RESOLUTION-ALIGNED CHANGE DETECTION VIA DIFFUSION SUPER-RESOLUTION AND RL-GUIDED UNSEEN-CLASS ADAPTATION

Anonymous authors
Paper under double-blind review

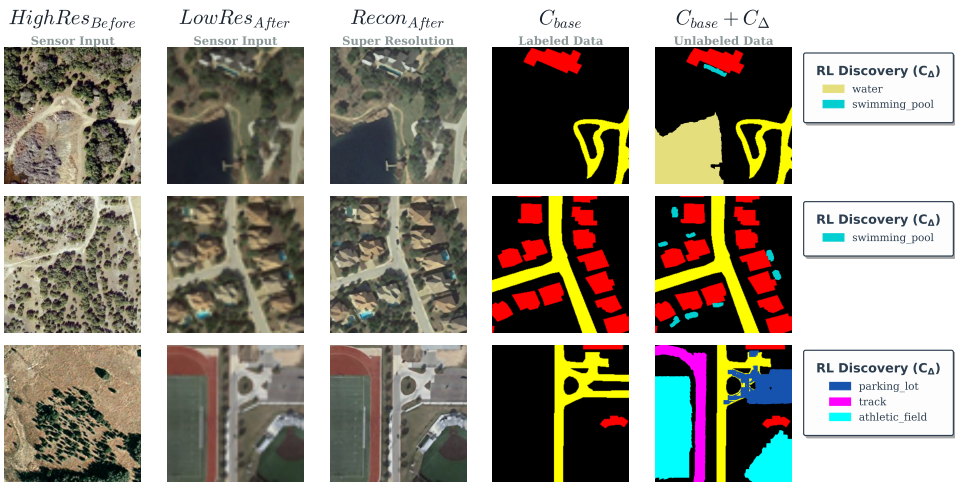

Figure 1: A diffusion-based super-resolution module reconstructs the low-resolution *after* image in the spatial domain of the high-resolution *before* image, enabling resolution-aligned change detection with the base detector C_{base}. For unlabeled image pairs containing held-out semantic classes, initial change candidates are sequentially refined, accepted, or rejected by an RL policy guided by a retrieval-augmented reward model. Accepted masks are used as pseudo-labels to adapt the detector, producing $C_{\text{base}} + \Delta C$ without requiring target-scene pixel-level annotations for the held-out classes.

ABSTRACT

Cross-resolution change detection (CD) is challenging because differences in spatial resolution between temporal observations introduce structural misalignment, inaccurate boundaries, and false change responses. An additional challenge arises when semantic classes excluded from base-detector training may appear after deployment without target-scene pixel-level annotations. We introduce a resolution-aligned adaptation framework that combines conditional diffusion-based super-resolution with reinforcement-learning-guided pseudo-label refinement. First, a conditional diffusion model reconstructs the low-resolution after image in the spatial domain of the high-resolution before image using noise-prediction, reconstruction, and perceptual objectives that preserve geometry and empirically reduce reconstruction-induced artifacts. A shared-weight Siamese detector with cross-temporal attention then predicts changes from the aligned image pair. To adapt beyond the original label space, candidate-mask refinement is formulated as a finite-horizon decision process. An RL policy iteratively refines, accepts, or rejects candidate regions using a retrieval-augmented reward combining geometric consistency, image-text alignment, and language-guided semantic feedback. Accepted masks provide pseudo-labels for target-mask-free adaptation to semantic classes held out from base-detector training. Experiments on synthetic and real cross-resolution benchmarks demonstrate improved reconstruction fidelity, change-detection accuracy, held-out-class coverage, mask quality, and post-adaptation detector performance over matched baselines.

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.

