CHAIN-OF-CHANGE DETECTION FOR MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional remote-sensing change detection (CD) models typically follow a common design: Siamese encoders extract features from each date, an interaction module compares them, and a task-specific decoder produces a single output. We refer to this observe-then-compare structure as the CD-specific inductive bias. In practice, these models remain task-specific: binary change detection, semantic change detection, and change captioning are typically addressed by separate models. Multimodal large language models (MLLMs) unify these tasks in a single model, yet retain this bias only through encoder-side feature interaction. Within the LLM, supervision defines the target outputs but leaves observation and comparison implicit, allowing shortcut learning and disagreement between a model's masks and captions about the same change. To address this, we introduce **Chain-of-Change Detection (CoCD)**, which formulates each step of the CD process as an explicitly supervised task and passes the decoded prediction of each stage to the next as evidence. CoCD first *observes* each date by predicting semantic masks for the queried classes, then *compares* the image pair to localize change guided by these masks, and finally *describes* the change from the predicted change mask and its semantic transitions. CoCD connects these stages with two proposed modules, the Semantic State Painter routing the observed semantic states to comparison and the Change Evidence Injector routing the localized change and its transitions to description, extending the CD-specific inductive bias beyond encoder-side feature interaction. Since no existing benchmark provides all three outputs for the same queried classes, we construct CoCD-Dataset from HRSCD and WUSU with aligned per-date semantic masks, change masks, and grounded captions. Experiments show improved change localization, caption quality, and cross-output consistency over unified baselines and a sequencing-only control. Inference-time interventions further demonstrate that downstream predictions use the routed evidence. Code is available at `https://anonymous.4open.science/r/cocd_release-3CE3/`.

1 INTRODUCTION

Remote-sensing change detection (CD) identifies surface changes from images of the same region acquired at different times, supporting urban monitoring, disaster assessment, and land management (Chen & Shi, 2020; Gupta et al., 2019; Daudt et al., 2019). Conventional CD models typically combine Siamese encoders for per-date feature extraction, interaction modules for temporal comparison, and task-specific decoders (Figure 1(a)) (Daudt et al., 2018; Fang et al., 2023). We refer to this observe-then-compare structure as the CD-specific inductive bias, which explicitly promotes cross-temporal comparison. Conventional methods typically address three tasks with separate models: binary change detection (BCD), which localizes changes (Chen et al., 2022; Bandara & Patel, 2022); semantic change detection (SCD), which identifies before-and-after class transitions (Zheng et al., 2022; Ding et al., 2022; Liu et al., 2025); and remote-sensing image change captioning (RSICC), which describes changes in natural language (Liu et al., 2023; 2022).

Multimodal large language models (MLLMs) provide a foundation for unifying these tasks. Early CD MLLMs addressed RSICC, visual question answering, and dialogue through their native language interface (Noman et al., 2025; Xue et al., 2026; Deng et al., 2026). General-purpose segmentation MLLMs extended this interface to dense prediction by conditioning mask decoders on the hidden states of special output tokens (Lai et al., 2024; Ren et al., 2024; Yuan et al., 2025). Building on

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

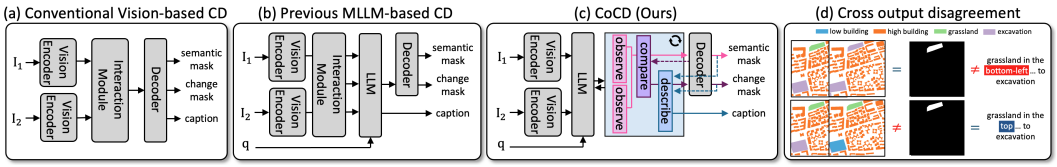

Figure 1: Comparison of conventional CD, previous MLLM-based CD, and CoCD, with the inconsistencies CoCD addresses. **(a)** Conventional CD with dedicated observation, comparison, and decoding components. **(b)** Previous MLLM-based CD, which leaves observation and comparison implicit within the LLM. **(c)** CoCD, which executes a supervised observe–compare–describe sequence and routes decoded predictions to later stages. **(d)** Two inconsistencies: the caption locates change where the mask marks none (top), or the change mask contradicts the semantic masks (bottom).

this interface, recent CD MLLMs support semantic and change masks alongside language responses, moving separate tasks toward a unified output interface (Zhang et al., 2026; Li et al., 2026).

However, a unified output interface does not explicitly structure the underlying CD process. Existing CD MLLMs retain the CD-specific inductive bias only through encoder-side feature interaction (Figure 1(b)). The encoder extracts and compares image features, which the LLM and mask decoders use to generate captions and masks. Supervision specifies the target outputs but does not require the model to identify what is present at each date, determine what has changed between them, and describe those changes in sequence. These intermediate steps remain implicit and can be bypassed. For example, the change decoder may respond to appearance differences without explicitly identifying per-date semantic content, while the LLM may generate a change description from encoder features without consulting the predicted mask. Such shortcuts can compromise both predictive accuracy and cross-output consistency, so that outputs of the same model disagree about the same change: the caption with the change mask, or the semantic masks with the change mask (Figure 1(d)).

We introduce **Chain-of-Change Detection (CoCD)**, which formulates each step of the CD process as a directly supervised task and passes its decoded prediction to the next stage as evidence (Figure 1(c)). CoCD first *observes* each date by predicting semantic masks for the queried classes. It then *compares* the original image pair to localize changes, using the predicted semantic masks as soft spatial guidance. Finally, it *describes* the changes based on the predicted change mask and the semantic transitions. Two modules implement the evidence pathways between stages. The **Semantic State Painter (SSP)** spatially weights LLM class embeddings with decoded class probabilities to construct a dense semantic prompt for change localization. The **Change Evidence Injector (CEI)** renders the predicted change mask and enumerates the corresponding locations and class transitions, providing visual and textual evidence to the LLM before caption generation. Because these pathways carry decoded predictions rather than hidden states, each stage builds on outputs that are themselves directly supervised, closing the shortcuts described above. CoCD thus extends the CD-specific inductive bias from encoder-side feature interaction to a sequence of supervised, interdependent predictions.

Supervising the full sequence requires per-date semantic masks, a binary change mask, and a caption aligned with the same queried classes. We construct CoCD-Dataset from HRSCD (Daudt et al., 2019) and WUSU (Shi et al., 2023), whose full-scene land-cover annotations at both dates enable us to derive query-scoped per-date semantic targets and change masks directly. We then prompt Qwen3-VL-32B (Bai et al., 2025) to express the mask-derived transitions and locations as diverse captions and verify each caption against these facts.

Experiments show improved change localization, captioning, and cross-output consistency over general-purpose segmentation and CD-specific MLLMs. Matched controls and interventions assess the benefit and functional use of intermediate predictions. Our contributions are:

- We propose CoCD, a supervised observe–compare–describe framework for MLLM-based CD that passes each stage's decoded prediction to the next as evidence, extending the CD-specific inductive bias beyond encoder-side feature interaction.
- We introduce SSP and CEI to route decoded semantic predictions to change localization and predicted change evidence to caption generation.
- We construct CoCD-Dataset, aligning per-date semantic masks, binary change masks, and grounded captions for a common set of queried classes.

2 Related Work

Remote-sensing change detection. Conventional CD compares bi-temporal features and decodes binary change masks (Chen et al., 2022; Bandara & Patel, 2022; Fang et al., 2023), semantic transitions (Daudt et al., 2019; Zheng et al., 2022; Ding et al., 2022; Liu et al., 2025), or natural-language descriptions (Liu et al., 2022; 2023). Binary masks locate changes, semantic outputs identify the classes involved, and captions summarize their content and spatial context. These tasks share an observation–comparison structure but are typically optimized for separate outputs (Zhu et al., 2025a; Chen et al., 2024). CoCD connects these complementary outputs within a common query scope.

MLLMs for masks and change understanding. CD MLLMs supply input-side temporal evidence to the LLM, through bi-temporal feature interaction (Noman et al., 2025), difference perception (Deng et al., 2026), or, in ChangeVG (Xue et al., 2026), auxiliary cues such as object counts and locations injected as prompts. Segmentation MLLMs instead extend the output interface, querying dense masks through special tokens (Lai et al., 2024; Ren et al., 2024; Yuan et al., 2025). UniChange adds timepoint- and change-specific tokens for joint BCD and SCD, while Delta-LLaVA couples segmentation and question answering via change-enhanced attention (Zhang et al., 2026; Li et al., 2026). CoCD retains this token-conditioned interface but feeds each stage's decoded prediction to the next as explicit input. Unlike ChangeVG's cues, which come from a separate branch at the input, CoCD's evidence is the model's own supervised masks, inserted after observation and comparison.

Multimodal chain-of-thought reasoning. Chain-of-thought prompting elicits textual intermediate steps (Wei et al., 2022). Visual CoT (Shao et al., 2024), MINT-CoT (Chen et al., 2025), and VLM-R^3 (Jiang et al., 2025) incorporate localized regions, selected visual tokens, or cropped image evidence into reasoning. CoCD shares this emphasis on explicit visual evidence but differs in what the evidence is and where it flows: its intermediate steps are decoded task predictions with their own dense supervision rather than regions selected from the input, and they re-enter not only the LLM but also the vision decoder as a dense prompt.

3 Method

3.1 Problem Formulation

Given two images of the same region at an earlier and a later date, denoted I_t with $t \in \{1, 2\}$, and a user instruction, the model predicts a full-scene semantic map $\hat{S}_t$ for each date, a binary change mask $\hat{M}_\Delta$, and a change caption $\hat{C}$, all scoped to the queried classes. The instruction specifies a set of *queried classes* $\mathcal{Y}_q \subseteq \mathcal{Y}_d$, where $\mathcal{Y}_d$ is the land-cover vocabulary of source dataset d and $K = |\mathcal{Y}_q| \geq 2$. Scoping outputs to the queried classes lets the user restrict detection and description to the changes of interest.

Let G_t denote the ground-truth semantic map of date t. For semantic supervision, labels outside $\mathcal{Y}_q$ are mapped to background. The query-scoped binary change target is

$$M_\Delta(x) = \mathbf{1}[G_1(x) \in \mathcal{Y}_q \wedge G_2(x) \in \mathcal{Y}_q \wedge G_1(x) \neq G_2(x)]\,, \tag{1}$$

where $\mathbf{1}[\cdot]$ is the indicator function and x indexes pixels. For example, the query $\{\text{excavation}, \text{low building}\}$ includes transitions in both directions between these two classes and excludes transitions involving any other class. Caption targets describe changes under the same scope. Their construction is detailed in Section 3.2.

