performance across all tested backbone families, with consistent gains in BLEU-4, METEOR, ROUGE-L, and CIDEr-D in most cases. Although the effect on mIoU is smaller and not always positive, the overall trend indicates that the proposed routing mechanism is not limited to a particular encoder design. Table 7(b) shows a similar pattern across different ConvNeXt scales. From Tiny to Large, CoDeR consistently improves most captioning metrics, while detection performance is generally maintained and occasionally shows modest improvements. These results suggest that the benefit of CoDeR is broadly preserved across backbone families and model scales, supporting the general applicability of the proposed routing design.

7.3 Additional Ablation Studies

Effect of region-token pooling size. Table 8(b) studies the pooling size used to construct region-level tokens in the caption branch. Compared with using no region-level tokens, introducing pooled region tokens generally improves both detection and captioning performance, confirming the usefulness of intermediate-granularity representations. Among the tested settings, a pooling size of 2 gives the best overall performance on most metrics. Although a pooling size of 4 remains competitive, it does not surpass the overall performance of size 2, suggesting that an appropriate region granularity is more effective than overly coarse pooling.

Effect of augmentation type. Table 9(a) analyzes the effect of geometric augmentation and directional caption remapping. Applying geometric augmentation alone improves several captioning metrics, but the gains are not consistent across all metrics. In contrast, combining geometric augmentation with caption remapping yields the best overall results for both detection and captioning. This indicates that, when direction-sensitive transformations are applied to the input images, maintaining caption-image consistency through remapping (see Table 10) is important for effective training.

Table 6: Ablation of the CoDeR routing scale on LEVIR-MCI [25]. Each subtable reports performance when same-scale routing is applied at different receiving scales in one direction, while the opposite direction is fixed to its default setting. **Bold** indicates the best result in each column.

(a) **Cap → seg: routing scale**

Scale	mIoU	BLEU-4	METEOR	ROUGE-L	CIDEr-D
1/4	87.75	64.93	40.62	75.61	139.84
1/8	**87.79**	**66.27**	**41.08**	**76.36**	**141.79**
1/16	87.48	**66.27**	40.49	76.17	139.86
1/32	87.78	65.72	40.71	75.93	139.80

(b) **Seg → cap: routing scale**

Scale	mIoU	BLEU-4	METEOR	ROUGE-L	CIDEr-D
1/8	87.76	65.54	40.38	75.22	138.04
1/16	**87.79**	**66.27**	**41.08**	**76.36**	**141.79**
1/32	87.59	66.13	40.23	75.50	138.45

Table 7: Robustness of CoDeRNet across backbone choices on LEVIR-MCI [25]. (a) Backbone-family ablation: performance changes with and without CoDeR across different backbones. (b) Backbone-size ablation: performance changes with and without CoDeR across ConvNeXt variants from Tiny to Large. **Bold** indicates the best result in each column. In the CoDeRNet rows, colored values indicate improvements (gain) or degradations (drop) relative to the corresponding baseline. B4, MET., R-L, and CID. denote BLEU-4, METEOR, ROUGE-L, and CIDEr-D, respectively.

(a) Backbone family ablation

Backbone	CoDeR	mIoU	BLEU-4	MET.	R-L	CID.
ConvNeXt-B [27]	×	87.67	64.41	39.93	74.98	136.18
	✓	**87.79**	**66.27**	**41.08**	**76.36**	**141.79**
SwinV2-B [26]	×	87.49	63.55	40.24	75.26	136.93
	✓	87.34	64.45	40.30	75.31	138.20
MiT-B5 [42]	×	87.74	63.92	40.15	75.49	137.54
	✓	87.78	65.38	40.54	75.54	140.28
DaViT-B [10]	×	87.27	63.32	39.70	74.94	136.85
	✓	87.37	64.59	40.34	75.24	137.03
EfficientNetV2-M [38]	×	87.05	63.31	40.31	75.66	137.97
	✓	87.10	65.17	40.72	75.99	140.12
TIPSv2-B/14 [3]	×	86.56	64.31	40.27	75.30	138.00
	✓	86.73	65.23	40.91	75.74	140.86