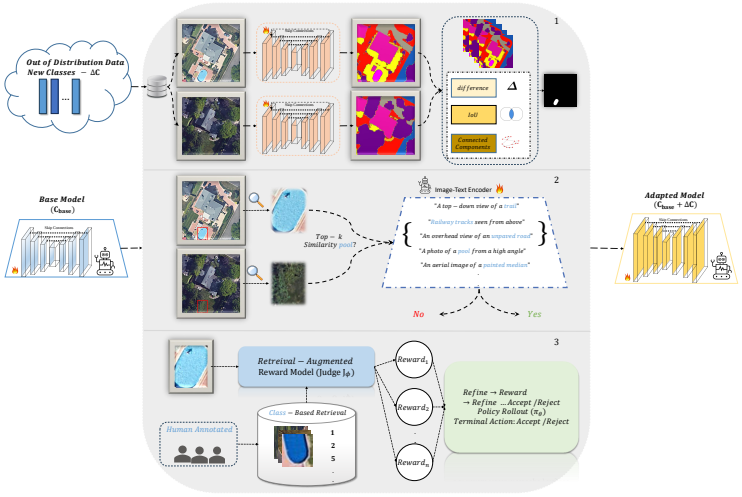

Figure 2: The base model first generates candidate change masks from semantic segmentation differences, followed by connected-component extraction and geometric filtering. Each candidate initializes a finite-horizon RL episode in which the policy sequentially refines the mask or issues a terminal accept/reject decision using retrieved support examples and retrieval-augmented reward feedback. Accepted terminal masks are retained as pseudo-labels and used to adapt the detector beyond its original semantic label space.

1 Introduction

Change detection (CD) in remote sensing aims to identify and localize changes between images of the same geographic area acquired at different times. In practical settings, however, the *after* image may be available only at low resolution (LR), while the corresponding *before* image is available at high resolution (HR). This discrepancy degrades spatial alignment, boundary localization, and semantic interpretation (Chen et al., 2023; Tian et al., 2022; Yang et al., 2025). Directly downsampling the HR image removes useful structural detail, whereas conventional upsampling of the LR image may introduce artifacts that are incorrectly interpreted as change. Although recent methods address cross-resolution CD through resolution-aware architectures or generative modeling (Wang et al., 2025; Zheng et al., 2021; Wu et al., 2023), reliable comparison across mismatched spatial scales remains challenging. A second difficulty arises when semantic change classes excluded from base-detector training may appear after deployment. Obtaining pixel-level annotations for such classes is costly, while supervised detectors remain restricted to their original label space. **How can a model align cross-resolution observations while adapting to semantic classes held out from base-detector training without target-scene pixel-level annotations?**

We address these challenges through a resolution-aligned framework combining diffusion-based super-resolution with reinforcement-learning-guided adaptation. First, a conditional diffusion model reconstructs the LR *after* image in the spatial domain of the HR *before* image. Building on the strong reconstruction capabilities of diffusion and super-resolution models (Lorenz et al., 2022; Saharia et al., 2022; Ledig et al., 2017), our formulation combines noise-prediction, reconstruction, and perceptual objectives to preserve geometric structure and reduce reconstruction-induced artifacts that may otherwise produce false change responses. A shared-weight Siamese detector with cross-temporal attention then predicts changes from the aligned image pair. Experiments on (Chen & Shi, 2020; Shen et al., 2021; Liu et al., 2020) show that this formulation consistently outperforms bicubic upsampling and matched resolution-normalization baselines. To extend the detector beyond its original label space, we formulate candidate-mask refinement as a finite-horizon decision process. Inspired by vision-language evaluation and reward-guided adaptation (Wu et al., 2026; Lee et al., 2024; Team et al., 2023), an RL policy sequentially refines, accepts, or rejects candidate regions using a retrieval-augmented reward that combines geometric consistency, image-text alignment, and language-guided semantic feedback. Accepted masks are retained as pseudo-labels for adapting the detector to held-out classes without new pixel-level annotations. End-to-end evaluation measures

candidate coverage, semantic correctness, mask quality, and detector performance before and after adaptation. **Our contributions are threefold:**

Cross-resolution adaptation setting. We study change detection under the joint challenges of resolution-mismatched bi-temporal imagery and semantic classes held out from base-detector training, providing a unified setting for resolution alignment and target-mask-free adaptation.

Retrieval-augmented sequential mask refinement. We formulate candidate refinement and selection as a finite-horizon decision process in which an RL policy iteratively improves masks and makes terminal accept/reject decisions using geometric, image-text, and retrieval-augmented semantic feedback.

Integrated resolution-aligned adaptation framework. We combine SR-based resolution alignment, cross-temporal change detection, and policy-guided pseudo-label refinement in a unified pipeline, and evaluate how these stages affect reconstruction quality, change localization, held-out-class discovery, and downstream adaptation.