3.2 CoCD-Dataset

CoCD requires aligned supervision for semantic observation, change localization, and description. No existing dataset, however, provides full-scene semantic masks at both dates, change masks, and captions for the same image pairs. HRSCD and WUSU provide the first two, so we build CoCD-Dataset on them.

Each patch pair first yields a *full query* over the entire class vocabulary of its source dataset, so that every scene has semantic and change targets for all classes. We then add *subset queries* that name

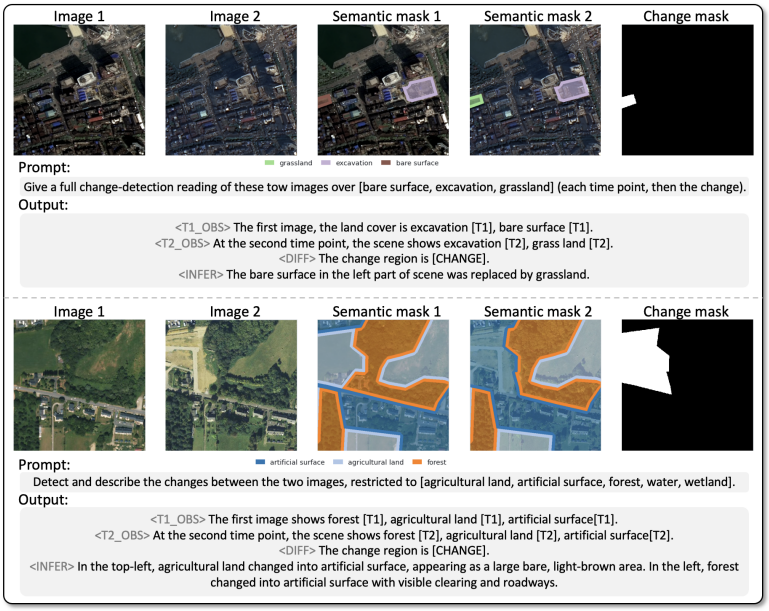

Figure 2: Examples of state-complete, query-scoped traces in CoCD-Dataset, showing the image pair, class-scoped instruction, dense targets, and full structured output. The `[T1]`, `[T2]`, and `[CHANGE]` tokens in the output correspond to the two semantic masks and the change mask, and `<INFER>` contains the grounded caption.

only some classes and derive their targets by Eq. 1. Subset class sets are built from the classes present in the patch, by dropping a changed class, isolating a single transition, dropping unchanged classes, or adding an absent class. CoCD-Dataset contains 11,898/510/1,606 train/validation/test queries. Construction details and statistics are in Appendix A.3.

To generate descriptions, we group in-scope changed pixels by their before–after class pair. After filtering small isolated components, we summarize each transition by its area and location on a 3×3 grid. Qwen3-VL-32B verbalizes these facts, incorporating visual details from the image pair without altering the specified transition types or locations. Each candidate caption is parsed back into transition–location tuples and checked for unsupported transitions, incorrect locations, and omitted major facts. Failed candidates are regenerated, with a deterministic template as fallback. No-change descriptions are sampled from a predefined template pool. Caption-generation prompts, settings, and verification details are provided in Appendix A.4. Figure 2 illustrates the resulting examples, each pairing the input images with per-date semantic states, a binary change mask, and a grounded caption.

3.3 Chain-of-Change Detection (CoCD)

As shown in Figure 3, CoCD connects a large language model with a vision encoder, a vision backbone, and a vision decoder. The LLM emits a `[T1]` and a `[T2]` token for each queried class and a single `[CHANGE]` token for the pair, whose hidden states serve as mask queries for the decoder. The decoder runs once per date for the semantic masks and once for the change mask, with the same architecture but different inputs (Appendix A.2). This token-conditioned interface follows UniChange (Zhang et al., 2026), adapted to Qwen3-VL and SAM2.

CoCD organizes these predictions into an observe–compare–describe sequence: the per-date semantic masks are the outputs of observation, the change mask is the output of comparison, and the final caption describes the predicted changes and transitions. Each stage is directly supervised and receives the decoded prediction of the previous stage as explicit input, closing the shortcuts that bypass earlier stages. The Semantic State Painter (SSP; Section 3.3.1) routes soft semantic predictions from the vision decoder as guidance for change localization, so the change mask is conditioned on what the model observed at each date. The Change Evidence Injector (CEI; Section 3.3.2) returns the predicted change mask and its class transitions to the LLM as evidence for the description stage, so the caption is grounded in where the model localized change and what changed there. Through these

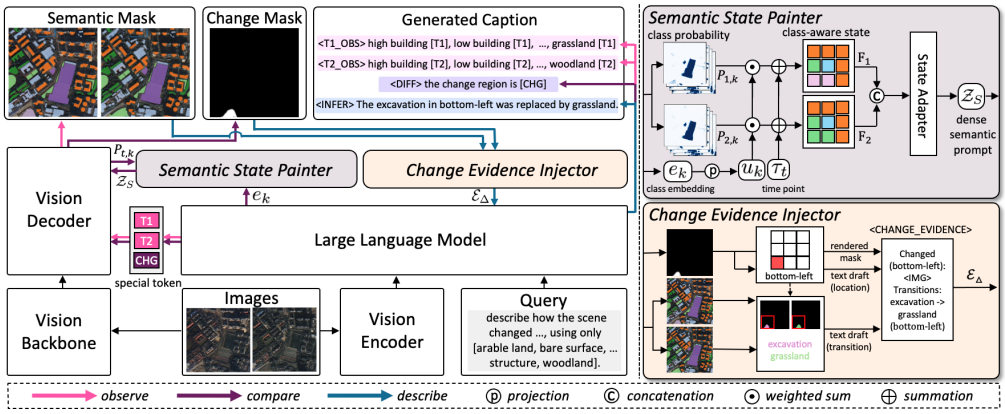

Figure 3: Overview of the CoCD framework. During **Observe**, LLM hidden states at class-specific `[T1]` and `[T2]` tokens condition the vision decoder to produce semantic masks for the queried classes. SSP (Section 3.3.1) combines the decoded class probabilities with LLM class embeddings to form a dense semantic prompt. During **Compare**, the decoder uses this prompt and `[CHANGE]` (CHG in figure) together with bi-temporal visual features to predict a change mask. CEI (Section 3.3.2) then packages the rendered change mask (`<IMG>` in figure), grid-cell location text, and transition text derived from the change mask and semantic states into `<CHANGE_EVIDENCE>`. During **Describe**, the LLM conditions on this visual and textual evidence to generate the final change caption.

two pathways, the CD-specific inductive bias is carried beyond encoder-side feature interaction into the execution sequence of the LLM and its decoders.

3.3.1 Semantic State Painter (SSP)

SSP paints LLM class-name embeddings into spatial maps using the decoder's predicted class probabilities. By deriving class representations from the queried names, the same module can construct semantic priors for different user-specified class sets.

Let $P_{t,k}(x)$ denote the probability of queried class k at pixel x output by the semantic decoder of date t, normalized jointly over the K queried classes and background. The semantic mask $\hat{S}_t$ is its argmax. For class name k tokenized into m_k tokens $w_{k,1}, \ldots, w_{k,m_k}$, let $E_{\text{in}}(\cdot)$ denote the LLM input embedding lookup. The class embedding e_k is the mean of the token embeddings, and u_k is its learned 64-dimensional projection:

$$e_k = \frac{1}{m_k} \sum_{r=1}^{m_k} E_{\text{in}}(w_{k,r}), \qquad u_k = W_e e_k. \tag{2}$$

After resizing the probability maps to 128×128, each pixel becomes a soft class-aware state

$$F_t(x) = \sum_{k=1}^{K} P_{t,k}(x) u_k + \tau_t, \tag{3}$$

where τ_t is a learned timepoint embedding. Within SSP, stop-gradient is applied to $P_{t,k}$ and e_k, while W_e, τ_t, and the state adapter are learned. Weighting by probabilities rather than painting the argmax class preserves the decoder's uncertainty in the state. Appendix B.3 compares the two choices.

The two states are concatenated, resized to the 64×64 decoder grid by $\mathcal{R}_{64}$, and passed through a state adapter to form the dense semantic prompt:

$$\mathcal{Z}_S = f_{\text{adapt}}(\mathcal{R}_{64}([F_1 \| F_2])), \quad f_{\text{adapt}} = \text{Conv}_{3\times3} \to \text{LN2d} \to \text{GELU} \to \text{Conv}_{1\times1}. \tag{4}$$

The adapter maps the 128 concatenated channels to 256 channels, producing the dense semantic prompt $\mathcal{Z}_S$. For vision-backbone features $V_t = f_{\text{backbone}}(I_t)$, the vision decoder combines this prompt with bi-temporal visual evidence:

$$\begin{aligned} B_\Delta &= \text{Fuse}_{\text{vis}}(V_1, V_2) + \mathcal{Z}_S, \\ \text{Fuse}_{\text{vis}}(V_1, V_2) &= \mathcal{C}_1(\mathcal{C}_3(\mathcal{C}_1([V_1 \| V_2 \| |V_1 - V_2|]))), \\ \mathcal{C}_k &= \text{Conv}_{k\times k} \to \text{LN2d} \to \text{GELU}. \end{aligned} \tag{5}$$

The vision decoder then predicts the change mask using the bi-temporal image features, the combined dense prompt B_Δ, and a change query q_Δ derived from the LLM hidden state at `[CHANGE]`:

$$\hat{M}_\Delta = D_{\text{chg}}(V_1, V_2, q_\Delta, B_\Delta). \tag{6}$$

In effect, SSP provides the comparison stage with what the model observed at each date, including the decoder's uncertainty about the observations. Query construction, visual fusion, and decoder details are provided in Appendix A.2.

3.3.2 Change Evidence Injector (CEI)

CEI injects evidence derived from dense predictions directly into the LLM to support natural-language generation, making predicted change regions and semantic transitions explicit context for the final description. Given the predicted change mask and per-date semantic states, it constructs

$$\mathcal{E}_\Delta(\hat{M}_\Delta, \hat{S}_1, \hat{S}_2) = \left[R(\hat{M}_\Delta), f_{\text{cell}}(\hat{M}_\Delta), f_{\text{tr}}(\hat{M}_\Delta, \hat{S}_1, \hat{S}_2)\right]. \tag{7}$$

Here, R renders the change mask as a 256×256 white-on-black image, preserving its spatial layout. f_{cell} names occupied cells on a 3×3 grid, while f_{tr} lists the dominant directed class transitions within the predicted change region. The mask provides spatial detail beyond the coarse grid, while the text makes locations and class transitions explicit.

The evidence is inserted within `<CHANGE_EVIDENCE>` immediately before the final description, so the LLM can ground its caption in the decoded results.

