# INSTRUCTCD: A UNIFIED INSTRUCTION-FOLLOWING FRAMEWORK FOR REMOTE SENSING CHANGE DETECTION

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

## ABSTRACT

Remote-sensing change detection has diverged into four paradigms: binary, semantic, referring, and reasoning, with each addressed by specialized models and increasingly driven by natural-language instructions. Existing unified models cover only part of this spectrum. They merge binary and semantic detection, or couple detection with textual question answering, yet none handle all four paradigms in a single model, and none combine variable-category referring with rejection and multi-level reasoning. We present InstructCD, an instruction-following framework that recasts every paradigm as one interface: a language query produces a set of `<change>` tokens, and each token decodes to a mask, so the number and meaning of the outputs follow the instruction. A Multimodal Language Model conditions the queries, a Temporal Scale Adapter encodes pre-change, post-change, and difference evidence, and a Flexible Mask Decoder emits a variable number of masks. To evaluate the harder paradigms, we build directional semantic queries, a category-referring set, and InstructCD-29K, a reasoning benchmark with four difficulty levels whose supervision is recomputed deterministically from ground truth. Under one protocol in which every baseline uses the same or a larger image encoder, InstructCD achieves the best primary metric on all four paradigms, and a single jointly trained model rivals per-task specialists.

## 1 INTRODUCTION

Change detection compares two co-registered remote-sensing images of the same place acquired at different times and localizes the differences between them. It underlies urban expansion monitoring, post-disaster damage assessment, agricultural and ecological surveys, and land-use enforcement, where the quantity of interest is not the content of a single image but how the scene has changed. Yet different applications care about different changes, and a single fixed change map cannot serve them all. Change detection is therefore shifting toward returning the changes a user specifies in a natural-language instruction.

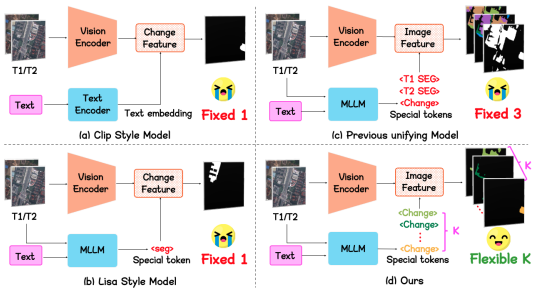



Figure 1: InstructCD follows a language instruction and returns the matching change masks, covering all four paradigms with one set of weights.

Along this shift the task has diverged into four increasingly expressive paradigms. Binary change detection marks every changed pixel with a single foreground class and answers whether and where a scene changed (Chen & Shi, 2020; Ji et al., 2019). Semantic change detection additionally names the land-cover transition on each side of a change, so a pixel is labeled not only as changed but as, for example, vegetation turning into building (Yang et al., 2022). Category-referring change detection narrows the output to the categories a user names, returning only the changes that involve them. Reasoning change detection goes further and resolves an implicit description, such as a function or a cause, so the model must infer which changes are meant rather than being told a category.



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

Because these paradigms developed independently, each comes with its own architectures, datasets, and metrics, and a system that needs several of them must train and maintain several models. Several works have therefore pursued unification, but each covers only part of the spectrum. UniChange merges binary and semantic detection under a Multimodal Language Model, yet uses a fixed query and cannot follow a language-specified target (Zhang et al., 2026). Decoding-the-Delta unifies change detection with textual question answering, so its outputs are sentences rather than the four kinds of masks (Li et al., 2026). Referring and reasoning have been studied in isolation, by RCDNet for single-category referring (Korkmaz et al., 2026) and by ReasonCD for reasoning (Huang et al., 2025). No prior change-detection model places all four paradigms in one network, and none combines variable-category referring with rejection and multi-level reasoning in the same model.

We observe that the four paradigms share a single structure: a language query maps to a set of change masks. InstructCD makes this structure explicit. Every target is represented by a `<change>` token, and the number and semantics of the tokens are dictated by the instruction. Binary detection asks for one token. Semantic detection enumerates thirty directional tokens. Referring detection emits one token per queried category, with an empty mask when a category is absent. Reasoning detection emits one token per described target. A single decoding mechanism serves all of them, so the paradigms differ only in their instructions.

On top of a shared image encoder, InstructCD realizes this interface with three components, shown in Figure 2: a Multimodal Language Model whose `<change>` hidden states become mask queries, a Temporal Scale Adapter that provides pre-change, post-change, and difference evidence, and a Flexible Mask Decoder that predicts one mask per query. To evaluate the harder paradigms, we construct directional semantic queries over SECOND, a category-referring set, and InstructCD-29K, whose supervision is recomputed deterministically from ground truth.

Our contributions are as follows.

- **A unified formulation.** We recast the four paradigms as one mapping from a language query to an ordered, variable-size tuple of change masks, where each `<change>` token indexes one mask.
- **One shared architecture.** Language-conditioned `<change>` queries drive a Flexible Mask Decoder over the pre-change, post-change, and difference evidence from a Temporal Scale Adapter, decoding any number of per-target masks jointly with the same network across all four paradigms.
- **New capabilities and benchmarks.** We extend unified change detection to variable-category referring and multi-level reasoning by building the SECOND-Ref and InstructCD-29K benchmarks. With every baseline retrained on the same or a larger image encoder, InstructCD achieves the best primary metric on all four paradigms, and a single jointly trained model rivals per-task specialists.

## 2 Related Work

### 2.1 Specialized Change Detection

Binary change detection has progressed from fully convolutional Siamese networks (Caye Daudt et al., 2018) to attention-based (Chen & Shi, 2020) and Transformer-based models (Chen et al., 2022; Bandara & Patel, 2022; Fang et al., 2023), with dense (Fang et al., 2022) and hierarchical-attention variants (Han et al., 2023), and recent state-space (Chen et al., 2024a) and foundation-model backbones (Chen et al., 2024b). Semantic change detection adds land-cover transitions and dedicated benchmarks (Yang et al., 2022), and is addressed by state-space (Chen et al., 2024a) and SAM-adapted models (Mei et al., 2024). These methods are accurate within their paradigm but are tied to fixed output heads and cannot transfer across paradigms.

### 2.2 Language-Driven and Unified Change Detection

Multimodal Language Models have been introduced to make change detection controllable. UniChange unifies binary and semantic detection (Zhang et al., 2026), Change-LISA adds language-

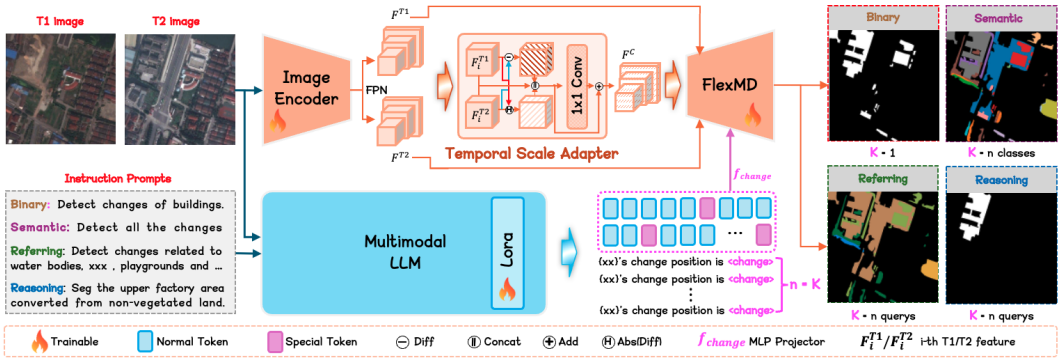



Figure 2: Overview of InstructCD, which consists of a shared image encoder and three components. The encoder and the Temporal Scale Adapter yield pre-change, post-change, and difference evidence, a multimodal LLM emits <change> tokens that a projector turns into mask queries, and FlexMD decodes one mask per query.

guided reasoning on a frozen backbone (Jia et al., 2026), ReasonCD targets implicit change of interest (Huang et al., 2025), and Decoding-the-Delta couples detection with question answering (Li et al., 2026). Change-Agent interprets changes through cooperating modules (Liu et al., 2024), RCDNet performs single-category referring (Korkmaz et al., 2026), and ChangeCLIP aligns visual and textual representations (Dong et al., 2024). Each covers a subset of the paradigms, whereas InstructCD covers all four and combines variable-category referring with rejection and multi-level reasoning in one model.

### 2.3 Reasoning and Referring Segmentation

InstructCD inherits the embedding-as-mask formulation from reasoning segmentation, where a special token from a language model conditions a mask decoder (Lai et al., 2024; Yang et al., 2023). PixelLM decodes masks for several targets from language-model token embeddings (Ren et al., 2024), and GSVA emits multiple segmentation tokens together with a rejection token for absent targets (Xia et al., 2024). In remote sensing, the idea underlies geospatial pixel reasoning (Li et al., 2025) and language-guided segmentation that handles multiple targets (Xin et al., 2026). Our decoder builds on promptable segmentation backbones (Kirillov et al., 2023; Ravi et al., 2025), and our language side builds on instruction-tuned multimodal models (Liu et al., 2023; Bai et al., 2025). These methods segment single images. InstructCD decodes multiple masks from bitemporal change evidence, fixes the slot order so that each slot has a defined meaning such as one of thirty semantic directions, and represents an absent target by an empty mask in its own slot rather than by a separate rejection token.