2 Related Work

Change Detection and Datasets. Remote-sensing change detection (*CD*) has progressed from image differencing to fully convolutional and Siamese architectures (Daudt et al., 2018; Liu et al., 2020). Existing reviews summarize advances in model design, supervision, and applications (You et al., 2020; Li et al., 2018; Zhu et al., 2017; Chen et al., 2025). Public benchmarks, including LEVIR-CD and its variants (Chen & Shi, 2020; Liu et al., 2024; 2022a), S2Looking (Shen et al., 2021; Van Etten et al., 2021), and RSCD (Li et al., 2023), enable systematic evaluation of structural and land-cover changes. Most existing pipelines assume that the bi-temporal inputs (I_b, I_a) have approximately matched spatial resolution. Their performance can therefore degrade when the low-resolution *after* image satisfies $\mathrm{res}(I_a) \ll \mathrm{res}(I_b)$, motivating methods designed explicitly for cross-resolution observations.

Diffusion-Based Super Resolution with Attention. Diffusion-based super-resolution (*SR*) models reconstruct high-quality images by learning the reverse of a progressive noising process (Lorenz et al., 2022; Saharia et al., 2022). Their strong perceptual reconstruction ability has made them effective for image restoration. However, most existing methods optimize single-image reconstruction and do not explicitly consider the temporal consistency required for downstream *CD*, where hallucinated structures may be incorrectly interpreted as changes. We adopt a conditional diffusion-based SR formulation as the resolution-alignment stage of our framework. Our contribution does not lie in introducing a new diffusion SR objective, but in studying its use within cross-resolution change detection and its interaction with downstream held-out-class adaptation.

Cross-Resolution and Generative Change Detection. Resolution-aware *CD* methods commonly employ multi-scale representations (Tian et al., 2022; Yang et al., 2025; Wang et al., 2025) or continuous cross-resolution mappings (Chen et al., 2023). Other approaches rely on image reconstruction (Noh et al., 2022), cross-resolution differencing (Zheng et al., 2021), transformer-based translation (Xu et al., 2023), or GAN-based generation under weak supervision (Wu et al., 2023). Ref-Diff (Dong et al., 2024) further incorporates change priors into conditional diffusion for reference-based remote-sensing super-resolution, highlighting the connection between temporal change information and cross-resolution reconstruction. Transformer and attention-based architectures improve temporal fusion and long-range context modeling (Bandara & Patel, 2022; Zhang et al., 2020; Fang et al., 2021; Chen et al., 2021; Feng et al., 2022; 2023). Representative convolutional baselines include FC-EF, FC-Siam-Diff, and FC-Siam-Conc (Daudt et al., 2018), while SUNet and SRCDNet explicitly address cross-resolution inputs (Shao et al., 2021; Liu et al., 2022b). Rather than introducing a new super-resolution formulation, our framework uses diffusion-based reconstruction as a resolution-alignment stage before cross-temporal change detection, and couples this stage with subsequent target-mask-free adaptation to held-out semantic classes.

RL for Mask Refinement and Guided Selection. Reinforcement learning has been used for iterative mask generation (Liu et al., 2025), instance-segmentation refinement (Hilt et al., 2023), and task-aware sample selection (Saeed et al., 2021). Reward models have also been used to represent perceptual quality and guide adaptation (Wu et al., 2026). Related foundation-model approaches include AnyChange (Zheng et al., 2024), which performs training-free zero-shot change detection

through bitemporal feature matching, and AlignSAM (Huang et al., 2024), which uses reinforcement learning to adapt SAM through iterative prompting. In contrast, our method formulates adaptation to classes held out from base-detector training as a finite-horizon decision process over object-level change candidates. The policy sequentially refines each candidate and issues a terminal accept/reject decision using a retrieval-augmented reward combining geometric consistency, image-text alignment, and language-guided semantic feedback. Accepted masks are then used as pseudo-labels for adapting the detector beyond its original label space.

3 METHODOLOGY

Our framework addresses two coupled deployment challenges: spatial-resolution mismatch between bi-temporal observations and adaptation to semantic classes excluded from task-specific detector training. The pipeline follows the execution order shown in Figure 2: (1) diffusion-based resolution alignment, (2) base change detection, (3) candidate generation and filtering, (4) retrieval-augmented sequential refinement, (5) terminal accept/reject selection, (6) semantic assignment, and (7) detector adaptation.