3.3.3 Training and Inference

Using current model predictions as CEI evidence during teacher-forced training would require a preliminary forward pass to decode the semantic and change masks and construct the evidence, followed by a training forward pass with that evidence inserted. To avoid this additional forward pass while exposing the description stage to imperfect model-generated evidence, we reuse decoded predictions from the previous epoch through a per-sample rolling cache:

$$\mathcal{E}_\Delta^{(i,e)} = \mathcal{E}_\Delta\left(\hat{M}_\Delta^{(i,e-1)}, \hat{S}_1^{(i,e-1)}, \hat{S}_2^{(i,e-1)}\right), \tag{8}$$

where i indexes the training sample and e the epoch. Ground truth initializes the cache at epoch 0 and fills cache misses. Appendix B.6 compares this choice with ground-truth evidence throughout training, which likewise requires no additional forward pass.

At inference, the model (1) generates both observation spans and decodes $\hat{S}_1, \hat{S}_2$, (2) forms $\mathcal{Z}_S$ and decodes $\hat{M}_\Delta$, (3) appends $\mathcal{E}_\Delta$, and (4) resumes generation for $\hat{C}$. No external planner or separate captioning model is used.

The objective is

$$\mathcal{L} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{bcd}} + 0.5\mathcal{L}_{\text{inst}} + \mathcal{L}_{\text{pix}}, \tag{9}$$

where $\mathcal{L}_{\text{bcd}}$ is BCE plus Dice for the dedicated change mask, $\mathcal{L}_{\text{inst}}$ is per-class BCE plus Dice for present semantic classes, and $\mathcal{L}_{\text{pix}}$ is a class-balanced softmax loss over queried classes and background. $\mathcal{L}_{\text{text}}$ trains the trace and caption, with inserted evidence interiors label-masked. Caption samples receive weight 1.5, final-caption tokens weight 2.0, and structural tokens weight 0.5.

4 Experiments

4.1 Setup

Training. All models are trained and evaluated on CoCD-Dataset. We leverage Qwen3-VL-4B-Instruct (Bai et al., 2025), together with a SAM2.1 Hiera-L vision backbone and SAM2-family mask decoders (Ravi et al., 2025). We freeze the pretrained Qwen weights and train rank-8 LoRA adapters (Hu et al., 2022) on its language and vision projections, while fully fine-tuning the SAM2.1 vision backbone and mask decoders. We also optimize the newly introduced token and query embeddings, query-projection and visual-fusion modules, and the learnable components of SSP. Models are trained for 10 epochs using 4 A6000 GPUs with an effective batch size of 32. Additional optimization details are provided in Appendix A.1.

Table 1: Comparison with general segmentation MLLMs adapted to two-frame CD ($*$) and prior CD MLLMs on CoCD-Dataset. UniChange-style (†) is our adapted Qwen3-VL/SAM2 implementation of UniChange.

Group	Method [components]	oIoU	F1	sem-mIoU	κ	BLEU-4	CIDEr	MMC	CMC
General seg-MLLM	LISA* (Lai et al., 2024) [LLaVA-1.5-7B + SAM ViT-H]	16.07	27.69	42.16	50.23	24.49	154.40	30.60	47.19
	Sa2VA* (Yuan et al., 2025) [Qwen3-VL-4B + SAM2]	12.78	22.67	43.30	52.34	27.10	155.94	27.41	39.23
	PixelLM* (Ren et al., 2024) [LLaVA-1.5-7B + native decoder]	24.08	38.82	46.69	56.93	24.12	141.01	40.54	49.86
CD MLLM	ChangeVG (Xue et al., 2026) [Qwen3-VL-4B]	–	–	–	–	22.30	114.18	–	–
	DeltaVLM (Deng et al., 2026) [InstructBLIP/Vicuna-7B]	–	–	–	–	17.81	73.43	–	–
	UniChange-style† (Zhang et al., 2026) [Qwen3-VL-4B + SAM2]	25.00	40.00	45.78	59.75	21.98	131.57	37.98	35.18
Ours	**CoCD [Qwen3-VL-4B + SAM2]**	**31.33**	**47.71**	**52.61**	**64.02**	**30.00**	**170.26**	**55.77**	**66.71**

Table 2: Separating the effects of task chaining and decoded-state feedback. *Trace* denotes sequential task execution. *Enc. diff* denotes DeltaVLM's encoder-side difference-perception module. *State re-entry* denotes feedback of intermediate decoded predictions through SSP and CEI.

Method	Trace	Enc. diff	State re-entry	oIoU	F1	sem-mIoU	κ	BLEU-4	CIDEr	MMC	CMC
Parallel				25.00	40.00	45.78	59.75	21.98	131.57	37.98	35.18
Naive-Chain	✓			25.59	40.75	51.43	62.61	21.96	125.14	48.74	45.11
Naive-Chain + encoder diff	✓	✓		25.67	40.85	50.87	61.94	23.14	136.83	48.88	43.87
CoCD	✓		✓	**31.33**	**47.71**	**52.61**	**64.02**	**30.00**	**170.26**	**55.77**	**66.71**

Metrics. All outputs are evaluated under the class scope specified by each query. We evaluate BCD using overall IoU (oIoU) and F1 of the predicted change mask. For per-date semantic segmentation, we report foreground-class mIoU and Cohen's κ (Cohen, 1960). We do not report a separate SCD score, since an SCD map follows from pairing the two semantic maps within the predicted change region. Caption quality is evaluated using BLEU-4 (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015), with the latter computed using the CIDEr-D variant. Because these n-gram metrics also reward matching the reference phrasing, Appendix B.1 additionally scores the parsed transitions and locations of each caption against ground truth. We also report two consistency scores: *Mask–Mask Consistency (MMC)*, the agreement between the predicted change mask and the change implied by the two semantic maps, and *Caption–Mask Consistency (CMC)*, the agreement between the caption's stated transitions and those found inside the predicted change mask. Both measure internal agreement rather than correctness against ground truth and are defined in Appendix A.5.

4.2 Comparison with Unified Baselines

Table 1 compares CoCD with two groups of unified baselines: general segmentation MLLMs adapted to two-frame CD, and prior CD MLLMs. We adapt the former to two-frame input with `[T1]`, `[T2]`, and `[CHANGE]` roles while retaining their native mask decoders. ChangeVG and DeltaVLM are text-only in this comparison, so dense metrics, MMC, and CMC are unavailable. All baselines are trained on CoCD-Dataset, with adaptation details in Appendix A.6. The UniChange-style row is our matched Qwen3-VL/SAM2 reimplementation, not the released checkpoint (Appendix B.5).

CoCD achieves the best localization, semantic segmentation, caption quality, and consistency scores in Table 1. Among the baselines, strength in one output does not carry over to the others: Sa2VA and LISA lead the adapted segmentation MLLMs in caption quality but trail in localization, PixelLM shows the reverse, and the UniChange-style baseline scores below 40 on both consistency measures. Over this matched baseline, CoCD raises oIoU from 25.00 to 31.33 and BLEU-4 from 21.98 to 30.00, and improves every other metric as well. The following controls examine whether these gains stem from task sequencing or from decoded-state feedback.

4.3 Effects of Task Chaining and State Feedback

Table 2 compares four variants with the same MLLM, vision backbone, decoder family, data, and training setup. Parallel predicts each task from an independent prompt. Naive-Chain places observation, comparison, and description in a single sequence, allowing later stages to attend to earlier tokens. We augment Naive-Chain with DeltaVLM's difference-perception module (Deng et al., 2026) to provide difference-aware vision-encoder features to the LLM. CoCD instead connects the stages through SSP and CEI.

Table 3: Inference-time state interventions (GT: ground truth). Interventions occur after semantic prediction, so semantic mIoU is invariant.

Evidence at inference	oIoU	sem-mIoU	BLEU-4	CIDEr
GT, both	**73.76**	52.61	**38.43**	**248.6**
GT semantic only	**73.76**	52.61	37.59	238.3
GT change only	31.33	52.61	31.55	197.8
Predicted (CoCD)	31.33	52.61	30.00	170.3
State route off	27.58	52.61	29.62	166.4
Blank change	31.33	52.61	16.11	104.6
Blank semantic	2.47	52.61	20.60	121.0
Blank, both	2.47	52.61	16.11	104.6

Table 4: Joint SSP–CEI ablation. Spatial: mask rendering and location text; Full: spatial evidence plus transition text. None without SSP is Naive-Chain, and Full with SSP is CoCD.

CEI evidence	SSP	oIoU	F1	sem-mIoU	κ	BLEU-4	CIDEr	MMC	CMC
None	–	25.59	40.75	51.43	62.61	21.96	125.14	48.74	45.11
None	✓	26.48	41.87	50.04	61.90	22.32	130.28	54.40	43.75
Spatial	–	26.82	42.29	51.13	62.27	22.88	131.89	47.52	58.96
Spatial	✓	28.27	44.08	51.04	62.40	24.22	142.12	54.33	60.52
Full	–	27.37	42.98	50.50	62.41	27.79	148.53	48.46	**67.27**
Full (**CoCD**)	✓	**31.33**	**47.71**	**52.61**	**64.02**	**30.00**	**170.26**	**55.77**	66.71

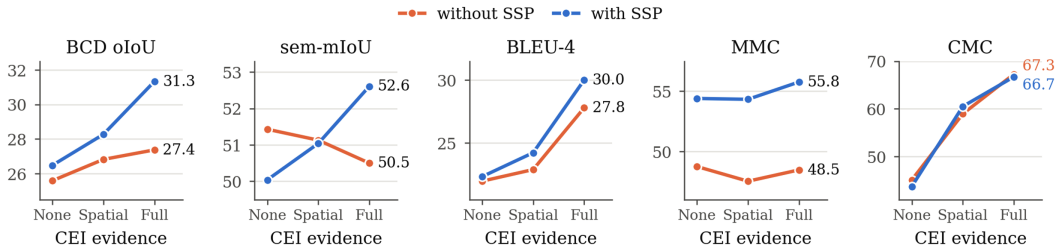

Figure 4: SSP–CEI interaction across five metrics from Table 4. Each panel compares performance with and without SSP as caption evidence changes from None to Spatial to Full.

Task chaining improves per-date semantics and both consistency scores over Parallel, but leaves localization (25.00 to 25.59 oIoU) and caption quality (21.98 to 21.96 BLEU-4) essentially unchanged. Encoder-side interaction raises BLEU-4 modestly to 23.14 while leaving localization unchanged and slightly reducing semantic scores and CMC. CoCD improves every metric over both chain controls, reaching 31.33 oIoU and 30.00 BLEU-4. Sequencing alone thus improves cross-output consistency but provides little improvement in change localization or caption quality. Explicit decoded-state feedback improves both, as the shortcut hypothesis predicts. The gains hold across three MLLM backbones (Appendix B.4).