## 3 Method

### 3.1 Unified Formulation

We formulate every change-detection task as one instruction-following mapping. The inputs are a bitemporal image pair, the pre-change image $I_1$ and the post-change image $I_2$, together with a natural-language query $q$. Given the triplet $(I_1, I_2, q)$, InstructCD returns an ordered tuple of change masks of variable length

$$\mathcal{M} = (M_1, \ldots, M_K), \qquad M_k \in [0,1]^{H \times W}, \tag{1}$$

where $K$ is the number of changes named by $q$ rather than a constant fixed by the architecture. Figure 1 contrasts this variable-size output with prior designs that return a fixed number of masks, and Figure 2 shows the pipeline that realizes it.

A strict index correspondence ties the language side to the mask side. For every $k$, the $k$-th textual target in the response, the $k$-th <change> token, the $k$-th query, and the $k$-th mask $M_k$ all denote the same change, so the order of $\mathcal{M}$ is fixed by the response and no matching step is needed.

Table 1: The unified interface. Every response restates the $k$-th target and closes it with the $k$-th <change>, as in "trees change areas are <change>", so the token order follows the query. A mask is empty when its target is absent, including L0 negatives. Instructions are shortened from the working examples in the technical appendix.

| Paradigm | Test set | Example instruction | $K$ | Mask $M_k$ |
|---|---|---|---|---|
| Binary | WHU-CD, official | detect building-related changes | 1 | all changed pixels |
| Semantic | SECOND, 1,694 | detect changes of the six classes | 30 | the $k$-th from-to direction |
| Referring | SECOND-Ref, 1,694 | detect trees, buildings, and playgrounds | $\lvert\mathcal{C}_q\rvert \le 6$ | changes involving class $k$ |
| Reasoning | InstructCD-29K, 2,930 | fire-damaged buildings on the left | $K_{\text{desc}}$ | the $k$-th described target |

The binary, semantic, category-referring, and reasoning paradigms differ only in how the instruction sets $K$, as

$$K_{\text{bin}} = 1, \qquad K_{\text{sem}} = 30, \qquad K_{\text{ref}} = |\mathcal{C}_q|, \qquad K_{\text{rea}} = K_{\text{desc}}, \tag{2}$$

where the semantic setting enumerates all 30 ordered from-to directions over 6 classes, $\mathcal{C}_q$ is the set of categories the user queries, and $K_{\text{desc}}$ is the number of targets implied by the description. The rest of the model is identical across the four paradigms, and Table 1 lists the instruction, mask meaning, and absent-target behavior of each paradigm. Because the decoder operates on a query set of size $K$, all $K$ masks are decoded jointly from one set of visual features. Union-collapse baselines instead emit one merged mask and lose the per-target identity, and per-class baselines recover it only through $K$ separate inference passes.

### 3.2 Language-Side Query Generation

A multimodal large language model (MLLM) $\Phi$, instantiated as Qwen2.5-VL-7B with LoRA (Bai et al., 2025), reads the bitemporal images $I_1, I_2$ together with the natural-language query $q$. It produces a textual response in which $K$ special <change> tokens are interleaved with ordinary language tokens, written $(y_1, \ldots, y_T) = \Phi(I_1, I_2, q)$ for a response of length $T$, where each <change> token marks one requested change target. The response follows the template of Table 1, which depends only on the query, so the position and number of the tokens are set by the query rather than by the ground truth. In training and evaluation we feed this template response and read all hidden states in one teacher-forced pass. For every per-task InstructCD model in Table 2, free autoregressive generation reproduces this template response and gives the same accuracy as teacher forcing, and for the jointly trained model it lowers only the semantic scores, by at most 0.18 (Table 4). We reuse a token embedding as a segmentation prompt, following Lai et al. (2024).

For the $k$-th <change> token, we take the decoder hidden state at the position that predicts it, denoted $h_k \in \mathbb{R}^{3584}$. A two-layer multi-layer perceptron $f_{change}$ projects it to a change query of width $C = 256$, and stacking the $K$ queries in response order yields the change-query set

$$F_k^q = f_{change}(h_k) \in \mathbb{R}^C, \qquad F^Q = \left[\, F_1^q; \ldots; F_K^q \,\right] \in \mathbb{R}^{K \times C}, \tag{3}$$

which we pass to the Flexible Mask Decoder (FlexMD) shown in Figure 3.

### 3.3 Temporal-Scale Visual Encoding

A shared-weight SAM2.1 Hiera-L encoder (Ravi et al., 2025) reads the pre-change image $I_1$ and the post-change image $I_2$, and a feature pyramid unifies its stages to a common width $C{=}256$. After a high-resolution projection we keep three scales, of side 128, 64, and 32, giving bitemporal features $F^{T1}$ and $F^{T2}$, each a triple of scale maps.

At each scale $s$, the Temporal Scale Adapter forms the temporal residual $\Delta_s = F_s^{T2} - F_s^{T1}$ and builds a change stream $F_s^C$ from the pre-change map, the post-change map, and the signed and absolute residuals, as in

$$F_s^C = \mathrm{ConvRes}\big(\left[\, F_s^{T1};\ F_s^{T2};\ \Delta_s;\ |\Delta_s| \,\right]\big), \tag{4}$$

where $[\,\cdot\,;\cdot\,]$ is channel concatenation, $|\cdot|$ is the element-wise absolute value, and ConvRes is a convolution followed by residual blocks. The signed residual keeps the direction of a change and the absolute residual keeps its magnitude.

The adapter thus passes three aligned streams to the decoder at every scale, the pre-change stream $F^{T1}$, the post-change stream $F^{T2}$, and the change stream $F^{C}$, so the decoder reads the appearance of both dates together with a direct signal of where and how strongly the scene has changed.

### 3.4 Flexible Mask Decoder

FlexMD first tokenizes the three visual streams. Each of $F^{T1}$, $F^{T2}$, and $F^{C}$ passes a $1\times1$ convolution, a pooling step, and a flatten, and the three are concatenated into a token bank $F^{V} \in \mathbb{R}^{B\times N\times C}$ of $N{=}3{\times}(64^2{+}32^2{+}16^2){=}16128$ tokens, where $B$ is the batch size. The change queries form $F^{Q} \in \mathbb{R}^{B\times K\times C}$ with $C{=}256$.

A Bidirectional Feature Interaction Module (BFIM) of four layers couples the two token sets. With $F_i^{Q}$ and $F_i^{V}$ the states at layer $i$, initialized as $F_0^{Q}{=}F^{Q}$ and $F_0^{V}{=}F^{V}$, each layer updates the queries by self-attention and query-to-visual cross-attention, and updates the visual tokens by visual-to-query cross-attention, every attention followed by a feed-forward network:

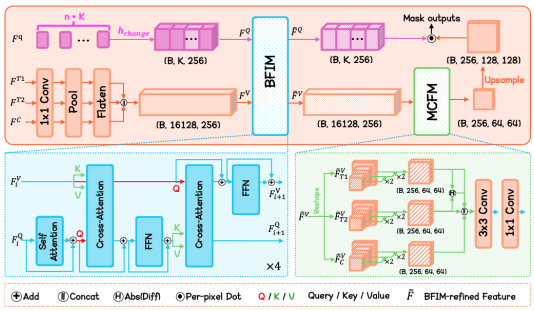



Figure 3: The Flexible Mask Decoder (FlexMD). BFIM refines change queries and visual tokens. MCFM fuses pre-change, post-change, and difference features into one mask for each query.

$$
\begin{aligned}
\hat{F}_i^{Q} &= F_i^{Q} + \mathrm{SA}(F_i^{Q}),\\
F_{i+1}^{Q} &= \mathrm{FFN}\big(\hat{F}_i^{Q} + \mathrm{CA}(\hat{F}_i^{Q},\, F_i^{V})\big),\\
F_{i+1}^{V} &= \mathrm{FFN}\big(F_i^{V} + \mathrm{CA}(F_i^{V},\, F_{i+1}^{Q})\big),
\end{aligned}
\tag{5}
$$

where SA is self-attention and $\mathrm{CA}(\mathrm{q},\mathrm{kv})$ is cross-attention whose first argument supplies the query. After four layers we obtain the refined sets $\tilde{F}^{Q}$ and $\tilde{F}^{V}$.

A Multi-scale Change Fusion Module (MCFM) turns $\tilde{F}^{V}$ into a dense feature. It reshapes $\tilde{F}^{V}$ back to the three streams, upsamples each stream top-down to $64\times64$ to give $\tilde{F}_{T1}^{V}$, $\tilde{F}_{T2}^{V}$, and $\tilde{F}_{C}^{V}$, and fuses them by

$$
\begin{aligned}
G &= \big[\,\tilde{F}_{T1}^{V};\ \tilde{F}_{T2}^{V};\ |\tilde{F}_{T1}^{V}{-}\tilde{F}_{T2}^{V}|;\ \tilde{F}_{C}^{V}\,\big],\\
D &= \mathrm{Up}_2\big(\mathrm{Conv}_{1\times1}\,\mathrm{Conv}_{3\times3}\,G\big),
\end{aligned}
\tag{6}
$$

where $G$ concatenates the four terms along the channel axis into $4C$ channels, the two convolutions reduce them to $C$ channels at $64\times64$, and $\mathrm{Up}_2$ is a $\times2$ upsampling, yielding a dense change feature $D \in \mathbb{R}^{B\times C\times128\times128}$ that carries pre-change, post-change, and change evidence at mask resolution.

Each refined query reads out one mask by a per-pixel dot product with $D$, a sigmoid $\sigma$, and upsampling to the input size, as in

$$
M_k = \mathrm{Up}\big(\sigma(\langle\tilde{F}_k^{Q},\, D\rangle)\big), \qquad k = 1,\dots,K,
\tag{7}
$$

so one evaluation of $D$ produces all $K$ masks at $128\times128$, upsampled to $512\times512$.