(b) Backbone size ablation

Size	CoDeR	mIoU	BLEU-4	MET.	R-L	CID.
Tiny	×	87.05	65.26	39.91	75.47	134.69
	✓	87.48	65.76	41.02	76.13	141.50
Small	×	87.22	64.29	40.18	75.28	137.32
	✓	87.32	65.30	40.05	75.33	138.40
Base	×	87.67	64.41	39.93	74.98	136.18
	✓	87.79	**66.27**	**41.08**	**76.36**	**141.79**
Large	×	87.81	64.91	40.02	75.36	138.93
	✓	**87.85**	65.73	40.65	76.01	140.23

Effect of change fusion composition. Table 9(b) examines how the composition of the shared change representation affects performance. Starting from the temporal features $\{t_0, t_1\}$, adding the signed change cue Δ already improves the captioning-related metrics, indicating that directional temporal difference provides useful semantic information. Further adding the magnitude cue $|\Delta|$ leads to the best overall performance across all reported metrics. These results suggest that the signed and magnitude change cues play complementary roles, and that combining both is more effective than using temporal features alone.

7.4 Detailed Formulation of Feature Interaction Operators

This section provides the detailed formulations of the feature interaction operators compared in Table 4(b), which serve as controlled baselines for cross-branch communication. The purpose of this comparison is not to faithfully reproduce the original architectures of prior methods, since they were introduced for different tasks, modalities, feature resolutions, and network designs. Instead, each operator is implemented as an abstraction of its core interaction principle within the same CoDeRNet framework. All variants are inserted at the same cross-task transfer locations and use the same encoder, task decoders, feature dimensions, and training protocol. This controlled setting allows the comparison to focus on the effect of the feature interaction design itself rather than on differences in the surrounding architecture.

Let $F_t, F_s \in \mathbb{R}^{B\times C\times H\times W}$ denote the target and source features, respectively. The target feature F_t indicates the branch representation being updated, while the source feature F_s provides complementary information from the other branch. For example, when the segmentation branch receives information from the captioning branch, F_t corresponds to the segmentation feature and F_s corresponds to the captioning feature. The reverse direction is defined analogously. We use $\phi(\cdot)$ to denote a lightweight source adapter implemented with a 1×1 convolution, and $\odot$ denotes element-wise multiplication. All operators are implemented as lightweight feature interaction blocks with comparable parameter counts and FLOPs.

Table 8: Ablation studies on LEVIR-MCI [25]. (a) Single-scale versus multi-scale CoDeR application. The default design applies CoDeR at one receiving scale per direction (seg $\leftarrow$ cap at $1/8$, cap $\leftarrow$ seg at $1/16$), while the multi-scale variants apply independent CoDeR blocks at all four scales for the indicated direction(s). (b) Region-token pooling size ablation for caption decoding. **Bold** indicates the best result in each column.

(a) Single-scale vs. multi-scale

Type	Setting	mIoU	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Multi	both (all $\leftrightarrow$ all)	87.66	65.81	40.18	75.93	138.83
	seg only (all $\leftrightarrow$ $1/16$)	87.75	65.44	41.03	76.22	**142.18**
	cap only ($1/8 \leftrightarrow$ all)	**87.86**	64.56	40.83	75.93	140.32
Single (Ours)	($1/8 \leftrightarrow 1/16$)	87.79	**66.27**	**41.08**	**76.36**	141.79

(b) Region pooling size

M	mIoU	BLEU-4	METEOR	ROUGE-L	CIDEr-D
None	87.56	65.95	40.80	75.99	139.22
1	87.73	66.27	40.55	75.58	138.88
2	**87.79**	**66.27**	**41.08**	**76.36**	**141.79**
4	87.77	65.67	**41.08**	76.22	141.57

Table 9: Additional ablation studies on LEVIR-MCI [25]. **(a)** Effect of augmentation type, where G denotes geometric augmentation and R denotes directional caption remapping. **(b)** Effect of change fusion composition, where F is formed from temporal features at t_0 and t_1 together with signed and magnitude change cues Δ and $|\Delta|$. **Bold** indicates the best result in each column.