4.4 Do Downstream Tasks Use the Routed States?

Table 3 reports interventions on one trained CoCD checkpoint at inference, with weights and generation settings fixed. *State route off* bypasses only the state adapter, so $\mathcal{Z}_S$ is not added to the visual prompt. A blank semantic state is instead an all-background counterfactual. Ground-truth replacement is an oracle diagnostic, not deployable performance.

Bypassing SSP lowers oIoU from 31.33 to 27.58, and replacing its semantic guidance with an all-background state collapses it to 2.47, whereas ground-truth guidance raises it to 73.76. On the caption side, blanking the change evidence drops BLEU-4 from 30.00 to 16.11, while ground-truth evidence lifts it to 38.43. Both downstream stages therefore depend on the routed evidence and track its quality rather than bypassing it, which is the dependency the observe–compare–describe sequence is meant to enforce.

4.5 Complementary Roles of SSP and CEI

Table 4 and Figure 4 cross SSP with three CEI settings: None, Spatial (mask rendering and location text), and Full (adding transition text). Each configuration is trained separately with the same task chain and training recipe.

With SSP enabled, progressively adding spatial evidence and transition text improves localization, semantic segmentation, caption quality, and CMC. Without SSP, richer CEI evidence still improves captioning but slightly reduces semantic mIoU (51.43 to 50.50). The two routes also interact: the oIoU gain from SSP grows from 0.89 with no CEI evidence to 1.45 with spatial evidence and 3.96 with full evidence. Full CoCD is the strongest configuration on every task metric. CMC alone is marginally higher without SSP (67.27 vs. 66.71), which we return to below.

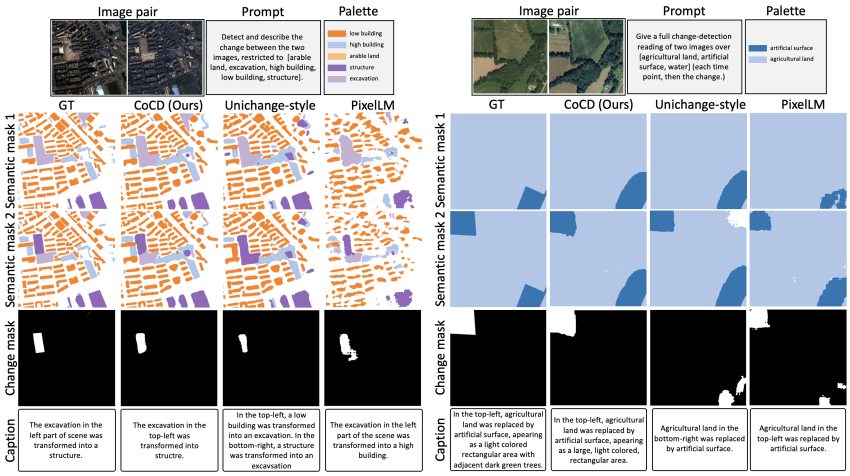

Figure 5: Selected qualitative comparisons with PixelLM* and the UniChange-style parallel baseline on WUSU (left) and HRSCD (right). Each example shows the input image pair, per-date semantic predictions, binary change masks, and generated captions, alongside ground-truth annotations. For CoCD, only the final `<INFER>` caption is shown.

The consistency scores separate their roles. MMC moves with SSP (48.74 to 54.40 without CEI evidence) and barely with CEI, while CMC moves with CEI evidence (45.11 to 58.96 to 67.27 without SSP) and barely with SSP. The two routes thus strengthen complementary forms of agreement, between semantic and change masks and between captions and mask-derived evidence, matching the roles assigned to SSP and CEI in Section 3.

4.6 Qualitative Results

Figure 5 shows one example from each source dataset. **Left (WUSU):** the UniChange-style baseline localizes the excavation-to-structure change and predicts it in its semantic maps, yet its caption describes a low-building-to-excavation transition, a caption–mask inconsistency. **Right (HRSCD):** the same baseline predicts the agricultural-land-to-artificial-surface transition in the top-left through its semantic maps but places the change mask in the bottom-right, where its semantic maps are unchanged, a mask–mask inconsistency. PixelLM correctly captions the target transition but overpredicts the change region. In both cases, CoCD produces a caption, change mask, and semantic maps that agree on the same transition, addressing the two failure modes illustrated in Figure 1(d). More examples are in Appendix D.

5 Conclusion

We presented CoCD, a supervised observe–compare–describe sequence spanning the LLM and dense decoders, in which each stage receives the decoded prediction of the previous one. This carries the CD-specific inductive bias beyond encoder-side feature interaction into the execution sequence itself and closes the shortcuts that leave masks and captions free to disagree. Supported by the aligned supervision in CoCD-Dataset, matched experiments show gains beyond task chaining alone, while inference interventions demonstrate that downstream predictions depend on the routed evidence. These results support explicit, supervised execution for more accurate and consistent change understanding.

Limitations. Training requires per-date semantic labels, which restricts supervision to two in-domain datasets with closed vocabularies and leaves broader generalization to be established. Upstream prediction errors can propagate to later stages, and caption evaluation is limited to controlled language and a single reference per sample. Future work will broaden supervision and evaluation settings while improving robustness to intermediate prediction errors.

Ethics Statement

This work studies land-cover change detection and description using existing remote-sensing datasets. Although intended to support applications such as environmental monitoring, the same capabilities could be misused for intrusive surveillance or unsupported claims about activities in sensitive locations. Predicted masks and captions may contain errors or reflect biases in the source imagery, class taxonomies, and automatically generated supervision. In particular, agreement between a caption and predicted masks does not guarantee factual correctness. Outputs should therefore be independently validated and subject to human review before use in consequential decisions. Any release of derived annotations or other third-party assets will be subject to the applicable dataset and model licenses and redistribution restrictions.

Reproducibility Statement

The main text and Appendix document the model architecture, training objective, data construction, and evaluation protocols. Appendix A.1 specifies optimization settings and training resources, Appendix A.2 details mask decoding and the integration of SSP and CEI, Appendix A.3 describes how CoCD-Dataset is built from HRSCD and WUSU, Appendix A.4 gives the caption-generation prompt and verification procedure, Appendix A.5 defines the consistency metrics, and Appendix A.6 lists all baseline adaptations. Anonymized code for data preparation, training, inference, and evaluation is available at `https://anonymous.4open.science/r/cocd_release-3CE3/`. The data-preparation scripts rebuild CoCD-Dataset from the public HRSCD and WUSU releases, and we will additionally release the constructed dataset and trained weights upon publication, subject to the source datasets' licenses and redistribution restrictions.

AI Use Statement

Generative AI tools were used to assist with related-work discovery and retrieval, brainstorming, experimental design feedback, code editing, and manuscript drafting and revision. We also used Qwen3-VL-32B to generate synthetic change captions for CoCD-Dataset from ground-truth-derived semantic transitions and spatial information. The caption-generation and verification procedures are described in Section 3.2 and Appendix A.4. The authors reviewed and revised the AI-assisted research and writing, and generated captions were checked against the mask-derived facts using the reported verification procedure.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Wele Gedara Chaminda Bandara and Vishal M. Patel. A transformer-based siamese network for change detection. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 207–210, 2022. doi: 10.1109/IGARSS46834.2022.9883686.

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.

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.

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. doi: 10.1109/TGRS.2024.3417253.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. MINT-CoT: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2324.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. LMSYS Org blog, 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *IEEE International Conference on Image Processing (ICIP)*, pp. 4063–4067, 2018. doi: 10.1109/ICIP.2018.8451652.

Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187: 102783, 2019. doi: 10.1016/j.cviu.2019.07.003.

Pei Deng, Wenqian Zhou, and Hanlin Wu. DeltaVLM: Interactive remote sensing image change analysis via instruction-guided difference perception. *Remote Sensing*, 18(4):541, 2026. doi: 10.3390/rs18040541.

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.

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.

Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xBD: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2019. doi: 10.1109/TGRS.2018.2858817.

Chaoya Jiang, Yongrui Heng, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. VLM-R^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2143.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024. doi: 10.1109/CVPR52733.2024.00915.

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. *arXiv preprint arXiv:2604.14044*, 2026.

Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large-scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022. doi: 10.1109/TGRS.2022.3218921.

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. doi: 10.1109/TGRS.2023.3321752.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024. doi: 10.1109/CVPR52733.2024.02484.

Xuanguang Liu, Chenguang Dai, Lei Ding, Zhenchao Zhang, Yujie Li, Xibing Zuo, Mengmeng Li, Hanyun Wang, and Yuzhe Miao. GSTM-SCD: Graph-enhanced spatio-temporal state space model for semantic change detection in multi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 230:73–91, 2025. doi: 10.1016/j.isprsjprs.2025.09.003.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Mubashir Noman, Noor Ahsan, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. CDChat: A large multimodal model for remote sensing change description. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 7033–7037, 2025. doi: 10.1109/IGARSS55030.2025.11243118.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16, 2020. doi: 10.1109/SC41405.2020.00024.

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*, 2025.

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*, pp. 26374–26383, 2024. doi: 10.1109/CVPR52733.2024.02491.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8612–8642, 2024. doi: 10.52202/079017-0275.

Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. S2Looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24):5094, 2021. doi: 10.3390/rs13245094.

Sunan Shi, Yanfei Zhong, Yinhe Liu, Jue Wang, Yuting Wan, Ji Zhao, Pengyuan Lv, Liangpei Zhang, and Deren Li. Multi-temporal urban semantic understanding based on GF-2 remote sensing imagery: From tri-temporal datasets to multi-task mapping. *International Journal of Digital Earth*, 16(1):3321–3347, 2023. doi: 10.1080/17538947.2023.2246445.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Junxiao Xue, Quan Deng, Xuecheng Wu, Kelu Yao, Xinyi Yin, Fei Yu, Wei Zhou, Yanfei Zhong, Yang Liu, and Dingkang Yang. Towards comprehensive interactive change understanding in remote sensing: A large-scale dataset and dual-granularity enhanced VLM. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–16, 2026. doi: 10.1109/TGRS.2025.3650151.

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.

Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying SAM2 with LLaVA for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

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*, pp. 42169–42179, 2026.

Zhuo Zheng, Yanfei Zhong, Shiqi Tian, Ailong Ma, and Liangpei Zhang. ChangeMask: Deep multi-task encoder–transformer–decoder architecture for semantic change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183:228–239, 2022. doi: 10.1016/j.isprsjprs.2021.10.015.