### 3.5 Training

We train the whole model end-to-end with three equally weighted terms, $\mathcal{L} = \mathcal{L}_{\mathrm{CE}}+\mathcal{L}_{\mathrm{BCE}}+\mathcal{L}_{\mathrm{Dice}}$, where $\mathcal{L}_{\mathrm{CE}}$ is the language-modeling loss on the assistant response, teaching $\Phi$ to produce the correct text and its $K$ `<change>` tokens, and $\mathcal{L}_{\mathrm{BCE}}$ and $\mathcal{L}_{\mathrm{Dice}}$ are per-mask losses against the ground-truth masks $M_k^{*}$.

We aggregate the mask terms hierarchically, averaging the $K$ slots inside a sample before averaging over the batch, so a $K{=}30$ semantic sample does not outweigh a $K{=}1$ binary sample. Absent slots are kept and supervised with an all-zero mask $M_k^{*}{=}\mathbf{0}$, and their weight warms up from $0.05$ to $0.25$, which trains the rejection that referring and reasoning depend on.

The mask loss flows back through the `<change>` queries into $\Phi$, so all components are trained jointly, with the encoder unfrozen at a smaller learning rate and $\Phi$ adapted with LoRA. One SAM2.1 Hiera-L encoder, one LoRA-adapted $\Phi$, and one FlexMD serve all four paradigms.

# 4 UNIFIED TASKS AND BENCHMARKS

## 4.1 BINARY AND SEMANTIC CHANGE DETECTION

For binary detection, each pair yields one sample with a single <change> token and the building-change mask as its target. We use WHU-CD in the Open-CD 512-tile format.

We recast semantic change detection over SECOND as a set of directional targets: each changed pixel carries an ordered from-to transition, pixels of the same transition form one mask, and enumerating the six classes gives thirty ordered directions. Absent directions receive all-zero masks, so the model must both find the transitions that occur and suppress those that do not. A standard semantic change map is recovered per pixel. If all thirty direction probabilities fall below $0.5$, the pixel is unchanged and both dates receive the no-change label, and otherwise the pixel takes the from and to classes of its most probable direction.

## 4.2 CATEGORY-REFERRING CHANGE DETECTION

A query lists $K$ categories and the model returns the changes related to each, where a category is related if it appears before or after the change. The mask of a category therefore covers every transition that starts or ends in it. When A→B and B→C both occur, the query for A and C returns the A→B pixels in the first mask and the B→C pixels in the second. InstructCD answers with one <change> token per category and decodes all $K$ masks jointly. Based on SECOND, we curate SECOND-Ref, a training set of $20{,}000$ queries built by enumerating candidate category combinations per image, including absent categories, and reweighting them by class rarity. For evaluation we sample $K$ from one to six and score the union of the selected masks, giving $1{,}694$ queries with $142$ all-negative cases. We detail the construction in the technical appendix.

## 4.3 REASONING CHANGE DETECTION: INSTRUCTCD-29K

InstructCD-29K organizes referring difficulty into a four-level curriculum built from SECOND and xBD, with $29{,}310$ samples split into $23{,}450$ training, $2{,}930$ validation, and $2{,}930$ test cases with no base-image leakage. L1 names a direct class transition, L2 adds attribute, spatial, and compositional constraints, L3 turns to functional phenomena and disaster causality, and L0 is an adversarial negative level whose correct answer is an empty mask. Reasoning here means resolving the described targets from attributes, spatial relations, functional phenomena, or disaster causes that are grounded in the annotations, and open-domain commonsense inference beyond these cues is outside the scope of the benchmark. Two invariants keep the benchmark faithful: every supervised fact is recomputed deterministically from ground-truth annotation while GPT-5.5 only synthesizes and paraphrases the referring phrase from these facts, and each response carries exactly as many <change> tokens as masks, so a sample can be audited automatically. The technical appendix details the levels and the distractor design.

All splits are formed per source image pair, keyed by source dataset and image, before any instruction is generated, so every instruction, paraphrase, and mask derived from one pair shares its split. Every SECOND test sample of every paradigm, including the SECOND part of the InstructCD-29K test set, comes from the official SECOND test split, and every SECOND training sample comes from the official training split. No test image therefore reaches training through any paradigm, including in joint training.

# 5 EXPERIMENTS

## 5.1 SETUP AND PROTOCOL

We compare methods under one protocol: the same test split, the same $512\times512$ bitemporal inputs and ground truth, the same $0.5$ threshold with no test-time augmentation, and an adapter that maps every method to a common mask format. Every baseline except Change-LISA is retrained on the same splits with the same SAM2.1 Hiera-L image encoder as InstructCD, and Change-LISA keeps its larger SAM ViT-H encoder, which we fine-tune, so no baseline in Table 2 relies on a smaller image encoder than InstructCD. Because every baseline is retrained with this encoder on our splits,

Table 2: Results under the unified protocol (referring and reasoning $\times 100$). All baselines share the SAM2.1 Hiera-L encoder of InstructCD except Change-LISA, which keeps and fine-tunes SAM ViT-H. InstructCD gives identical results under teacher-forced and autoregressive decoding. NCA is NCA@0.01 on negative queries, – marks an unaddressed paradigm, and *Flex* marks joint decoding of a variable number of masks. Best in bold, second best underlined.

| Method | Binary WHU-CD | | Semantic SECOND | | | | Referring SECOND-Ref | | | Reasoning InstructCD-29K | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1↑ | IoU↑ | OA↑ | mIoU↑ | Fscd↑ | SeK↑ | F1↑ | IoU↑ | NCA↑ | F1↑ | Ovr↑ | L1↑ | L2↑ | L3↑ | NCA↑ | Flex |
| *Binary-CD specialists* | | | | | | | | | | | | | | | | |
| BIT | 85.67 | 74.93 | – | – | – | – | – | – | – | – | – | – | – | – | – | ✗ |
| ChangeFormer | 92.30 | 85.70 | – | – | – | – | – | – | – | – | – | – | – | – | – | ✗ |
| Changer | 86.09 | 75.57 | – | – | – | – | – | – | – | – | – | – | – | – | – | ✗ |
| TTP | 93.74 | 88.22 | – | – | – | – | – | – | – | – | – | – | – | – | – | ✗ |
| *Semantic-CD specialists* | | | | | | | | | | | | | | | | |
| SCanNet | – | – | 87.19 | 72.87 | 63.28 | 23.05 | – | – | – | – | – | – | – | – | – | ✗ |
| Bi-SRNet | – | – | 87.51 | 73.13 | 62.67 | 22.73 | – | – | – | – | – | – | – | – | – | ✗ |
| MambaSCD | – | – | 87.85 | 72.81 | 62.48 | 22.41 | – | – | – | – | – | – | – | – | – | ✗ |
| SCD-SAM | – | – | 85.68 | 71.23 | 60.84 | 20.79 | – | – | – | – | – | – | – | – | – | ✗ |
| *Unified (partial)* | | | | | | | | | | | | | | | | |
| UniChange | 93.80 | 88.32 | 87.83 | 72.27 | 63.07 | 22.28 | – | – | – | – | – | – | – | – | – | ✗ |
| *Language-driven (referring + reasoning)* | | | | | | | | | | | | | | | | |
| ChangeCLIP | – | – | – | – | – | – | 58.95 | 45.69 | 84.51 | 11.36 | 8.13 | 11.44 | 0.02 | 13.40 | **98.91** | ✗ |
| RCDNet | – | – | – | – | – | – | 61.48 | 48.55 | **92.25** | 39.65 | 31.94 | 44.74 | 16.79 | 30.57 | 98.36 | ✗ |
| Change-LISA | – | – | – | – | – | – | 62.14 | 48.86 | 76.06 | 49.87 | 39.01 | 44.68 | 30.94 | **40.20** | 97.64 | ✗ |
| **InstructCD (ours)** | **94.18** | **89.00** | **87.99** | **73.16** | **64.09** | **23.59** | **63.79** | **50.57** | 80.99 | **52.67** | **40.74** | **44.87** | **37.06** | 38.66 | 95.45 | ✓ |

its numbers can differ from the published ones. UniChange, for example, reports 90.41 IoU on WHU-CD in its original setting and reaches 88.32 under our protocol. Binary detection uses the Open-CD changed-class IoU at 0.5, and semantic detection uses the SECOND official test set with the canonical class order.

Every paradigm uses the changed-class intersection over union (IoU) and its F1 score. Semantic change detection adds the overall accuracy (OA), the mean IoU (mIoU) over the no-change and semantic classes, the semantic change F-score (Fscd) on changed pixels, and the separated kappa (SeK), the primary SECOND metric, which removes the dominant no-change class. Reasoning averages IoU and F1 over the samples of each level, and the overall score averages over all L1 to L3 samples, which weights the levels by their test counts of 999, 791, and 590. On negative queries we measure the false-positive area (FPA), the predicted foreground as a fraction of the image, and the negative-consistency accuracy (NCA@0.01), the fraction of negative queries with FPA below 0.01.

### 5.2 Binary and Semantic Change Detection

Table 2 reports F1 and IoU on WHU-CD, and Figure 4(a) shows qualitative results. InstructCD reaches 94.18 F1 and 89.00 IoU, 0.38 F1 and 0.68 IoU above UniChange, the strongest baseline. A language-driven interface therefore does not sacrifice accuracy on the purely geometric binary task.

On SECOND, with qualitative results in Figure 4(b), InstructCD ranks first on every metric, with margins of 0.54 SeK and 0.81 Fscd over SCanNet, while OA and mIoU are on par with the best specialists at leads of 0.14 and 0.03 (Ding et al., 2024; 2022; Chen et al., 2024a; Mei et al., 2024). The response names the direction of every slot before its `<change>` token, so the semantic map is assembled from masks whose classes are read directly from the structured text.