3.1 PROBLEM FORMULATION AND SUPERVISION SETTING

Let $I_b \in \mathbb{R}^{H\times W\times 3}$ denote the high-resolution (*HR*) *before* image and $I_a^{\downarrow} \in \mathbb{R}^{H_{\downarrow}\times W_{\downarrow}\times 3}$ the corresponding low-resolution (*LR*) *after* observation, where $H_{\downarrow} < H$ and $W_{\downarrow} < W$. We assume that the two observations correspond to the same geographic region and that benchmark preprocessing provides scene-level spatial correspondence; our method addresses the resolution mismatch rather than performing explicit geometric registration.

Let $\mathcal{C}_{\text{all}}$ denote the unified semantic taxonomy. We partition it into disjoint task-specific seen and held-out sets,

$$\mathcal{C}_{\text{seen}} \cap \mathcal{C}_{\text{unseen}} = \emptyset, \qquad \mathcal{C}_{\text{seen}} \cup \mathcal{C}_{\text{unseen}} = \mathcal{C}_{\text{all}}. \tag{1}$$

Here, "unseen" is defined relative to the task-specific detector and adaptation pipeline. Classes in $\mathcal{C}_{\text{unseen}}$ are excluded from supervised training of the base detector, candidate-scoring modules, retrieval support set, and adaptation policy. During adaptation, no ground-truth target-scene masks for these held-out classes are used.

A separate frozen semantic model may contain prior knowledge of the unified taxonomy and is used only to assign semantic categories to accepted regions. Thus, the setting is target-mask-free rather than category-agnostic.

The model produces two outputs. A binary change head localizes changed pixels, while a semantic head assigns an after-image semantic category to detected change regions. The goal is therefore to recover both spatial change masks and semantic change categories under cross-resolution input conditions.

3.2 DIFFUSION-BASED RESOLUTION ALIGNMENT

The first stage reconstructs the *LR after* observation in the spatial domain of the *HR before* image. Let $I_a \in \mathbb{R}^{H\times W\times 3}$ denote the *HR* training target. We learn $G_\theta : I_a^{\downarrow} \mapsto \hat{I}_a \approx I_a$, where $\hat{I}_a$ is subsequently used by the change detector.

Conditional reconstruction. The low-resolution observation is first resized using fixed bicubic interpolation, $I_a^{\uparrow} = \text{Bicubic}(I_a^{\downarrow})$. At diffusion step t, the noisy HR sample x_t is concatenated with $I_a^{\uparrow}$ and processed by a conditional U-Net,

$$\hat{\epsilon} = \epsilon_\theta([x_t \| I_a^{\uparrow}], t). \tag{2}$$

The network uses residual encoder-decoder blocks, skip connections, multi-head self-attention, and timestep conditioning. Standard forward and reverse diffusion definitions and sampling details are provided in the appendix.

Training objective. The reconstruction model is optimized using

$$\mathcal{L}_{\text{SR}} = \lambda_1 \mathcal{L}_{\text{noise}} + \lambda_2 \mathcal{L}_{\text{rec}} + \lambda_3 \mathcal{L}_{\text{perc}}, \tag{3}$$