(a) Augmentation type

Aug.	mIoU	BLEU-4	METEOR	ROUGE-L	CIDEr-D
-	86.61	64.51	39.44	74.72	136.06
G	86.67	64.45	40.64	75.97	139.21
G+R	**87.79**	**66.27**	**41.08**	**76.36**	**141.79**

(b) Change fusion composition

F	mIoU	BLEU-4	METEOR	ROUGE-L	CIDEr-D
$\{t_0, t_1\}$	87.45	64.21	40.04	75.37	138.76
$\{t_0, t_1, \Delta\}$	87.47	65.69	40.53	75.87	140.86
$\{t_0, t_1, \Delta, \lvert\Delta\rvert\}$	**87.79**	**66.27**	**41.08**	**76.36**	**141.79**

Co-attention-style Interaction. The co-attention-style variant computes a spatial modulation map from the concatenated target and source features:

$$a = \sigma\left(\mathrm{Conv}_{1\times1}([F_t; F_s])\right),$$

where $[F_t; F_s]$ denotes channel-wise concatenation, $\sigma(\cdot)$ is the sigmoid function, and $a \in [0,1]^{B\times1\times H\times W}$ is a spatial attention map. The target feature is updated by

$$F_t' = F_t + a \odot \phi(F_s).$$

This operator represents a co-attention-style spatial gating mechanism, where the source feature is injected into the target branch according to a spatial weight computed from the joint target-source representation. Since the target and source features are mixed before computing a, the resulting modulation does not explicitly distinguish receiver-side need for support from sender-side reliability.

Cross-attention Interaction. The cross-attention variant uses the target feature as query and the source feature as key and value. After flattening the spatial dimensions, we define

$$Q = \mathrm{Flatten}(F_t), \quad K = \mathrm{Flatten}(F_s), \quad V = \mathrm{Flatten}(F_s),$$

where $Q, K, V \in \mathbb{R}^{B\times HW\times C}$. The target feature is updated as

$$F_t' = F_t + \mathrm{Reshape}\left(\mathrm{MHA}(Q, K, V)\right),$$

where $\mathrm{MHA}(\cdot)$ denotes multi-head attention and $\mathrm{Reshape}(\cdot)$ restores the spatial feature shape. This operator captures token-wise target-source relations through attention weights. For a controlled comparison with the other lightweight operators, we use the attention operation as the interaction module without adding a full Transformer block with additional feed-forward layers.

Exchange-style Interaction. The exchange-style variant performs a soft channel-wise interpolation between the target feature and the adapted source feature. It first computes a channel-wise exchange gate:

$$p = \sigma\left(\mathrm{Conv}_{1\times1}(\mathrm{GAP}(F_t))\right),$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling and $p \in [0,1]^{B\times C\times1\times1}$. The target feature is updated by

$$F_t' = (1-p) \odot F_t + p \odot \phi(F_s).$$

This formulation can be interpreted as an exchange-inspired soft mixing operation: channels with small values of p mainly preserve the target representation, while channels with large values of p receive more information from the source branch. Unlike spatial gating, this operator controls feature transfer at the channel level and uses the same exchange weight across spatial locations. Unlike hard channel replacement strategies, this formulation implements a soft exchange-style interaction within the same comparison framework.

Table 10: Directional caption remapping under geometric augmentation on LEVIR-MCI. To preserve caption–image consistency, direction-sensitive expressions are remapped according to the applied geometric transform. Two-word directional phrases are matched prior to single-word directional levels.