### 5.3 Category-Referring Change Detection

Table 2 reports the overall referring results under the union protocol, Table 3 varies the number of queried categories, and Figure 4(c) shows qualitative results. InstructCD is best on F1 (63.79), on IoU (50.57), and at every $K$. Its margin over the strongest baseline is 0.45 IoU at $K{=}1$ and 0.99 to 2.43 for $K{\geq}2$, which suggests that joint decoding helps more once the instruction names more than one target. Each mask takes its category from the structured response, which names the

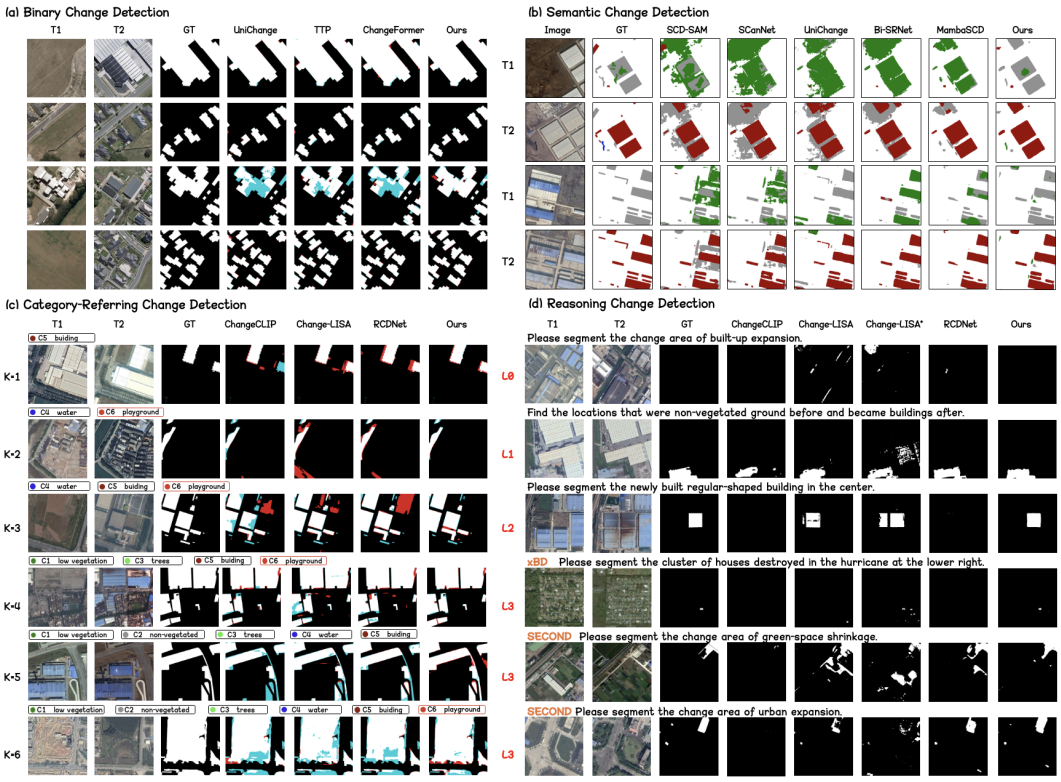



Figure 4: Qualitative results of the four paradigms. (a) Binary change detection on WHU-CD. (b) Semantic change detection on SECOND, with pre-change (T1) and post-change (T2) maps. (c) Category-referring change detection on SECOND, one row per $K$=1–6 with the queried classes above each row. (d) Reasoning change detection on InstructCD-29K, one row per level with the instruction above each row.

Table 3: Category-referring IoU ($\times$100) versus the number of queried categories $K$.

| Model | $K$=1 | $K$=2 | $K$=3 | $K$=4 | $K$=5 | $K$=6 |
|---|---|---|---|---|---|---|
| **InstructCD (ours)** | **41.60** | **48.01** | **49.83** | **52.57** | **53.62** | **54.15** |
| Change-LISA (union) | 40.92 | 45.58 | 48.77 | 50.55 | 51.97 | 52.09 |
| RCDNet | 41.15 | 43.99 | 46.77 | 51.58 | 52.14 | 52.44 |
| ChangeCLIP (union) | 35.78 | 41.30 | 44.68 | 48.56 | 49.91 | 49.73 |
| ChangeCLIP (single) | 36.08 | 39.85 | 43.13 | 46.66 | 48.00 | 48.24 |

category before the corresponding <change> token. On the 142 all-negative queries, RCDNet and ChangeCLIP reach higher NCA, 92.25 and 84.51 against 80.99, at lower IoU.

### 5.4 Reasoning Change Detection

Table 2 reports results on InstructCD-29K, and Figure 4(d) shows qualitative results. InstructCD leads on overall IoU (40.74) and F1 (52.67). Its margin is 6.12 IoU on L2, where a target is pinned down only by jointly resolving what, where, and which one, and 0.13 on L1, while Change-LISA is 1.54 higher on L3. On the 550 L0 negatives, InstructCD reaches 95.45 NCA with a mean false-positive area of 0.215 percent, below Change-LISA at 97.64, RCDNet at 98.36, and ChangeCLIP at 98.91. These baselines reach lower overall IoUs of 39.01, 31.94, and 8.13, yet false positives on absent targets remain a weakness that joint training reduces, raising NCA to 97.45 (Table 4).

### 5.5 Joint Training and Ablations

Table 4 compares one model trained on all four paradigms with per-task models of the same architecture and size. The joint model stays within 1.25 points of them on every positive metric and raises NCA from 80.99 to 83.80 on referring and from 95.45 to 97.45 on reasoning. Generating the

Table 4: Joint training under teacher-forced (TF) and autoregressive (AR) decoding ($\times 100$). Per-task models are trained on one paradigm each, and the joint model is one model trained on all four paradigms with the same architecture and size. NCA is NCA@0.01. Best in bold.

| | Binary | | Semantic | | | | | Referring | | | Reasoning | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | IoU | F1 | SeK | Fscd | mIoU | Fbcd | IoU | IoU | F1 | NCA | IoU | F1 | NCA |
| Per-task (TF) | **89.00** | **94.18** | **23.59** | **64.09** | **73.16** | **73.57** | **58.19** | **50.57** | **63.79** | 80.99 | **40.74** | **52.67** | 95.45 |
| Joint (TF) | 88.51 | 93.90 | 22.96 | 64.01 | 72.63 | 72.80 | 57.24 | 49.32 | 62.61 | **83.80** | 40.49 | 51.45 | **97.45** |
| Joint (AR) | 88.51 | 93.90 | 22.78 | 63.87 | 72.53 | 72.66 | 57.06 | 49.32 | 62.61 | **83.80** | 40.49 | 51.45 | **97.45** |

Table 5: Ablations ($\times 100$). The upper block scales the SAM2.1 Hiera encoder, and the lower block replaces modules. Best in bold.

| Config. | Binary | | Semantic | | Ref. | Reason. | |
|---|---|---|---|---|---|---|---|
| | F1 | IoU | SeK | mIoU | IoU | F1 | IoU |
| *Backbone scale (SAM2.1 Hiera)* | | | | | | | |
| Hiera-T | 91.94 | 85.09 | 21.85 | 72.34 | 48.49 | 51.76 | 39.74 |
| Hiera-S | 92.09 | 85.34 | 22.40 | 72.38 | 48.17 | 52.13 | 40.13 |
| Hiera-B+ | 92.22 | 85.57 | 22.66 | 72.35 | 49.47 | 52.04 | 39.94 |
| Hiera-L | 94.18 | 89.00 | 23.59 | 73.16 | 50.57 | 52.67 | 40.74 |
| *Module ablation (main line)* | | | | | | | |
| Full | **94.18** | **89.00** | **23.59** | 73.16 | **50.57** | **52.67** | **40.74** |
| −TSA | 93.31 | 87.46 | 23.54 | **73.40** | 49.48 | 52.09 | 40.32 |
| −FlexMD | 93.46 | 87.73 | 22.83 | 72.68 | 49.19 | 47.51 | 35.79 |
| −TSA−FlexMD | 93.50 | 87.79 | 22.79 | 72.83 | 49.43 | 48.50 | 36.41 |

response autoregressively gives the same binary, referring, and reasoning scores as teacher forcing and lowers the semantic scores by at most $0.18$. The ablations replace rather than delete each module: −TSA uses the absolute difference $|F_s^{T2} - F_s^{T1}|$ at every scale, and −FlexMD uses a plain FPN decoder that reads the same queries. Table 5 shows that removing the Flexible Mask Decoder lowers the semantic, referring, and reasoning scores more than removing the Temporal Scale Adapter, most on reasoning, where overall IoU decreases from $40.74$ to $35.79$.

Scaling the shared encoder from Hiera-T to Hiera-L improves every paradigm. Model scale alone does not account for the margins in Table 2: the −TSA−FlexMD row keeps the same encoder and MLLM but uses a plain decoder, and it falls below the strongest baseline on every binary, semantic, and reasoning metric in Table 5. Removing the Temporal Scale Adapter lowers six of the seven metrics, and semantic mIoU is the only exception.

### 5.6 Failure Cases and Cost

Three failure patterns recur. Rare transitions remain hard: in the per-transition analysis of the technical appendix, frequent conversions such as bare ground to building localize well, while rare ones sit at the tail with wide confidence intervals. Single-category referring is the weakest referring setting at $41.60$ IoU for $K{=}1$ against $54.15$ for $K{=}6$, compositional L2 queries are the hardest reasoning level at $37.06$, and absent targets still draw residual foreground. For cost, the teacher-forced path processes one $512{\times}512$ pair in about $231$ ms on one H100 GPU with $19.9$ GiB peak memory, and the latency grows by $0.32$ percent from one to six masks because all masks share one pass.