Duowang Zhu, Xiaohu Huang, Haiyan Huang, Hao Zhou, and Zhenfeng Shao. Change3D: Revisiting change detection and captioning from a video modeling perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24011–24022, 2025a. doi: 10.1109/CVPR52734.2025.02236.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025b.

SUPPLEMENTARY MATERIAL

CONTENTS

A Additional Experimental Details **15**

A.1 Implementation 15

A.2 Vision Decoder Details 15

A.3 CoCD-Dataset Construction and Statistics 17

A.4 Caption Generation and Verification 18

A.5 Cross-Output Consistency Protocol 19

A.6 Baseline Adaptations 20

B Supporting Results **21**

B.1 Caption Content Accuracy against Ground Truth 21

B.2 Dedicated Change Decoder versus Semantic XOR 21

B.3 Soft versus Hard Semantic States in SSP 22

B.4 MLLM Backbone Dependence 22

B.5 Relation to UniChange 22

B.6 Ground-Truth versus Previous-Epoch Evidence 23

C Additional Dataset Visualizations **24**

D Additional Qualitative Results **25**

A Additional Experimental Details

A.1 Implementation

Our main implementation uses Qwen3-VL-4B-Instruct (Bai et al., 2025) and SAM2.1 (Ravi et al., 2025), with LoRA (Hu et al., 2022) for language and vision adaptation. We train for 10 epochs and select the final checkpoint for every adopted arm. Per-device batch size is 1 on four GPUs with gradient accumulation 8, yielding an effective batch size of 32. Training uses BF16, gradient checkpointing, DeepSpeed ZeRO-2 (Rajbhandari et al., 2020), and maximum gradient norm 1.0. AdamW (Loshchilov & Hutter, 2019) uses weight decay 0.01, betas $(0.9, 0.999)$, cosine decay, and 5% warmup. Qwen language LoRA has rank 8, scale 16, dropout 0.05, and learning rate 5×10^{-5} on query/key/value/output projections. Qwen vision-encoder LoRA uses the same rank, scale, and dropout with learning rate 2×10^{-5} on QKV and output projections. New heads, adapters, and embedding rows use 10^{-4}, while pretrained decoder modules use 2×10^{-5}.

Qwen inputs are resized to 512 pixels, SAM inputs to 1024, and dense targets to 512. Change-mask evidence is rendered at 256. The maximum sequence length is 4096, and evaluation batch size is 1. Geometric augmentation is applied with probability 0.5. When a transform invalidates the original directional caption coordinates, direction-bearing language labels are masked.

A.2 Vision Decoder Details

This subsection documents the decoder implementation shared by Parallel, Naive-Chain, and CoCD. It adapts UniChange's token-conditioned mask interface (Zhang et al., 2026) to Qwen3-VL and SAM2.1, with the aim of obtaining a comparable baseline rather than introducing a new decoder architecture. Appendix B.5 summarizes the differences from UniChange and compares performance. The three configurations differ only in the dense prompt for change decoding: CoCD adds SSP's semantic prompt (Section 3.3.1) to the visual prompt, whereas Parallel and Naive-Chain use the visual prompt alone.

The following description uses the notation of Section 3. Decoding roles are indexed by $r \in \{1, 2, \Delta\}$, where the semantic roles $r = 1, 2$ coincide with the dates t. Batch dimensions are omitted, $\|$ denotes channel or vector concatenation, and $\mathcal{R}_s$ denotes spatial interpolation to $s \times s$. For reference, V denotes vision-backbone features, $\hat{S}_t$ the predicted per-date semantic maps, F_t SSP's per-date soft class-aware states, $\mathcal{Z}_S$ its dense semantic prompt, and B_Δ the combined dense prompt.

Multi-scale vision-backbone features. The two images, resized to 1024×1024, are processed by the shared SAM2.1 Hiera-L vision backbone (Ravi et al., 2025). The feature pyramid supplied to the decoder is

$$(V_t^{64}, V_t^{128}, V_t^{256}) = f_{\text{backbone}}(I_t), \qquad V_t^s \in \mathbb{R}^{256 \times s \times s}. \tag{10}$$

The main feature V_t^{64}, denoted by V_t in the main text, enters mask-decoder attention. The high-resolution features V_t^{128} and V_t^{256} restore spatial detail during upsampling.

LLM-to-mask queries. A role-specific MLP constructs each mask query from three inputs: the final-layer hidden state $h_{r,k}$ of the corresponding `[T1]`, `[T2]`, or `[CHANGE]` token, a stage-anchor hidden state a_r, and a class context $c_{r,k}$, where $k \in \{1, \ldots, K\}$ indexes the queried class for the semantic roles (the change role has a single query, so k is dropped). Each query is formed as

$$\widetilde{q}_{r,k} = \text{MLP}_r([h_{r,k} \| a_r \| c_{r,k}]), \tag{11}$$

$$q_{r,k} = \text{Proj}_{\text{sam}}(\widetilde{q}_{r,k}). \tag{12}$$

The stage anchors a_1, a_2, and a_Δ are the final-layer hidden states at the span-opening tokens `<T1_OBS>`, `<T2_OBS>`, and `<DIFF>` (Figure 2). For the semantic roles, the class context is the class embedding of Eq. 2, $c_{r,k} = e_k$. For the change role, c_Δ is the mean of e_k over the queried classes. For Qwen3-VL-4B, each role-specific MLP has dimensions $3 \cdot 2560 \to 2560 \to 2560$, and the shared SAM projection has dimensions $2560 \to 2560 \to 256$. Both use GELU between their linear layers. The mask queries form one sequence per decoding role:

$$Q_t = [q_{t,0}, q_{t,1}, \ldots, q_{t,K}], \quad t \in \{1, 2\}, \qquad Q_\Delta = [q_\Delta], \tag{13}$$

where $q_{t,0}$ is a separate learned background query and $q_{t,1}, \ldots, q_{t,K}$ are the foreground queries for the K queried classes.

Visual comparison and the base dense prompt. Using the convolution blocks $\mathcal{C}_k$ defined in Eq. 5 and writing $V_r = V_r^{64}$ for every role r, the fused main feature and visual dense prompt are

$$V_\Delta = \mathcal{C}_3(\mathcal{C}_1([V_1 \| V_2 \| |V_1 - V_2|])), \tag{14}$$

$$B_{\text{vis}} = \mathcal{C}_1(V_\Delta) = \text{Fuse}_{\text{vis}}(V_1, V_2). \tag{15}$$

Each block has independent parameters and 256 output channels, yielding $V_\Delta, B_{\text{vis}} \in \mathbb{R}^{256\times64\times64}$. The high-resolution branches produce V_Δ^{128} and V_Δ^{256} by fusing each feature pair separately, without an absolute-difference input:

$$V_\Delta^s = \mathcal{C}_3(\mathcal{C}_1([V_1^s \| V_2^s])), \qquad s \in \{128, 256\}. \tag{16}$$

Each high-resolution branch has its own parameters and outputs 256 channels. The fused features V_Δ^s are used during change-decoder upsampling, while the date-specific features V_1^s and V_2^s are used during semantic-decoder upsampling for their respective dates. The shared base dense prompt B_{vis} provides both semantic and change prediction with information from the two dates and their absolute feature difference $|V_1 - V_2|$.

Shared spatial decoding procedure. For each role r, we write the spatial decoding procedure as

$$\ell_r = \mathcal{D}_r\big(V_r^{64}, B_r, (V_r^{128}, V_r^{256}), Q_r\big). \tag{17}$$

The operator $\mathcal{D}_r$ comprises two-way attention, upsampling with high-resolution features, and a mask head. SAM2's dense-prompt interface adds the prompt to the main image embedding before two-way attention:

$$(Q_r', X_r') = \text{TwoWayTransformer}(Q_r, V_r^{64} + B_r, \text{PE}), \tag{18}$$

where PE denotes spatial positional encoding, and Q_r' and X_r' are the updated queries and spatial features. Two transposed convolutions then upsample the spatial features, adding the corresponding high-resolution features through channel-alignment 1×1 convolutions C_s:

$$U_r^{128} = \text{GELU}\big(\text{LN}\big(\text{Deconv}_1(X_r') + C_{128}(V_r^{128})\big)\big), \tag{19}$$

$$U_r^{256} = \text{GELU}\big(\text{Deconv}_2(U_r^{128}) + C_{256}(V_r^{256})\big). \tag{20}$$

The channel–resolution sequence is $256 \times 64 \times 64 \to 64 \times 128 \times 128 \to 32 \times 256 \times 256$. A role-specific hypernetwork maps each updated query $q_{r,k}'$ in Q_r' to a 32-dimensional mask-weight vector $\omega_{r,k}$:

$$\omega_{r,k} = \text{HyperMLP}_r(q_{r,k}'), \qquad \ell_{r,k} = \mathcal{R}_{512}\big(\omega_{r,k}^\top U_r^{256}\big). \tag{21}$$

The transformer and upsampling modules are shared across the three roles, while HyperMLP_r is separate for T_1, T_2, and change. Thus $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_\Delta$ share one decoding structure and differ in their inputs and mask heads, as summarized in Table 5.

Semantic decoding. For $t \in \{1, 2\}$, we apply Eq. 17 to date-specific features, the shared visual prompt $B_t = B_{\text{vis}}$, and the semantic query sequence Q_t:

$$\ell_t = \mathcal{D}_t\big(V_t^{64}, B_{\text{vis}}, (V_t^{128}, V_t^{256}), Q_t\big). \tag{22}$$

The shared prompt provides auxiliary visual context from the other date to help resolve ambiguities in per-date classification, while date-specific features keep each prediction grounded in its own observation. Class probabilities are normalized jointly over foreground and background,

$$P_{t,k}(x) = \frac{\exp(\ell_{t,k}(x))}{\sum_{j=0}^{K} \exp(\ell_{t,j}(x))}, \qquad k \in \{0, \ldots, K\}, \tag{23}$$

and the predicted semantic map is

$$\hat{S}_t(x) = \underset{k\in\{0,\ldots,K\}}{\arg\max}\ P_{t,k}(x), \tag{24}$$

where $k = 0$ represents pixels outside the queried classes and $k = 1, \ldots, K$ index the queried classes.