Transform	Category	Rule / Example
		Remapping rules
H-Flip	Single-level	*left* $\leftrightarrow$ *right*
	Phrase-level	*upper left* $\leftrightarrow$ *upper right*
	Phrase-level	*lower left* $\leftrightarrow$ *lower right*
	Phrase-level	*top left* $\leftrightarrow$ *top right*
	Phrase-level	*bottom left* $\leftrightarrow$ *bottom right*
V-Flip	Single-level	*top* $\leftrightarrow$ *bottom*, *upper* $\leftrightarrow$ *lower*
	Phrase-level	*top left* $\leftrightarrow$ *bottom left*
	Phrase-level	*top right* $\leftrightarrow$ *bottom right*
	Phrase-level	*upper left* $\leftrightarrow$ *lower left*, *upper right* $\leftrightarrow$ *lower right*
90° CCW rot. ($\times k$)	Single-level	*top* $\rightarrow$ *left* $\rightarrow$ *bottom* $\rightarrow$ *right* $\rightarrow$ *top*
	Single-level	*upper* $\rightarrow$ *left*, *lower* $\rightarrow$ *right*
	Phrase-level	*top left* $\rightarrow$ *bottom left* $\rightarrow$ *bottom right* $\rightarrow$ *top right* $\rightarrow \cdots$
	Phrase-level	*upper left* $\rightarrow$ *lower left* $\rightarrow$ *lower right* $\rightarrow$ *upper right* $\rightarrow \cdots$
		Examples
Phrase-level	Original	*a building appeared in the **upper left** corner of the scene*
	H-Flip	*a building appeared in the **upper right** corner of the scene*
	V-Flip	*a building appeared in the **lower left** corner of the scene*
	Rot. 90° CCW	*a building appeared in the **lower left** corner of the scene*
	Rot. 180°	*a building appeared in the **lower right** corner of the scene*
	Rot. 270° CCW	*a building appeared in the **upper right** corner of the scene*
Single-level	Original	*the road on the **left** was removed*
	H-Flip	*the road on the **right** was removed*
	Rot. 90° CCW	*the road on the **bottom** was removed*

Gating-based Modulation. The gated modulation variant computes a channel-wise gate from the global context of both target and source features:

$$z = \sigma\left(\mathrm{Conv}_{1\times1}(\mathrm{GAP}([F_t; F_s]))\right),$$

where $z \in [0,1]^{B\times C\times1\times1}$. The target feature is updated as

$$F_t' = F_t + z \odot \phi(F_s).$$

This operator performs residual source feature injection controlled by a channel-wise gate. Compared with exchange-style interaction, it preserves the original target feature through a residual update rather than interpolating between the target and source features.

Confidence-Decomposed Routing (CoDeR). CoDeR separately computes target-side sufficiency and source-side reliability cues:

$$C_t = \mathcal{R}_t(F_t), \quad C_s = \mathcal{R}_s(F_s),$$

where $\mathcal{R}_t$ and $\mathcal{R}_s$ are lightweight routing estimators implemented with 1×1 convolution heads followed by sigmoid activation. The confidence-decomposed routing mask is defined as

$$M = (1 - C_t) \odot C_s.$$

The target feature is then updated by

$$F_t' = F_t + M \odot \phi(F_s).$$

Here, C_t and C_s are learned routing cues rather than externally supervised confidence labels, estimating target-side sufficiency and source-side reliability, respectively. The routing mask M becomes

Table 11: Additional detection evaluation. (a) Multi-class mIoU and change-class IoU (cIoU) on LEVIR-MCI [25]. Seg and Cap indicate change detection and change captioning training, respectively. cIoU is computed by merging building and road into a single change class. **Bold** indicates the best result in each metric column. (b) Image-level paired bootstrap analysis. Δ denotes the full-test-set performance difference between CoDeRNet and MCINet. Differences and 95% confidence intervals are expressed in percentage points. Positive values favor CoDeRNet.