## 6 Conclusion

We present InstructCD, which unifies binary, semantic, category-referring, and reasoning change detection through one interface that decodes each `<change>` token of a language query into a mask. It achieves the best primary metric on all four paradigms, and a single jointly trained model rivals per-task specialists. Limitations remain. The model assumes bitemporal pairs, struggles with off-nadir imagery, and is heavier than specialist detectors.

## AI USE STATEMENT

For the manuscript text, we used generative AI tools only for language polishing. GPT-5.5 also synthesizes and paraphrases the referring expressions of our benchmarks as part of their construction, described in Section 4 and the technical appendix, after all supervised facts are computed from the annotations. We did not use generative AI tools for research ideation, method design, experiments, or data analysis. We have reviewed all AI-assisted text and take responsibility for the content of this work.

## REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, 2025. URL https://arxiv.org/abs/2502.13923.

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*, pp. 207–210, 2022. doi: 10.1109/IGARSS46834.2022.9883686. URL https://doi.org/10.1109/IGARSS46834.2022.9883686.

Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. Fully Convolutional Siamese Networks for Change Detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 4063–4067, Athens, 2018. IEEE. doi: 10.1109/ICIP.2018.8451652. URL https://ieeexplore.ieee.org/document/8451652/.

Hao Chen and Zhenwei Shi. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sensing*, 12(10):1662, 2020. doi: 10.3390/rs12101662. URL https://www.mdpi.com/2072-4292/12/10/1662.

Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote Sensing Image Change Detection With Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. doi: 10.1109/TGRS.2021.3095166. URL https://ieeexplore.ieee.org/document/9491802/.

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, 2024a. doi: 10.1109/TGRS.2024.3417253. URL https://ieeexplore.ieee.org/document/10565926/.

Keyan Chen, Chengyang Liu, Wenyuan Li, Zili Liu, Hao Chen, Haotian Zhang, Zhengxia Zou, and Zhenwei Shi. Time Travelling Pixels: Bitemporal Features Integration with Foundation Model for Remote Sensing Image Change Detection. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 8581–8584, 2024b. doi: 10.1109/IGARSS53475.2024.10640593. URL https://doi.org/10.1109/IGARSS53475.2024.10640593.

Lei Ding, Haitao Guo, Sicong Liu, Lichao Mou, Jing Zhang, and Lorenzo Bruzzone. Bi-Temporal Semantic Reasoning for the Semantic Change Detection in HR Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. doi: 10.1109/TGRS.2022.3154390. URL https://ieeexplore.ieee.org/document/9721305/.

Lei Ding, Jing Zhang, Haitao Guo, Kai Zhang, Bing Liu, and Lorenzo Bruzzone. Joint Spatio-Temporal Modeling for Semantic Change Detection in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2024.3362795. URL https://ieeexplore.ieee.org/document/10443352/.

Sijun Dong, Libo Wang, Bo Du, and Xiaoliang Meng. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208:53–69, 2024. doi: 10.1016/j.isprsjprs.2024.01.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0924271624000042.

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, 2022. doi: 10.1109/LGRS.2021.3056416. URL https://ieeexplore.ieee.org/document/9355573/.

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. doi: 10.1109/TGRS.2023.3277496. URL https://ieeexplore.ieee.org/document/10129139/.

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. doi: 10.1109/JSTARS.2023.3264802. URL https://doi.org/10.1109/JSTARS.2023.3264802.

Zhenyang Huang, Xiao Yu, Yi Zhang, Decheng Wang, and Hang Ruan. ReasonCD: A Multimodal Reasoning Large Model for Implicit Change-of-Interest Semantic Mining, 2025. URL https://arxiv.org/abs/2512.19354.

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, 2019. doi: 10.1109/TGRS.2018.2858817. URL https://ieeexplore.ieee.org/document/8444434/.

Xiangyu Jia, Zhibo Chen, Shengyi Zhang, and Xiaojing Xue. Change-LISA: Language-Guided Reasoning for Remote Sensing Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–15, 2026. doi: 10.1109/TGRS.2026.3684817. URL https://ieeexplore.ieee.org/document/11482624/.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003, Paris, France, 2023. IEEE. doi: 10.1109/ICCV51070.2023.00371. URL https://ieeexplore.ieee.org/document/10378323/.

Yilmaz Korkmaz, Jay N. Paranjape, Celso M. de Melo, and Vishal M. Patel. Referring Change Detection in Remote Sensing Imagery. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 106–116, 2026. doi: 10.1109/WACV61042.2026.00019. URL https://doi.org/10.1109/WACV61042.2026.00019.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning Segmentation via Large Language Model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589, Seattle, WA, USA, 2024. IEEE. doi: 10.1109/CVPR52733.2024.00915. URL https://ieeexplore.ieee.org/document/10658574/.

Kaiyu Li, Zepeng Xin, Li Pang, Chao Pang, Yupeng Deng, Jing Yao, Guisong Xia, Deyu Meng, Zhi Wang, and Xiangyong Cao. SegEarth-R1: Geospatial Pixel Reasoning via Large Language Model, 2025. URL https://arxiv.org/abs/2504.09644.

Xiaohe Li, Jiahao Li, Kaixin Zhang, Yuqiang Fang, Leilei Lin, Hong Wang, Haohua Wu, and Zide Fan. Decoding the Delta: Unifying Remote Sensing Change Detection and Understanding with Multimodal Large Language Models, 2026. URL https://arxiv.org/abs/2604.14044.

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. doi: 10.1109/TGRS.2024.3425815. URL https://doi.org/10.1109/TGRS.2024.3425815.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Liye Mei, Zhaoyi Ye, Chuan Xu, Hongzhu Wang, Ying Wang, Cheng Lei, Wei Yang, and Yansheng Li. SCD-SAM: Adapting Segment Anything Model for Semantic Change Detection in Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. doi: 10.1109/TGRS.2024.3407884. URL https://ieeexplore.ieee.org/document/10543161/.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/45c1f6a8cbf2da59ebf2c802b4f742cd-Paper-Conference.pdf.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. PixelLM: Pixel Reasoning with Large Multimodal Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26374–26383, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ren_PixelLM_Pixel_Reasoning_with_Large_Multimodal_Model_CVPR_2024_paper.html.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. GSVA: Generalized Segmentation via Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3858–3869, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xia_GSVA_Generalized_Segmentation_via_Multimodal_Large_Language_Models_CVPR_2024_paper.html.

Zepeng Xin, Kaiyu Li, Luodi Chen, Wanchen Li, Yuchen Xiao, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. SegEarth-R2: Towards Comprehensive Language-guided Segmentation for Remote Sensing Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13199–13210, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Xin_SegEarth-R2_Towards_Comprehensive_Language-guided_Segmentation_for_Remote_Sensing_Images_CVPR_2026_paper.html.

Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Asymmetric Siamese Networks for Semantic Change Detection in Aerial Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022. doi: 10.1109/TGRS.2021.3113912. URL https://doi.org/10.1109/TGRS.2021.3113912.

Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. LISA++: An Improved Baseline for Reasoning Segmentation with Large Language Model, 2023. URL https://arxiv.org/abs/2312.17240.

Xu Zhang, Danyang Li, Xiaohang Dong, Tianhao Wu, Hualong Yu, Jianye Wang, Qicheng Li, and Xiang Li. UniChange: Unifying Change Detection with Multimodal Large Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42169–42179, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_UniChange_Unifying_Change_Detection_with_Multimodal_Large_Language_Model_CVPR_2026_paper.html.

# TECHNICAL APPENDIX

This technical appendix collects the material omitted from the main paper for space. It records the implementation details of the unified interface, a dataset overview and the construction details of the four benchmarks, additional quantitative results that cover joint training, rejection, and variable target counts, detailed ablations, a fairness and efficiency study of the baselines, and extended qualitative comparisons.

## A IMPLEMENTATION DETAILS

InstructCD serves all four paradigms through one response format. Every assistant response restates the referring phrase and closes with a run of `<change>` tokens whose count equals the number of supervised masks. The $k$-th text target, the $k$-th `<change>` token, the $k$-th decoder query, and the $k$-th mask denote the same object, and this order is a deterministic serialization contract rather than a post hoc match. The hidden state that precedes each `<change>` token passes through a projection $3584 \to 3584 \to 256$ to form the query that FlexMD decodes into one mask. The number of targets $K$ is set by the instruction, so all $K$ masks are decoded jointly under shared decoder parameters, with no per-count output head and no decomposition into repeated single-target inference.

Unless marked AR, all reported results use teacher forcing. The response template depends only on the query: binary and semantic responses are fixed, a referring response lists the queried categories in query order, including absent ones, and a reasoning response restates the referring phrase. We feed this template response and read the hidden states of all `<change>` tokens in one pass. Free autoregressive generation reproduces the template response and gives the same accuracy for every per-task model, and the main paper reports both decoding modes for the jointly trained model.

The paradigms differ only in how the instruction sets $K$ and what each mask denotes. Binary detection requests one target and returns one mask. Semantic detection enumerates the thirty ordered directions. A standard semantic-change map is reconstructed per pixel, which is unchanged when all thirty probabilities fall below $0.5$ and otherwise takes the from and to classes of the most probable direction. Category-referring detection lists $K$ classes with $K$ between one and six, and reasoning detection sets $K$ to the number of described targets. During training the mask loss is first averaged over the valid slots inside each sample and then aggregated across the batch, so a thirty-slot semantic sample does not receive thirty times the weight of a single-slot sample.