SSP integration. At inference, each date's semantic masks are decoded when its `[T1]` or `[T2]` mask-query tokens are emitted, and change decoding follows both semantic decodes in training and inference. For CoCD, SSP converts the current semantic probabilities into the soft class-aware states F_t of Eq. 3. In this implementation, $W_e : \mathbb{R}^{2560} \to \mathbb{R}^{64}$ and $F_t \in \mathbb{R}^{64\times128\times128}$. The time embedding $\tau_t \in \mathbb{R}^{64}$ is broadcast spatially, and background has no class vector in the weighted sum. As specified in Section 3.3.1, $P_{t,k}$ and e_k are stop-gradient inputs, while W_e, τ_t, and the state adapter are trainable. The two states are concatenated into 128 channels and resized to the decoder grid before the adapter:

$$\mathcal{Z}_S = f_{\text{adapt}}(\mathcal{R}_{64}([F_1 \| F_2])), \qquad B_\Delta = B_{\text{vis}} + \mathcal{Z}_S. \tag{25}$$

The adapter uses a 3×3 convolution from 128 to 256 channels, LayerNorm2d, GELU, and a 1×1 convolution from 256 to 256 channels. The final convolution is zero-initialized, so $\mathcal{Z}_S$ is initially zero and the model starts from the Naive-Chain behavior.

Change decoding. The same procedure receives the fused bi-temporal features, the combined prompt B_Δ, and the single-query sequence $Q_\Delta = [q_\Delta]$. Its single-channel output ℓ_Δ is converted to the binary change mask by sigmoid and thresholding:

$$\ell_\Delta = \mathcal{D}_\Delta\big(V_\Delta^{64}, B_\Delta, (V_\Delta^{128}, V_\Delta^{256}), Q_\Delta\big), \tag{26}$$

$$\hat{M}_\Delta(x) = \mathbf{1}[\sigma(\ell_\Delta(x)) \geq \eta_\Delta], \tag{27}$$

where σ is the sigmoid function and $\eta_\Delta = 0.5$ is the change-probability threshold. Parallel and Naive-Chain use the same procedure with $B_\Delta = B_{\text{vis}}$. In the main text, D_{chg} denotes the full process from bi-temporal feature fusion to the final change mask rather than $\mathcal{D}_\Delta$ alone. CEI then constructs visual and textual evidence for caption generation from $\hat{M}_\Delta$, $\hat{S}_1$, and $\hat{S}_2$.

Table 5: Inputs to the two semantic decoding calls and the change decoding call in CoCD. Parallel and Naive-Chain use $B_\Delta = B_{\text{vis}}$.

Decode	Main feature V_r^{64}	Dense prompt B_r	High-resolution features
T_1 semantic	V_1^{64}	B_{vis}	V_1^{128}, V_1^{256}
T_2 semantic	V_2^{64}	B_{vis}	V_2^{128}, V_2^{256}
Change	V_Δ^{64}	$B_\Delta = B_{\text{vis}} + \mathcal{Z}_S$	$V_\Delta^{128}, V_\Delta^{256}$

A.3 CoCD-Dataset Construction and Statistics

Source datasets. HRSCD (Daudt et al., 2019) provides 291 pairs of 0.5 m aerial orthoimages (10,000×10,000 pixels) over two French departments from 2005/2006 and 2012, with Urban Atlas land-cover maps at both dates over five classes: artificial surface, agricultural land, forest, wetland, and water. WUSU (Shi et al., 2023) provides 1 m GF-2 imagery of two Wuhan districts from 2015, 2016, and 2018, with land-cover maps at every date over eleven occurring classes: arable land, bare surface, excavation, grassland, high building, lake, low building, river, road, structure, and woodland.

Preprocessing. HRSCD pairs are cut into non-overlapping 512×512 patches. WUSU is distributed as 512×512 tiles, from which we form the 2015–2016 and 2016–2018 pairs. Each split is sampled so that about 25% of its patch pairs are no-change, giving 3,000/70/205 HRSCD and 1,010/63/152 WUSU patch pairs for train/validation/test. Splits are disjoint at the tile level: HRSCD is split into northing bands (60/15/25%) with a one-tile buffer between bands, and WUSU follows its official train/test split with a held-out validation subset. Class names are normalized to the lowercase phrases above, which are also used in instructions, captions, and SSP class embeddings.

Query construction. Each patch pair yields one *full query*, whose class set is the entire class vocabulary of its source dataset, and several *subset queries*, whose class sets are proper subsets of it. We generate one subset query per HRSCD pair and up to five per WUSU pair, since WUSU has more classes to combine but fewer patch pairs. At test time we generate three per HRSCD pair and subsample the WUSU subset queries so that both sources contribute equally many test queries. The class set of a subset query is derived from the classes present in the patch rather than sampled uniformly: we drop one changed class, keep only the two classes of a single transition, drop

some unchanged classes, or add one absent class, with probabilities 0.30/0.30/0.25/0.15, and reject duplicates and class sets with no present class at either date. For each query, classes outside its class set are mapped to background to form the semantic targets, the change target follows Eq. 1, and captions follow Appendix A.4. In total we obtain 14,014 queries (11,898/510/1,606).

Composition. Figure 6(a) shows the number of classes per query in each split. Two- and three-class queries dominate, five-class queries form a second mode from the HRSCD full query, and queries with more than five classes come only from WUSU. No-change queries make up 28.2%, 26.7%, and 26.7% of the three splits, slightly above the 25% of patch pairs because no-change pairs receive more subset queries on average. Figure 6(b) shows the frequency of directed class transitions. HRSCD has 11 transition types dominated by agricultural land to artificial surface, whereas WUSU has 59 types distributed across its more fine-grained land-cover classes, with excavation and low building as the most frequent source classes.

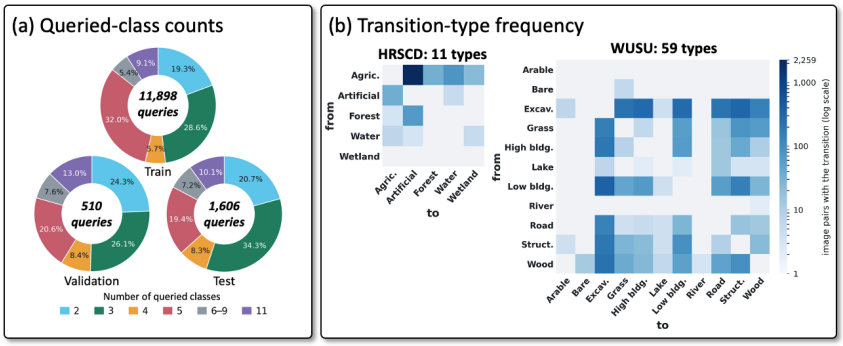

Figure 6: CoCD-Dataset statistics. **(a)** Number of queried classes per query in each split, with the number of queries at the center of each chart. **(b)** Frequency of directed class transitions across all image pairs, shown separately for HRSCD (left) and WUSU (right) on a log scale. The count above each matrix is the number of distinct transition types that occur at least once.

A.4 Caption Generation and Verification

As outlined in Section 3.2, reference captions for CoCD-Dataset are generated from the ground-truth transitions of each query. For each image pair and query, we retain transitions whose source and destination classes both belong to the queried class set and summarize each as a structured record of the before–after class pair and its location on a 3×3 grid. These records are the *fact tuples* supplied to the caption generator. Qwen3-VL-32B-Instruct (Bai et al., 2025) receives the fact tuples as text together with the two original images. It turns the supplied transitions into a caption. The images may contribute short visual details but may not introduce additional transitions or alter the specified locations. The semantic masks themselves are not image inputs to the generator. The same generation prompt is used for every query, with the fact tuples restricted to that query's class set.

Table 6: No-change caption template pool.

The scene remains unchanged. / No change is observed between the two images. / The land cover is the same in both images. / No land-cover change occurred between the two time points. / The two images show no change. / There is no detectable change between the earlier and later image. / The scene is unchanged across the two time points. / No change is present between the two images. / The land cover did not change between the two images. / Both time points show the same land cover. / Nothing changed between the earlier and later image. / The scene shows no land-cover transition. / No transition is observed between the two time points. / The earlier and later images depict the same land cover. / No changed regions are found between the two images. / The land cover remained stable across the two time points. / The scene stayed the same between the two images. / There is no land-cover change to report.

Generation uses temperature 0.7 and at most 200 output tokens. We extract the transitions and locations mentioned in each candidate caption and compare them with the supplied records. Captions with unsupported transitions, incorrect locations, or missing major transitions are rejected and regenerated, with at most three attempts in total. A scene is labelled no-change when no transition

among the queried classes remains after removing connected components smaller than 64 px. Its caption is then sampled uniformly from a fixed pool of 18 no-change templates (Table 6).

The generation prompt is reproduced below with the location and verb vocabularies written out in full. {FACT_TUPLES_JSON} denotes the supplied facts, while <IMAGE_T1> and <IMAGE_T2> mark the two source-image inputs.

```
System Prompt
You write ONE change caption for a pair of co-registered satellite
images (first=earlier, second=later). You are given the ground-truth
land-cover TRANSITIONS as fact tuples. Turn them into fluent
sentence(s) under a strict contract.

CONTRACT:
1. For every MAJOR tuple, state its transition using EXACTLY the
   given 'from', 'to' class words, the assigned 'verb', and (when
   present) the 'location' phrase. Do not substitute a synonym for
   any of these ('red-roofed buildings' is allowed, 'red-roofed
   complexes' is NOT). Minor tuples may be omitted.
2. Use the images ONLY to add a short visual modifier (colour,
   shape, arrangement, relative size) to an object the tuple already
   names. NEVER guess function or use ('residential', 'newly
   constructed', 'industrial').
3. Add a modifier only if you are certain; otherwise omit it.
4. One to two facts per sentence. No numbers, areas, or percentages.
5. Output ONLY the caption sentence(s). No tags, no special tokens
   ([T1]/[CHANGE]/<...>), no markdown, no lists.

LOCATION vocabulary (use these words verbatim): top-left, top,
top-right, left, center, right, bottom-left, bottom, bottom-right,
throughout the scene.
VERB vocabulary: replaced by, converted to, transformed into,
gave way to, changed into.

User Input Template
fact tuples: {FACT_TUPLES_JSON}
Describe the whole scene.
First (earlier) image:
<IMAGE_T1>
Second (later) image:
<IMAGE_T2>
Write the change caption now (only the sentence(s)).
```

Fact-tuple format. The following JSON illustrates the format of FACT_TUPLES_JSON, not a specific dataset record.

```
[
  {
    "from": "agricultural land",
    "to": "artificial surface",
    "verb": "replaced by",
    "importance": "major",
    "location": "bottom-left"
  }
]
```

A.5 Cross-Output Consistency Protocol

We evaluate two complementary relations among the same model's outputs: mask–mask consistency (MMC) between the semantic and change predictions, and caption–mask consistency (CMC) between the caption and the transitions inside the predicted change mask. Neither metric uses ground truth. Label 0 denotes background.