(a) Change-class IoU evaluation

Seg	Cap	CoDeR	mIoU	cIoU
✓	×	×	87.03	82.05
✓	✓	×	87.67	82.97
✓	✓	✓	**87.79**	**83.07**

(b) Paired bootstrap analysis

Dataset	Metric	Δ	95% CI
LEVIR-MCI	mIoU	$+1.36$	$[+1.07, +1.70]$
	cIoU	$+1.79$	$[+1.41, +2.17]$
WHU-CDC	cIoU	$+2.02$	$[+0.57, +3.69]$

large when the receiving branch requires support and the sending branch provides reliable evidence, encouraging source information to be transferred mainly where it is expected to be complementary. This differs from joint gating mechanisms that compute a single modulation weight from mixed target-source features, as CoDeR explicitly decomposes the transfer condition into receiver-side need and sender-side reliability terms.

7.5 Change-Class IoU Evaluation

To complement the multi-class mIoU evaluation on LEVIR-MCI, we additionally report change-class IoU (cIoU), which measures localization performance over the combined change regions. We merge the building and road classes into a single change class in both the predicted and ground-truth masks, and compute cIoU as the ratio of the total intersection to the total union across the test set. Unlike binary mIoU, which averages the IoUs of the background and change classes, cIoU evaluates only the change class. It therefore assesses change localization independently of the distinction between building and road changes.

Table 11(a) compares single-task training, joint training without CoDeR, and joint training with CoDeR. Joint training improves cIoU from 82.05 to 82.97 over the single-task baseline, and adding CoDeR further increases it to 83.07. The same trend holds for mIoU, with CoDeRNet achieving the highest scores on both metrics among the three settings. These results indicate improvements in both multi-class segmentation and localization of the merged change regions, although the incremental detection gains from CoDeR over joint training are modest: 0.12 percentage points in mIoU and 0.10 in cIoU.

7.6 Statistical Analysis

To assess the statistical evidence for the detection improvements, we conduct an image-level paired bootstrap analysis comparing CoDeRNet with MCINet [25] on LEVIR-MCI and WHU-CDC. We use one fixed checkpoint per method and evaluate both methods on the same test images under the same evaluation protocol. In each bootstrap replicate, we sample image indices with replacement, drawing as many images as in the original test set, and apply the same sampled indices to both methods. We then pool the pixel-level confusion counts over the sampled images, recompute the dataset-level metrics, and calculate the performance differences between the two methods. We repeat this procedure 10,000 times and obtain 95% confidence intervals using the 2.5th and 97.5th percentiles of the bootstrap distribution of performance differences.

Table 11(b) reports the performance differences computed on the full test sets and their bootstrap confidence intervals. Compared with MCINet, CoDeRNet improves mIoU and cIoU on LEVIR-MCI by 1.36 and 1.79 percentage points, respectively, and cIoU on WHU-CDC by 2.02 percentage points. The 95% confidence intervals lie entirely above zero for all detection metrics reported in the table, supporting the observed improvements on both datasets under image-level resampling.