Rejection is represented in the output space rather than in language. When the requested target is absent, the response still emits a single `<change>` token, but the corresponding slot is supervised with an all-zero mask. The empty-mask loss enters with a warm-up weight that rises from $0.05$ to $0.25$, which reduces the risk of a collapse to empty predictions early in training. This keeps the text, token, and mask interface uniform across positive and negative queries and makes refusal measurable through the false-positive area and the negative-consistency accuracy that the quantitative section defines.

The training objective combines an assistant cross-entropy term on the response text with a per-slot binary cross-entropy term and a Dice term on the masks. All four paradigms are served by one architecture, a SAM2.1 Hiera-L encoder, the Temporal Scale Adapter, and the FlexMD decoder behind a single query interface, so no paradigm needs a separate backbone or decoder.

## B DATASET OVERVIEW

Figure 5 summarizes the full InstructCD data pool across the four paradigms. The pool holds $57{,}616$ annotated records: $1{,}950$ binary (WHU-CD), $4{,}662$ semantic (SECOND directional), $21{,}694$ category-referring (SECOND-Ref), and $29{,}310$ reasoning (InstructCD-29K, drawn from SECOND and xBD). To avoid any confusion with the main paper, the figure counts *records over all splits*, whereas the interface table in the main paper lists *test-split* sizes: the semantic $4{,}662$ is $2{,}968$ train/val plus $1{,}694$ test images, the referring $21{,}694$ is $20{,}000$ training queries plus $1{,}694$ union-test queries, and the reasoning $29{,}310$ splits into $23{,}450/2{,}930/2{,}930$. The natural-language referring expressions and their paraphrases are synthesized with GPT-5.5, while the binary and semantic responses follow fixed templates.

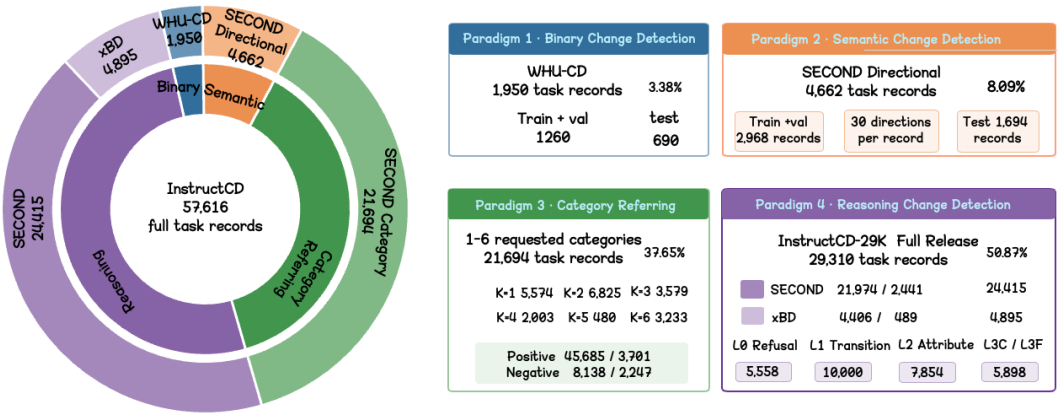



Figure 5: Overview of the InstructCD data pool across the four paradigms, with per-paradigm record counts and sources. Counts are over all splits. See the text for the correspondence to the main-paper test-split sizes.

## C Dataset Construction Details

Figure 7 shows one worked example per paradigm, each with the system prompt, the user instruction, and the assistant response that restates the reference and closes with the <change> tokens whose count equals the number of masks. The same interface serves all four paradigms; only the instruction and the number of emitted tokens change.

### C.1 Semantic Directional Targets

The semantic paradigm is built from SECOND, which provides bitemporal per-pixel semantic labels over six land-cover classes: low vegetation, non-vegetated ground, tree, water, building, and playground. We compare the pre-change and post-change semantic maps pixel by pixel, read an ordered from-to transition for every changed pixel, and aggregate pixels of the same transition into one mask. Enumerating the six classes gives thirty ordered directions, and each direction becomes one <change> target described by the phrase "{from} to {to} change areas are <change>", as in Figure 7. Directions that occur receive their true masks, and directions that do not occur receive all-zero masks, so the same all-query response trains both detection and suppression. The $k$-th direction phrase, the $k$-th <change> token, the $k$-th query, and the $k$-th mask always denote the same transition. We use the SECOND split of 2,671 training, 297 validation, and 1,694 official test pairs at $512 \times 512$, and reconstruct standard semantic-change maps per pixel, marking a pixel unchanged when all thirty probabilities fall below 0.5 and otherwise taking its most probable direction.

### C.2 SECOND-Ref: Category-Referring Set

SECOND-Ref reuses the SECOND bitemporal semantic labels to build category-level referring supervision. For each of the six classes we form a "class-related change" mask that marks every pixel whose transition involves that class either before or after the change. A query selects $K$ classes and the model returns one mask per class, with an empty mask for a queried class that has no related change in that image.

Enumerating all class combinations per image is redundant, so the training set is built by candidate enumeration followed by rule-based filtering under a fixed random seed. The candidate pool keeps, for every base image, all positive class combinations, single-class queries, mid-sized combinations, and queries with injected negative classes, and is then reweighted by class rarity. The resulting training set contains 20,000 queries over 53,823 referents, of which 45,685 are positive and 8,138 (15.12%) are empty, with 2.69 categories per query on average. For the union evaluation, each test image draws $K$ classes with $K$ sampled uniformly from one to six, and the ground truth is the union of the selected class masks. This yields 1,694 test queries over 5,948 selected referents (3,701

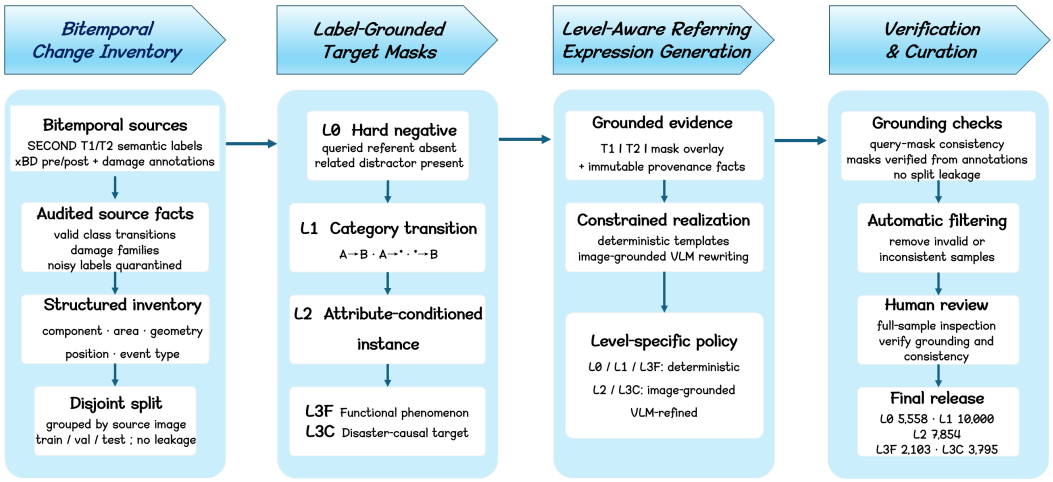



Figure 6: The four-stage InstructCD-29K synthesis pipeline: change inventory, label-grounded target masks, level-aware referring-expression generation, and verification.

positive, 2,247 empty), of which 142 are all-negative. The per-$K$ counts are 274, 295, 273, 270, 307, and 275 for $K = 1$ to 6.

### C.3 InstructCD-29K: Reasoning Benchmark

InstructCD-29K organizes referring difficulty into a four-level curriculum built from SECOND and xBD, with 29,310 samples split into 23,450 training, 2,930 validation, and 2,930 test cases under a base-image no-leak split. Table 6 lists the composition. Figure 6 shows the four-stage synthesis pipeline: a bitemporal change inventory is audited into structured facts and split disjointly; label-grounded target masks are computed for each level; referring expressions are then realized under a level-specific policy; and every sample passes grounding checks, automatic filtering, and human review before release.

Two invariants keep the benchmark faithful. First, the supervised layer and the language surface are strictly decoupled: every mask, class transition, spatial relation, and disaster event is recomputed deterministically from ground-truth annotation, namely the SECOND bitemporal semantic labels and the xBD building polygons with damage grades, while GPT-5.5 only synthesizes and paraphrases the referring phrase into fluent language and may not change any fact or introduce a name the image cannot verify. Second, the supervision is uniform: each response restates the referring phrase and ends with as many `<change>` tokens as masks, so a sample can be audited automatically by matching the token count to the mask count, and an L0 negative is a single `<change>` paired with an all-zero mask.

Two further controls protect the benchmark. Language diversity enters only after the supervised facts are fixed, and every candidate paraphrase is checked against the deterministic source before release, so uniqueness, spatial relations, lexical constraints, and token counts are all verified. Of 7,983 L2 candidates, 7,854 pass the full pipeline and 129 are removed, and xBD L3 candidates that fail the paraphrase check fall back to their deterministic descriptions. Leakage control operates at the level of the original bitemporal pair rather than the individual instruction. Every sample inherits a group identity keyed by its source dataset and its image, so all instructions, paraphrases, target combinations, and masks derived from one pair share the same split. The split is therefore free of shared visual evidence between training and evaluation, not merely free of repeated instruction text, and the SECOND official test data are used for evaluation only.