MMC. The predicted semantic maps imply a query-scoped change mask

$$\hat{M}_{\text{sem}}(x) = \mathbf{1}\left[\hat{S}_1(x) \neq \hat{S}_2(x) \wedge \hat{S}_1(x) > 0 \wedge \hat{S}_2(x) > 0\right]. \tag{28}$$

For each scene, MMC is the Dice overlap between this mask and the predicted change mask:

$$\text{MMC}_{\text{scene}} = \frac{2|\hat{M}_\Delta \cap \hat{M}_{\text{sem}}|}{|\hat{M}_\Delta| + |\hat{M}_{\text{sem}}|}. \tag{29}$$

The score is 1 when both masks are empty and 0 when only one is empty. MMC is computed directly from the binary predictions, independently of caption parsing and the CMC filtering below.

CMC. CMC is computed in three steps.

Step 1: mask-derived transitions. The predicted change mask $\hat{M}_\Delta$ determines which regions count as changed, and the semantic maps $\hat{S}_1, \hat{S}_2$ identify what changed within them. We remove connected components smaller than 64 pixels from $\hat{M}_\Delta$, label each remaining pixel by its class pair $(\hat{S}_1(x), \hat{S}_2(x))$, keep pixels whose two labels are nonzero and different, group them by class pair, and again remove components smaller than 64 pixels within each group. The remaining pixels of each pair form one mask-derived transition. Its location is determined from its own pixels: on a 3×3 grid, a cell is *occupied* if it holds at least 8% of the transition's pixels, and a transition with at least seven occupied cells also supports "throughout the scene."

Step 2: matching. The caption is parsed into $(\text{from}, \text{to}, \text{location})$ tuples with normalized class names and locations. A caption tuple matches a mask-derived transition when the two classes agree and the caption's location is one of the transition's occupied cells. A location of "throughout the scene" matches only a transition with at least seven occupied cells. Mask-derived transitions are processed from largest to smallest, each paired with at most one unused caption tuple, and no element is matched twice.

Step 3: scoring. Let N_C be the number of caption tuples, N_M the number of mask-derived transitions, and m the number of matched pairs. For scenes with mask-derived transitions,

$$P = \frac{m}{N_C}, \qquad R = \frac{m}{N_M}, \qquad \text{CMC}_{\text{scene}} = \frac{2PR}{P+R}. \tag{30}$$

Caption tuples that match no mask-derived transition lower precision, and mask-derived transitions that the caption does not mention lower recall. Because the reference is the model's own mask rather than ground truth, a change that the mask fails to detect does not enter N_M and is not penalized. CMC therefore measures agreement between outputs, not their correctness. When $P + R = 0$ or parsing fails, the score is 0. If the cleaned change mask is empty, the scene is treated as no-change: the score is 1 if the caption declares no change or states no transition, and 0 otherwise. A nonempty cleaned change mask with no valid transition scores 0, as does a caption that declares no change against a nonempty mask with valid transitions.

Both scores are reported as 100 times the mean per-scene value over the 1,606 test queries. They quantify whether a model's outputs agree with one another, not whether they are correct: a model can be consistently wrong and score highly.

A.6 Baseline Adaptations

We summarize the adaptations behind Table 1. All baselines are initialized from the pretrained weights of their constituent foundation models, not from the task-finetuned checkpoints released by the baseline methods, and are trained on CoCD-Dataset. Except for the changes below, we follow each method's official implementation and training recipe.

General segmentation MLLMs. LISA (Lai et al., 2024), PixelLM (Ren et al., 2024), and Sa2VA (Yuan et al., 2025) share one two-frame interface. We replace `[SEG]` with `[T1]`, `[T2]`, and `[CHANGE]` and add role-specific frame routing while retaining each model's native mask decoder. Semantic masks for `[T1]` and `[T2]` are decoded from the corresponding date's image features. Only the `[CHANGE]` path merges the two frames: at each selected feature level, a 1×1 convolution

maps the channel-wise concatenation of the two feature maps back to the original channel count, initialized with weights $[0.5I \mid 0.5I]$ and zero bias so that its initial output is their mean, and the merged features enter a single change-decoding call. The image encoders remain frozen, and we retain the original LoRA ranks (8 for LISA and PixelLM, 256 for Sa2VA). Model-specific details are as follows. In LISA and PixelLM the two images form separate visual-token sequences, the fusion operates on SAM features in LISA and on neck outputs in PixelLM, LoRA targets are expanded from query/value to query/key/value/output projections, and the text loss is replaced with our weighted cross-entropy. Sa2VA uses Qwen3-VL-4B with SAM2, a backbone configuration supported by its release, receives the image pair as a two-frame video through its frame-wise interface, and merges the two frames at each of its three FPN levels before the `[CHANGE]` decode.

CD MLLMs. For ChangeVG (Xue et al., 2026), we train the SegFormer-B1 dual branch on our data using its three-stage procedure. For DeltaVLM (Deng et al., 2026), we add rank-8, scale-16 LoRA to the otherwise frozen Vicuna-7B's query/key/value/output projections. The UniChange-style baseline is our reimplementation described in Appendix A.2, with differences from and comparison to the original in Appendix B.5.

B Supporting Results

B.1 Caption Content Accuracy against Ground Truth

CMC (Appendix A.5) measures agreement between a caption and the same model's masks, not correctness. To assess correctness, we score captions against the ground-truth transition tuples used to generate the reference captions (Appendix A.4), using the same parser as CMC. *Cap-Acc* is the per-scene F1 between caption tuples and ground-truth tuples, where a match requires the same directed class pair and a stated location inside the transition's ground-truth occupied cells. Recall is taken over ground-truth major transitions, so omitted transitions lower it, and precision over all caption tuples, so hallucinated transitions lower it. Cap-Acc is computed on test scenes with at least one ground-truth major transition, and a caption that declares no change or states no transition on such a scene scores 0 rather than being excluded. *No-change accuracy* is computed on scenes with no ground-truth change and is the fraction whose caption declares no change and states no transition. Both are averaged over scenes.

Table 7 reports both scores for all models in Tables 1 and 2. CoCD reaches 49.92 Cap-Acc, 8 points above the best external model and 19 above the matched UniChange-style baseline. Task chaining and encoder-side interaction improve Cap-Acc over Parallel by about 1 point, whereas decoded-state feedback adds a further 18. On no-change scenes, Parallel and Naive-Chain are slightly more accurate than CoCD.

Table 7: Caption content accuracy against ground-truth transitions on the CoCD-Dataset test set. Cap-Acc is computed on scenes with ground-truth change and No-change acc. on the remaining scenes.

Method	Cap-Acc	No-change acc.
LISA*	41.37	79.49
Sa2VA*	41.82	82.98
PixelLM*	40.32	77.62
ChangeVG	24.21	66.20
DeltaVLM	27.56	79.02
Parallel (UniChange-style†)	30.57	**86.71**
Naive-Chain	31.62	85.78
Naive-Chain + encoder diff	31.47	84.62
CoCD	**49.92**	84.15

B.2 Dedicated Change Decoder versus Semantic XOR

Table 8 compares the dedicated change decoder with change masks derived by XOR of the two predicted semantic maps. The dedicated decoder is more precise at similar recall. Notably, XOR

(27.44 oIoU) and a change decoder trained without SSP (27.37, Table 4) reach similar accuracy, whereas the decoder that combines both sources through SSP reaches 31.33, suggesting that semantic states and visual evidence are complementary rather than redundant for localizing change.

Table 8: Change-mask derivation on CoCD-Dataset.

Derivation	oIoU	F1	Precision	Recall
Dedicated `[CHANGE]` decoder	**31.33**	**47.71**	**64.51**	**37.86**
Semantic XOR	27.44	43.07	51.61	36.95

B.3 Soft versus Hard Semantic States in SSP

SSP paints class embeddings weighted by the decoder's class probabilities (Eq. 3). Table 9 compares this soft state with a hard variant that paints only the argmax class embedding at each pixel, under the full CoCD configuration. Soft states improve change localization by 3.5 oIoU and are better on seven of the eight metrics, while the hard variant attains slightly higher CMC.

Table 9: Soft versus hard semantic states in SSP on CoCD-Dataset, with all other components fixed.

SSP state	oIoU	F1	sem-mIoU	κ	BLEU-4	CIDEr	MMC	CMC
Soft (CoCD)	**31.33**	**47.71**	**52.61**	**64.02**	**30.00**	**170.26**	**55.77**	66.71
Hard	27.79	43.49	50.55	62.50	29.01	165.64	54.50	**68.78**

B.4 MLLM Backbone Dependence

Does the gain depend on the MLLM backbone? We repeat the Parallel–CoCD comparison on two additional MLLMs that differ from our main Qwen3-VL-4B (Bai et al., 2025) MLLM backbone along complementary axes. InternVL3-2B (Zhu et al., 2025b) changes the vision encoder and halves the model scale, while LLaVA-1.5-7B (Liu et al., 2024) changes the language-model lineage to Vicuna/LLaMA (Chiang et al., 2023; Touvron et al., 2023), uses a CLIP-336 vision encoder (Radford et al., 2021), and changes the input resolution. Data, optimization budget, LoRA placement, and all hyperparameters are held fixed. Only the MLLM backbone changes.

CoCD improves on Parallel in change localization, per-date semantics, caption quality, and cross-output consistency for each tested MLLM (Table 10).

Table 10: MLLM-backbone robustness under matched training. Data, optimization, LoRA placement, and hyperparameters are fixed. Only the MLLM backbone changes.

MLLM	Method	oIoU	sem-mIoU	κ	BLEU-4	CIDEr	MMC	CMC
Qwen3-VL-4B (main)	Parallel	25.00	45.78	59.75	21.98	131.57	37.98	35.18
	CoCD	**31.33**	**52.61**	**64.02**	**30.00**	**170.26**	**55.77**	**66.71**
	Δ	+6.33	+6.83	+4.27	+8.02	+38.69	+17.79	+31.53
InternVL3-2B	Parallel	26.87	46.66	61.52	21.44	128.98	43.25	39.50
	CoCD	**29.72**	**48.36**	**62.49**	**28.13**	**155.38**	**55.85**	**66.60**
	Δ	+2.85	+1.70	+0.97	+6.69	+26.40	+12.60	+27.10
LLaVA-1.5-7B	Parallel	22.66	33.59	44.21	20.48	128.69	32.20	35.11
	CoCD	**27.12**	**48.33**	**61.02**	**27.97**	**150.24**	**53.99**	**65.55**
	Δ	+4.46	+14.74	+16.81	+7.49	+21.55	+21.79	+30.44

B.5 Relation to UniChange

Architectural differences. The adaptation replaces UniChange's RSBuilding backbone and custom decoder with SAM2.1 Hiera-L and a SAM2-family decoder. The backbone replacement was motivated by our inability to access the pretrained weights during our experiments. It also adds stage-anchor and class context to mask queries and uses three decoding calls with convolutional temporal

fusion and a visual dense prompt, in place of UniChange's joint query refinement and differencing of attention-refined date features. Thus, "UniChange-style" denotes the inherited token-conditioned interface, not an architecturally identical reproduction.