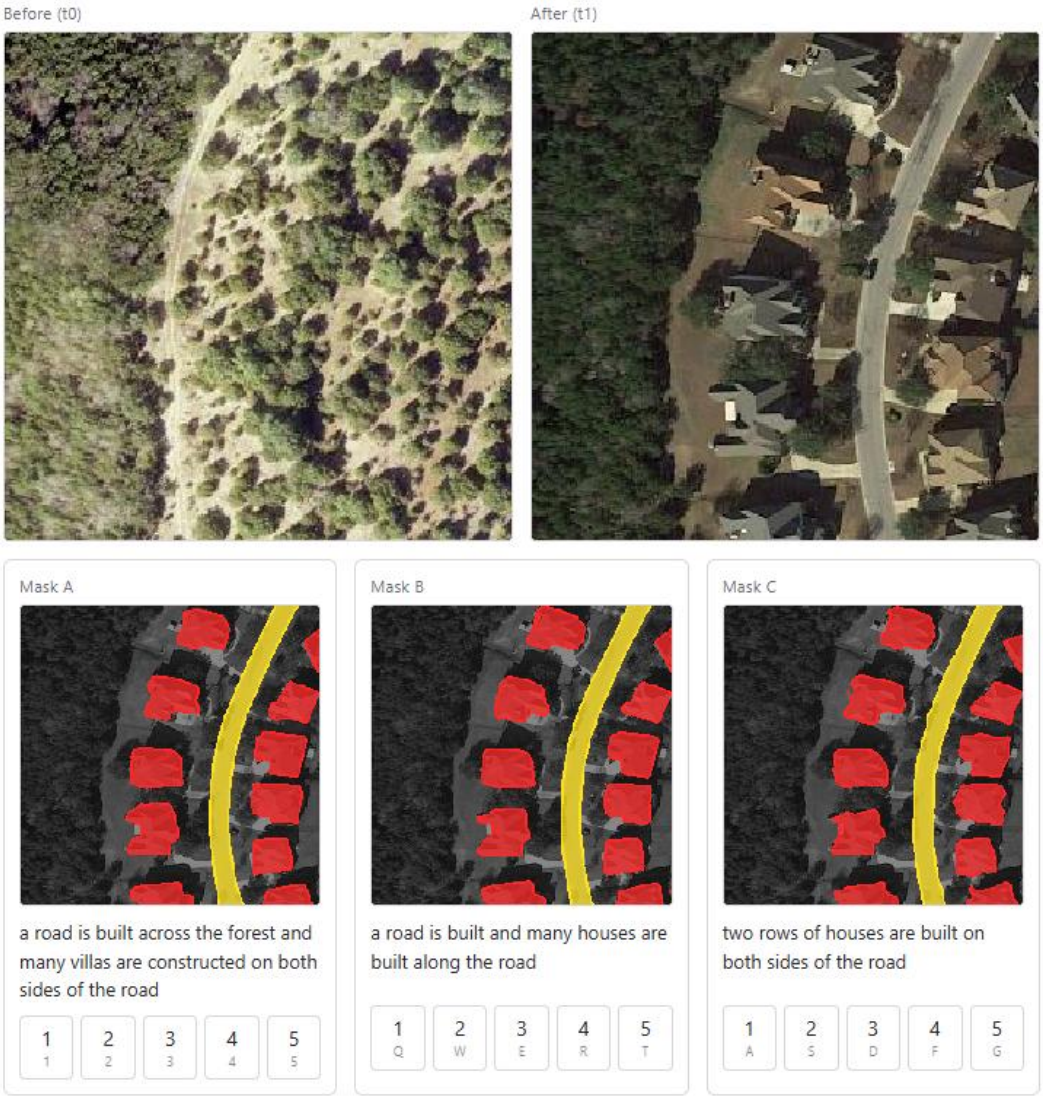


Figure 3: Screenshot of the cross-output-consistency (COC) evaluation interface. Bi-temporal images are displayed alongside three anonymized model outputs (A, B, and C), each consisting of a predicted change mask and a generated caption. Evaluators independently rated each caption on a five-point scale. Red and yellow indicate building and road changes, respectively.

8 Human Evaluation Protocol and Interface

This section provides additional details on the human evaluation protocol, including the evaluation interface, assessment procedure, and scoring criteria. Figure 3 illustrates the web-based interface used for the cross-output-consistency (COC) round of the blind human evaluation. Each evaluator was presented with bi-temporal images and three anonymized model outputs (A, B, and C), each consisting of a predicted change mask and a generated caption. Evaluators independently rated each output on a five-point scale without knowing the corresponding model identity.

The evaluation consisted of two separate rounds. In the factual-grounding (FG) round, captions were assessed against ground-truth change masks, without access to predicted masks. In the COC round, captions were assessed against their corresponding predicted change masks, while ground-truth masks were withheld.

Evaluation Instructions. The following instructions were provided to evaluators in the COC round.