**L1 — direct semantic change (SECOND).** The query names a change directly by its class transition, and the model localizes where that transition occurs. The referent and the region both come from the ground-truth labels, so this level needs no generative model to be grounded. The phrasing fixes both endpoints (A to B), only the source (A to any), or only the target (any to B).

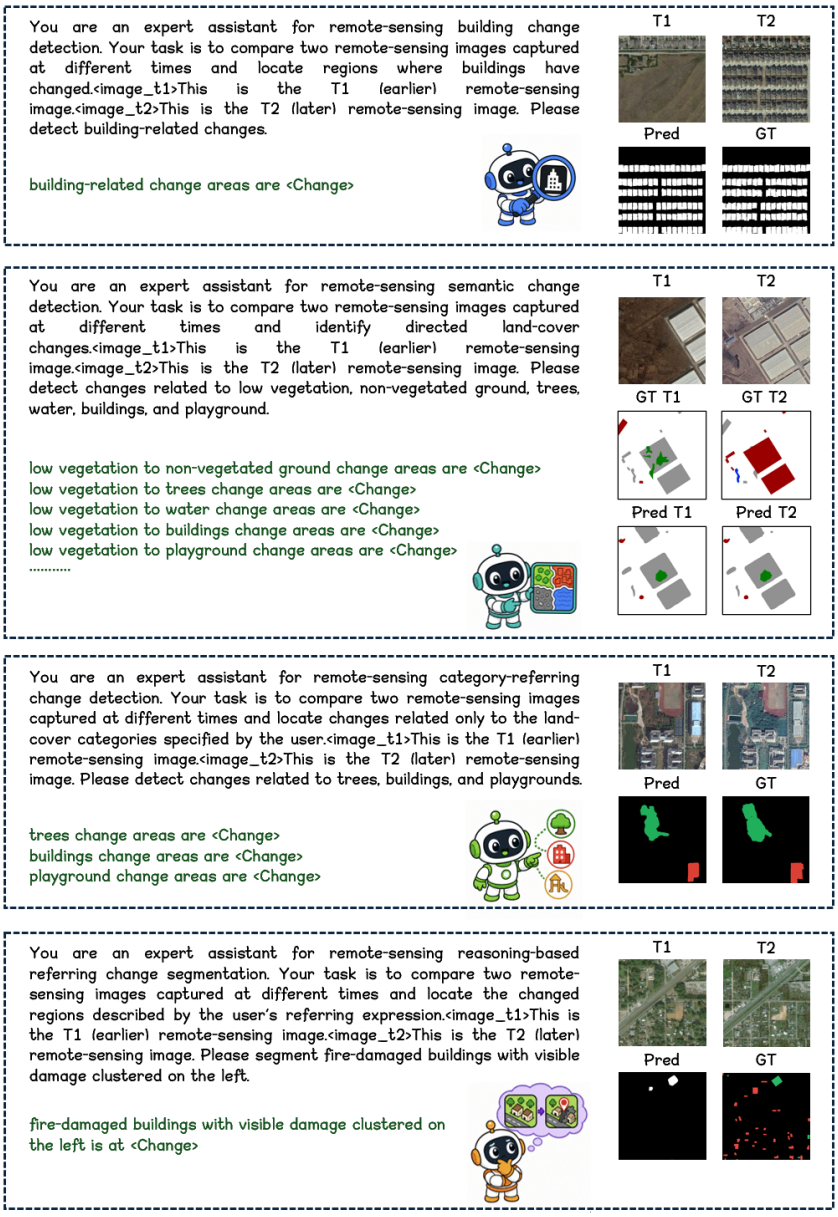



Figure 7: Working examples of the unified interface: each panel shows the system prompt, the instruction, the assistant response with its `<change>` tokens, and the pre/post images, prediction, and ground truth.

**L2 — attribute, context, and compositional reference (SECOND).** Visual and spatial constraints are layered on top of the class, as in "the largest newly added patch of irregular low vegetation in the lower left". The class name alone no longer pins down the target, so the model must jointly resolve what, where, and which one. Attributes such as position, size, shape, and neighborhood relations are read deterministically from the ground-truth masks and then paraphrased by GPT-5.5 under direct view of the highlighted target.

**L3 — functional and causal change (SECOND + xBD).** The query no longer names a class or a location but a functional phenomenon or an event cause. One line draws SECOND functional phenomena (2,103 samples) such as built-up expansion or greenspace loss, where the target is the union of all changes that constitute the phenomenon; each phenomenon must arise from more than one transition, so it cannot be reduced to a single L1 pair. The other line draws xBD disaster causality

Table 6: Composition of InstructCD-29K. The full set splits 23,450/2,930/2,930 (train/val/test) under a base-image no-leak split, with per-level test counts shown.

| Level | Source | Full | Test |
|---|---|---|---|
| L0 | SECOND + xBD | 5,558 | 550 |
| L1 | SECOND | 10,000 | 999 |
| L2 | SECOND | 7,854 | 791 |
| L3 | SECOND + xBD | 5,898 | 590 |
| Total | | 29,310 | 2,930 |

(3,795 samples), referring to buildings by the damaging event and its severity, with the mechanism taken from the authoritative event annotation rather than guessed from the image.

**L0 — empty / negative / reject (SECOND + xBD).** The adversarial counterpart of the positive levels, whose correct answer is an empty mask. Rather than simply removing the target, each L0 sample plants a distractor, a change that looks like the queried one but is not, so the model must resist surface similarity. Distractors are matched to each positive level: a wrong position for L2, a wrong source class or a present-but-unchanged class for L1, a reversed phenomenon for L3, and a wrong damage grade or event for xBD (about 2,943, 915, 600, and 1,100 samples respectively). Relational negatives that cannot be judged under the same rule as the positives are discarded to avoid the most harmful error, labeling a true positive as negative.

## D Extended Capability Probes

Beyond the benchmark metrics, Figure 8 probes behaviors that a unified instruction-following model should exhibit but that the aggregate scores do not isolate. The cases cover out-of-domain generalization to unseen scenes, complex spatial reasoning, out-of-domain and category composition of familiar concepts, refusal on unanswerable reasoning queries, and selective per-category acceptance or rejection within a single multi-category query. These qualitative probes complement the quantitative rejection metrics of this appendix by showing that acceptance and refusal are resolved per query and per category rather than globally.

## E Per-Transition Analysis

Figure 9 breaks down category-referring accuracy by the individual from-to transition on SECOND. The six land-cover classes form $6 \times 5 = 30$ ordered transitions, and $n$ counts the test image pairs in which a transition occurs in the ground truth rather than its pixel area. To keep a direction-level IoU from being dominated by a few images, we report only the 23 transitions with $n \geq 20$. Each pixel-pooled IoU sums intersection and union over all supporting pairs, and its 95 percent interval resamples those pairs 1,000 times with replacement, taking the 2.5 and 97.5 percentiles. Accuracy varies sharply: frequent, visually salient conversions such as bare ground to building localize well while rare transitions sit at the tail, a long-tail difficulty that the aggregate score averages away and that points to instance count and visual salience, not the referring interface, as the main bottleneck.

## F Joint Training Results

Table 7 compares a single model trained jointly on the four paradigms against the per-task models across the primary metrics of every paradigm under the unified protocol. The joint model stays within 1.25 points of each per-task model on positive-detection accuracy, for example 40.49 against 40.74 reasoning overall IoU, 49.32 against 50.57 referring IoU, 22.96 against 23.59 semantic SeK, and 88.51 against 89.00 binary IoU, and at the same time it raises the negative-consistency accuracy on the referring paradigm from 80.99 to 83.80 and on the reasoning paradigm from 95.45 to 97.45. One shared interface therefore serves all four paradigms at a small and consistent cost in positive accuracy while it strengthens rejection, which supports a single network over four separate ones.

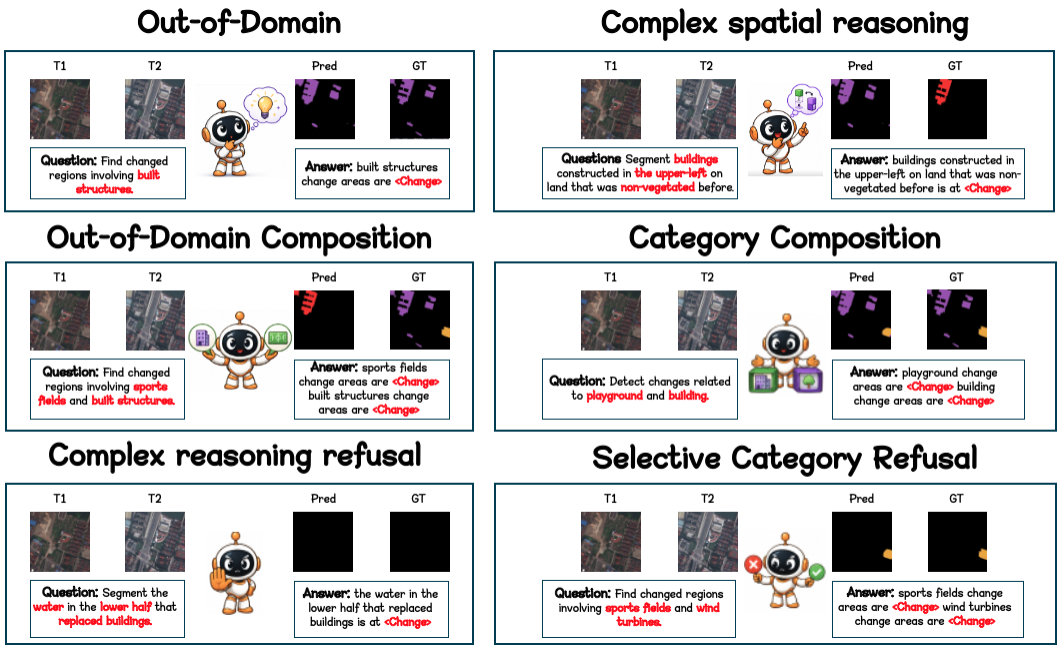



Figure 8: Extended capability probes: out-of-domain generalization, spatial reasoning, out-of-domain and category composition, refusal on unanswerable queries, and selective per-category acceptance/rejection. Check/cross marks denote accepted/rejected categories.