Performance of the starting point. To check that the adaptation preserves the strength of the original, we train and evaluate our UniChange-style (Parallel) implementation under UniChange's own joint-training protocol, using the same benchmark datasets and splits. It achieves results close to UniChange (Zhang et al., 2026) on WHU-CD (Ji et al., 2019), S2Looking (Shen et al., 2021), LEVIR-CD+ (Chen & Shi, 2020; Shen et al., 2021), and SECOND (Yang et al., 2022), with an average absolute IoU difference of about one point (Table 11). UniChange scores are taken from the original paper, not from our reimplementation. This establishes a competitive starting point despite the changes to the MLLM, vision backbone, and decoder.

Table 11: UniChange (Zhang et al., 2026) and our SAM2-based UniChange-style (Parallel) implementation on the original UniChange joint-training benchmarks.

Benchmark	Metric	UniChange	UniChange-style (Parallel)	Difference
WHU-CD	F1 / IoU	94.96 / 90.41	94.35 / 89.30	-.61 / -1.11
S2Looking	F1 / IoU	69.32 / 53.04	69.43 / 53.18	+.11 / +.14
LEVIR-CD+	F1 / IoU	88.19 / 78.87	87.54 / 77.85	-.65 / -1.02
SECOND	mIoU / SeK / IoU_c	72.85 / 23.02 / 57.62	72.98 / 24.99 / 58.83	+.13 / +1.97 / +1.21

B.6 Ground-Truth versus Previous-Epoch Evidence

We compare two ways to supply CEI evidence during training (Table 12). **Clean GT** constructs evidence from ground-truth semantic and change masks throughout all 10 epochs, teacher-forcing the mask evidence supplied to the caption stage. **Previous-epoch (CoCD)** instead uses cached predictions from the preceding training epoch, with GT used in the first epoch and for cache misses. Both retain SSP with current same-forward semantic predictions and require no additional evidence-generation forward pass.

At inference, both use only their own predictions from the current task chain. Naive-Chain, which uses neither SSP nor CEI, serves as a reference.

Table 12: GT versus previous-epoch predictions as training-time CEI evidence on CoCD-Dataset. Both use model-predicted evidence at test time.

Training evidence	oIoU	F1	sem-mIoU	κ	BLEU-4	CIDEr	MMC	CMC
No re-entry (Naive-Chain)	25.59	40.75	51.43	62.61	21.96	125.14	48.74	45.11
Clean GT	30.04	46.20	51.73	63.61	28.16	158.78	53.38	**70.94**
Previous-epoch (**CoCD**)	**31.33**	**47.71**	**52.61**	**64.02**	**30.00**	**170.26**	**55.77**	66.71

Both evidence sources outperform Naive-Chain across the reported metrics, so the gains over this reference do not require a rolling cache. Previous-epoch evidence yields better localization, semantic segmentation, caption quality, and MMC, while Clean GT achieves higher CMC. The cache therefore offers a practical way to train with predicted evidence without an additional forward pass.

C ADDITIONAL DATASET VISUALIZATIONS

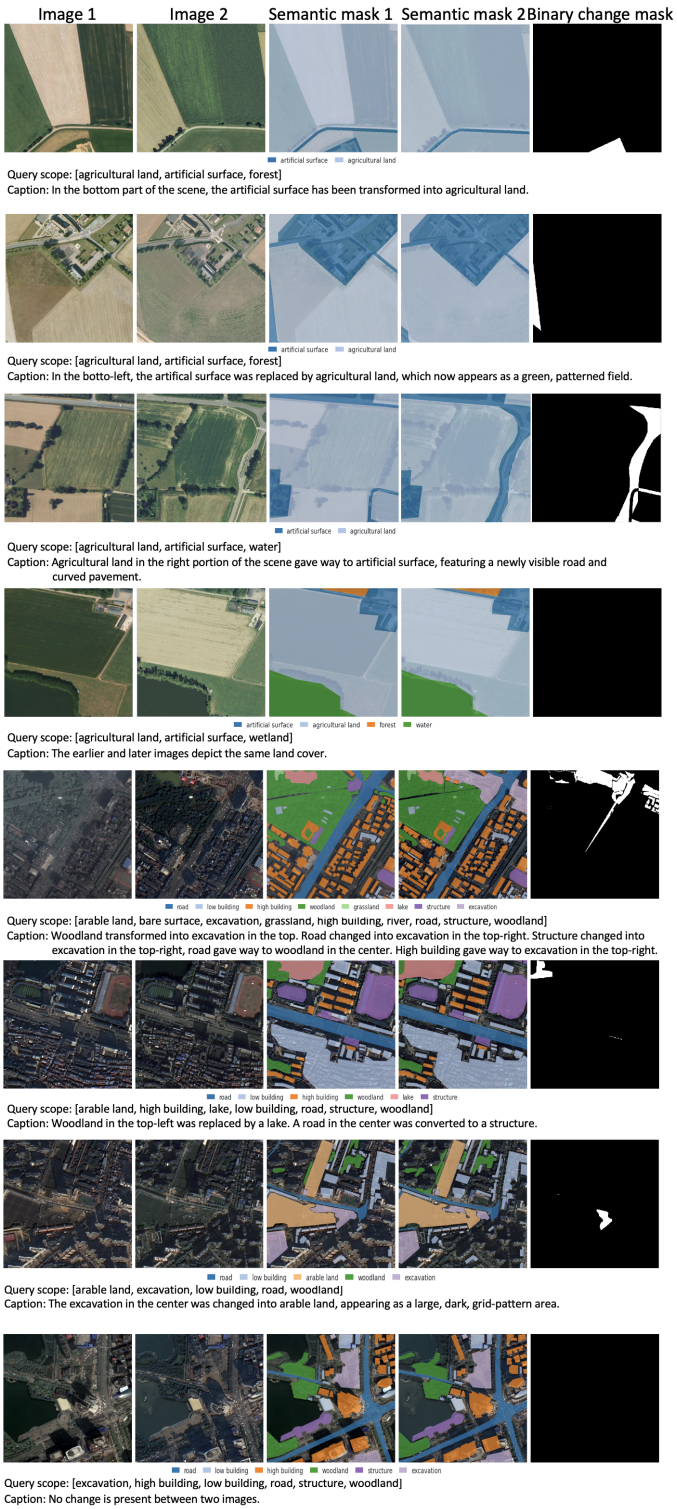

Figure 7: Additional examples from CoCD-Dataset. For compactness, only the queried class scope and final <INFER> caption are shown from each structured trace.

D ADDITIONAL QUALITATIVE RESULTS

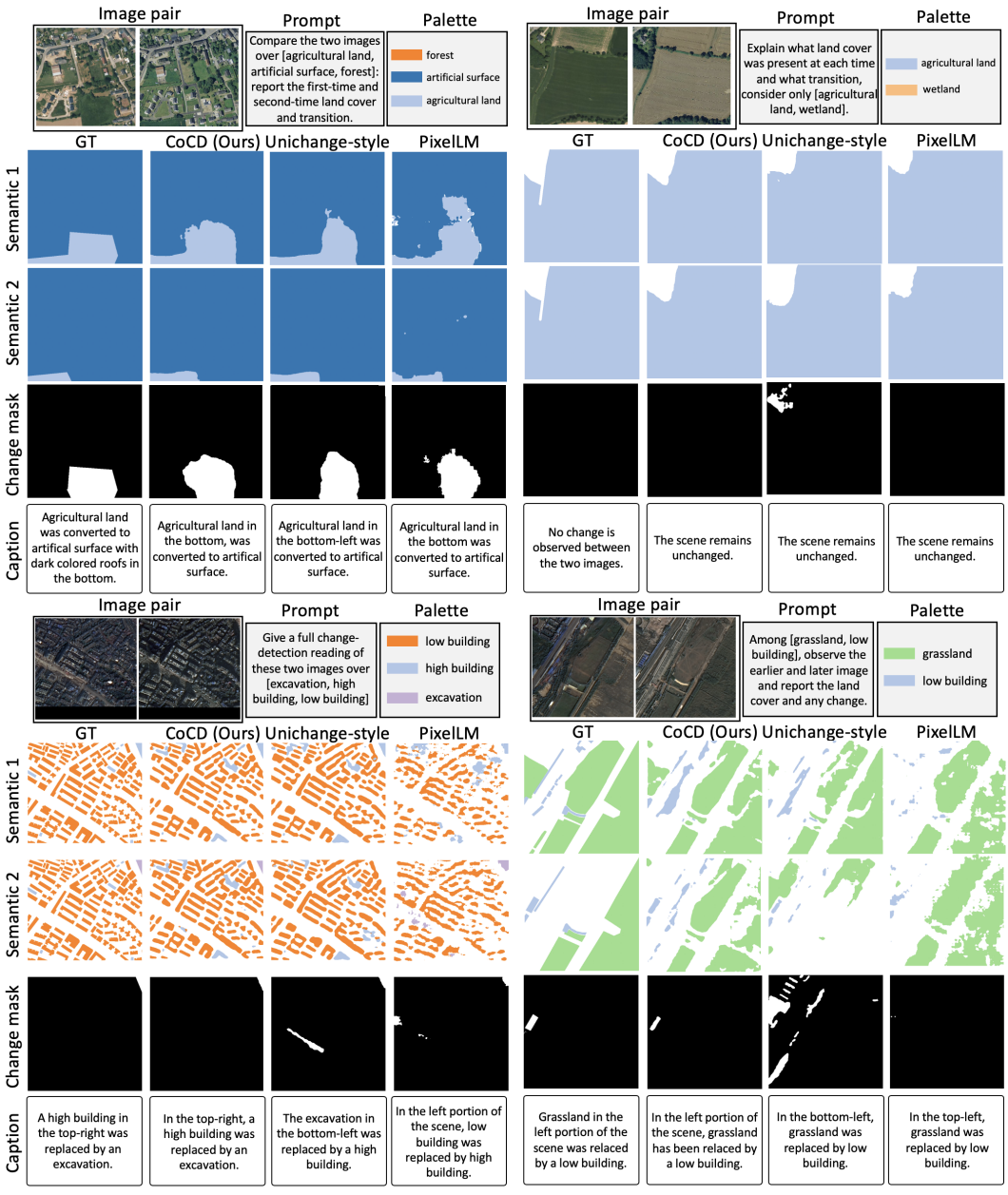

Figure 8: Additional qualitative results on CoCD-Dataset. Only the final <INFER> captions are shown. Intermediate execution spans and injected evidence are omitted.