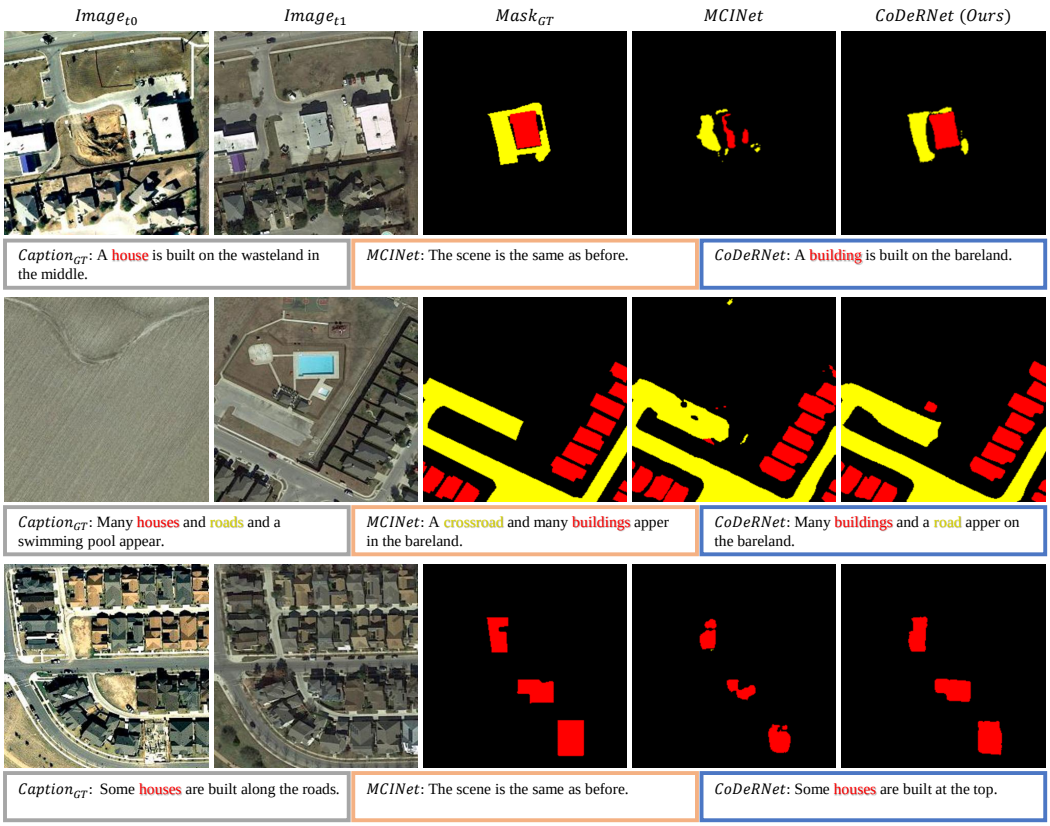


Figure 4: Qualitative comparison on LEVIR-MCI. In the captions, building-related words (e.g., *house*, *building*) and road-related words (e.g., *road*) are highlighted using colors consistent with the corresponding mask classes.

Scoring criteria. Rate how accurately each caption describes the changes marked by the predicted mask shown with it.

5 Accurately describes the building/road changes marked by the mask (object type and location both match).

4 Mostly consistent but partly inaccurate (wrong count, missing location, etc.).

3 Partially consistent (describes only some of the changes, or some object types are wrong).

2 Weak correspondence (fails to describe most of the changes marked by the mask).

1 No correspondence (unrelated to the mask, or says there is no change although the mask shows one).

Scope. Only building and road changes count. If other objects (swimming pools, vegetation, parking lots, etc.) are mentioned, neither add nor deduct points; leave them out of your judgement.

Note. House/building/villa and road/crossroad/street were each treated as equivalent categories. Each caption was evaluated independently, without ranking the three model outputs.

The FG round followed the same five-point scoring framework, with ground-truth masks replacing predicted masks as the reference for evaluating caption accuracy.