Table 7: Joint training versus per-task models across the primary metrics of all four paradigms, in percent and higher is better. Per-task models are trained on one paradigm each, and Joint is one model trained on all four paradigms. NCA is NCA@0.01, and the reasoning IoU and F1 columns are the overall scores over the L1 to L3 positives.

| Paradigm | Metric | Per-task | Joint | $\Delta$ |
|---|---|---|---|---|
| Binary (WHU-CD) | IoU ↑ | **89.00** | 88.51 | −0.49 |
| | F1 ↑ | **94.18** | 93.90 | −0.28 |
| Semantic (SECOND) | SeK ↑ | **23.59** | 22.96 | −0.63 |
| | $F_{scd}$ ↑ | **64.09** | 64.01 | −0.08 |
| | mIoU ↑ | **73.16** | 72.63 | −0.53 |
| | $F_{bcd}$ ↑ | **73.57** | 72.80 | −0.77 |
| | IoU ↑ | **58.19** | 57.24 | −0.95 |
| Referring (SECOND-Ref) | IoU ↑ | **50.57** | 49.32 | −1.25 |
| | F1 ↑ | **63.79** | 62.61 | −1.18 |
| | NCA ↑ | 80.99 | **83.80** | +2.81 |
| Reasoning (InstructCD-29K) | IoU ↑ | **40.74** | 40.49 | −0.25 |
| | F1 ↑ | **52.67** | 51.45 | −1.22 |
| | NCA ↑ | 95.45 | **97.45** | +2.00 |

## G Rejection Analysis

Table 8 reports rejection behavior on the 550 L0 negatives of the InstructCD-29K test split, placed next to the overall positive IoU because the two quantities are only informative when read together. Let $\hat{M}_i \in \{0,1\}^{H\times W}$ be the predicted mask for the $i$-th negative query. The false-positive area is $a_i = \|\hat{M}_i\|_1/(HW)$, the mean false-positive area is $\mathrm{FPA} = \frac{1}{N_0}\sum_i a_i$, and the negative-consistency accuracy is $\mathrm{NCA}@\tau = \frac{1}{N_0}\sum_i \mathbf{1}[a_i < \tau]$ with $\tau = 0.01$. Read on its own, either rejection metric rewards a model that predicts little foreground, so a method that segments almost nothing attains a near-perfect score at no cost. ChangeCLIP makes this explicit: it reaches the highest negative-

Table 8: Rejection on the 550 L0 negatives of InstructCD-29K, reported next to the overall positive IoU. The InstructCD row is the reasoning model, NCA is NCA@0.01, and FPA is a percentage.

| Method | Overall IoU ↑ | NCA ↑ | FPA ↓ |
|---|---|---|---|
| InstructCD | **40.74** | 95.45 | 0.215 |
| Change-LISA-FT | <u>39.01</u> | 97.64 | 0.094 |
| RCDNet | 31.94 | <u>98.36</u> | <u>0.076</u> |
| Change-LISA-F | 29.48 | 94.55 | 0.227 |
| ChangeCLIP | 8.13 | **98.91** | **0.075** |

Table 9: Union IoU of the referring model as a function of the number of requested categories $K$, from one shared checkpoint.

| $K$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Union IoU ↑ | 41.60 | 48.01 | 49.83 | 52.57 | 53.62 | 54.15 |

consistency accuracy of 98.91 and the lowest false-positive area of 0.075 percent, yet its overall positive IoU falls to 8.13, a refusal that reflects a failure to segment rather than an understanding of the query. RCDNet shows a milder form of the same effect, since its conservative single-class decoding pairs a low false-positive area of 0.076 percent with a positive IoU of only 31.94. The rejection score is therefore meaningful only at a fixed level of recall, and the informative comparison is the joint one. InstructCD has the best positive IoU of 40.74 at 95.45 negative-consistency accuracy and a 0.215 percent false-positive area, so its foreground on absent targets stays below one percent of the image. Change-LISA-FT, RCDNet, and ChangeCLIP reject more reliably at overall IoUs 1.73, 8.80, and 32.61 lower than InstructCD. False positives on absent targets therefore remain a weakness of InstructCD, which joint training reduces by raising its negative-consistency accuracy to 97.45.

## H Variable Target Counts

Table 9 reports the union IoU of the single referring checkpoint as the number of requested categories grows. Accuracy rises monotonically from 41.60 at $K = 1$ to 54.15 at $K = 6$, so co-requested categories act as mutual context rather than as interference, and the shared decoder handles a variable target count without a per-count output head.

## I Detailed Ablations

The two core modules are studied by replacement rather than deletion, so each ablated model keeps the same inputs and outputs and only exchanges the proposed design for a standard implementation. The −TSA row replaces the learned temporal adapter with a direct absolute difference $|F_s^{T2} - F_s^{T1}|$ at every scale, while the pre-change branch, the post-change branch, and the decoder remain. The −FlexMD row replaces the bidirectional query decoder with a plain FPN query decoder that fuses the same multi-scale evidence with convolutions and reads one mask per query through a pixelwise dot product, while the query count, the variable $K$ behavior, the backbone, and the temporal adapter remain. Table 10 reports both replacements on the reasoning benchmark. Replacing FlexMD lowers overall IoU from 40.74 to 35.79, which identifies the bidirectional query to evidence interaction as the main source of accuracy, whereas replacing the temporal adapter alone costs 0.42, so the temporal adapter contributes mainly in combination with FlexMD.

## J Baseline Fairness and Efficiency

Apart from Change-LISA, which keeps its SAM ViT-H encoder, every baseline is retrained with the SAM2.1 Hiera-L encoder of InstructCD on the same splits. The reasoning baselines are trained under a single-variable control. Change-LISA-F (official frozen) keeps the SAM ViT-H image encoder fixed, while Change-LISA-FT (fine-tuned) trains it, so the two differ only in the

Table 10: Module ablation on the reasoning benchmark by replacement. Each row exchanges one proposed module for a standard counterpart of equal input and output capability.

| Configuration | Overall IoU ↑ | Overall F1 ↑ |
|---|---|---|
| InstructCD (full) | **40.74** | **52.67** |
| Direct difference (−TSA) | <u>40.32</u> | <u>52.09</u> |
| Plain-FPN decoder (−FlexMD) | 35.79 | 47.51 |
| Direct difference + Plain-FPN | 36.41 | 48.50 |

Table 11: Single-variable control on Change-LISA over InstructCD-29K. Change-LISA-F and Change-LISA-FT differ only in whether the image encoder is trained. Here NCA is NCA@0.01 and FPA is a percentage.

| Setting | F1 ↑ | IoU ↑ | L2 ↑ | NCA ↑ | FPA ↓ |
|---|---|---|---|---|---|
| Change-LISA-F | 40.45 | 29.48 | 30.30 | 94.55 | 0.227 |
| Change-LISA-FT | <u>49.87</u> | <u>39.01</u> | <u>30.94</u> | **97.64** | **0.094** |
| InstructCD | **52.67** | **40.74** | **37.06** | <u>95.45</u> | <u>0.215</u> |

`--train_sam_encoder` flag while both still train the language adapter, projection layers, bitemporal fusion, and mask decoder. Both share the InstructCD-29K data, twenty epochs, and the same test split, and both select epoch fifteen. Table 11 shows that training the encoder raises overall IoU from $29.48$ to $39.01$ and L1 by $15.97$ and L3 by $10.53$, yet the compositional L2 level gains only $0.64$. InstructCD stays ahead of Change-LISA-FT by $1.73$ overall IoU and $6.12$ on L2. The main paper reports Change-LISA-FT; Change-LISA-F is provided here for completeness.

Table 12 measures conditional inference as $K$ grows. Each row uses one $512 \times 512$ bitemporal pair, batch size one, BF16, a single H100 80GB device, twenty warm-up passes, and one hundred timed passes, and measures the teacher-forced multi-mask path used for the reported results unless marked AR, without autoregressive text generation. Raising the number of masks from one to six increases latency by $0.32$ percent and peak memory by $16.3$ MiB, while mask throughput rises to $5.98$ times, because FlexMD reuses the shared bitemporal representation and pays only a small marginal cost for each additional identity-preserving mask.

Table 12: Conditional inference cost as the number of masks $K$ grows, on one H100 80GB device at $512 \times 512$ with batch size one.

| $K$ | Latency (ms) | Throughput (masks $s^{-1}$) | Peak mem. (GiB) |
|---|---|---|---|
| 1 | 231.42 | 4.32 | 19.933 |
| 2 | 230.76 | 8.67 | 19.938 |
| 3 | 231.02 | 12.99 | 19.941 |
| 4 | 231.24 | 17.30 | 19.944 |
| 5 | 231.59 | 21.59 | 19.947 |
| 6 | 232.16 | 25.84 | 19.949 |




Figure 9: Per-transition category-referring IoU on SECOND, sorted descending. Each bar is a pixel-pooled IoU with its supporting image-pair count $n$ and a 95 percent bootstrap interval.