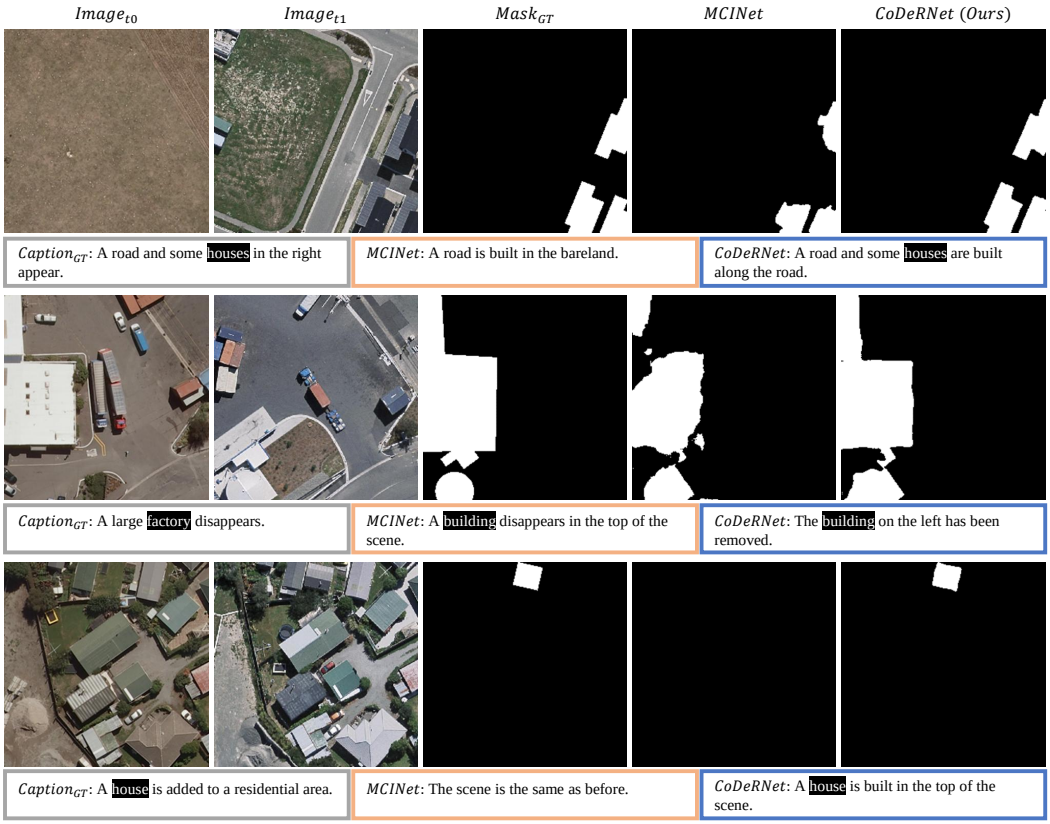


Figure 5: Qualitative comparison on WHU-CDC. In the captions, building-related words (e.g., *house*, *building*) are highlighted using colors consistent with the corresponding mask classes.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of this study are discussed in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not present formal theoretical results, theorems, or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the necessary details for reproducing the main experiments, including datasets, evaluation metrics, model configuration, training settings, and implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The implementation is publicly available at https://github.com/cekkec/CoDeRNet. The experiments are conducted on publicly available benchmark datasets, and the paper describes the data and evaluation protocol used in the experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies the datasets, metrics, training settings, and implementation details.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report sign tests for human evaluation scores in Section 4.5 and image-level paired bootstrap confidence intervals for detection performance in Section 7.6.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the main hardware and approximate training time used for the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public datasets and follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential applications and possible risks of incorrect predictions.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models or datasets requiring special safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the existing datasets, models, and software used in the experiments and follows their intended research-use settings.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The implementation is publicly released with documentation covering installation, data preparation, training, evaluation, and model checkpoints.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Section 4.5 describes the human evaluation protocol, including the evaluation criteria, rating scale, and blind assessment procedure. The Supplementary Material (Section 8) provides the evaluation interface screenshot and detailed scoring instructions given to participants. Seven subjects participated in the evaluation.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: The evaluation involved researchers assessing model-generated captions and change masks. No formal IRB approval or equivalent institutional ethics review was obtained.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core research does not use LLMs as part of the method or evaluation.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.