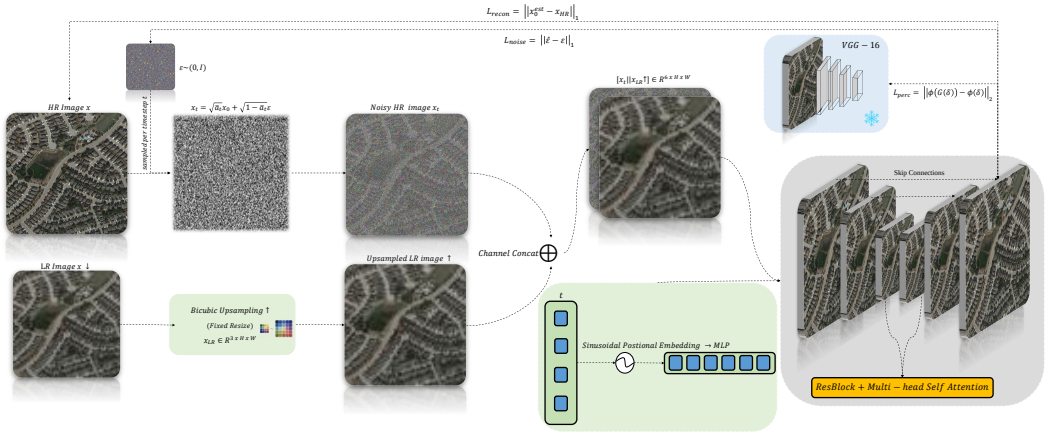

Figure 3: **Diffusion-based resolution alignment.** A conditional U-Net reconstructs the low-resolution *after* image in the high-resolution spatial domain. The noisy HR sample is conditioned on a bicubic-upsampled LR observation, while noise-prediction, reconstruction, and perceptual objectives encourage faithful spatial reconstruction.

where $\mathcal{L}_{\text{noise}}$ is the standard diffusion noise-prediction objective, $\mathcal{L}_{\text{rec}} = \|\hat{I}_a - I_a\|_1$ promotes pixel-level fidelity, and $\mathcal{L}_{\text{perc}}$ compares frozen VGG features of $\hat{I}_a$ and I_a.

These objectives encourage geometric and perceptual consistency and empirically reduce reconstruction-induced artifacts. The current formulation produces a single reconstruction and does not explicitly estimate predictive uncertainty; consequently, residual reconstruction errors may still propagate to the downstream change detector.

3.3 Resolution-Aligned Change Detection

The reconstructed image $\hat{I}_a$ and native *HR* image I_b are processed by a shared-weight Siamese detector. Resolution alignment allows the detector to compare structures in a common spatial domain rather than directly mixing observations with different pixel resolutions.

Shared feature extraction and temporal fusion. A Siamese encoder F_ψ produces

$$\mathbf{f}_b = F_\psi(I_b), \qquad \mathbf{f}_a = F_\psi(\hat{I}_a). \tag{4}$$

The deepest features are tokenized and fused using bidirectional cross-attention,

$$\mathbf{T}'_b = \text{Attn}(\mathbf{T}_b, \mathbf{T}_a), \qquad \mathbf{T}'_a = \text{Attn}(\mathbf{T}_a, \mathbf{T}_b), \tag{5}$$

followed by

$$\mathbf{T}_{\text{fused}} = [\mathbf{T}'_b, \mathbf{T}'_a, |\mathbf{T}'_b - \mathbf{T}'_a|]. \tag{6}$$

Multi-scale decoding. The fused representation is decoded with residual upsampling blocks and temporal skip features. Binary change heads are attached at 64×64, 128×128, and 256×256, and are trained using

$$\mathcal{L}_{\text{CD}} = \sum_{s \in \{64,128,256\}} w_s \left[\lambda_{\text{focal}} \mathcal{L}_{\text{focal}}(\hat{M}_s, M_s) + \lambda_{\text{dice}} \mathcal{L}_{\text{dice}}(\hat{M}_s, M_s) \right]. \tag{7}$$

A semantic head branches from the full-resolution representation to predict per-pixel semantic logits. The binary head therefore localizes change, while the semantic head provides task-specific category predictions.

3.4 Candidate Generation and Pre-Filtering

Adaptation begins by constructing object-level candidate regions before any policy rollout. A frozen segmentation model S predicts semantic maps for the aligned pair,

$$L_b = S(I_b), \qquad L_a = S(\hat{I}_a), \tag{8}$$

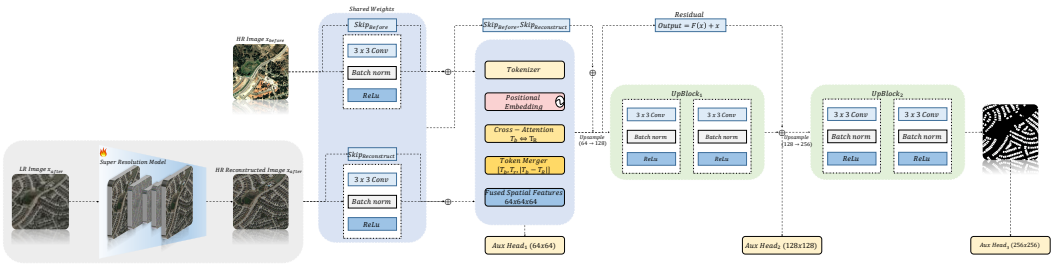

Figure 4: **Resolution-aligned change detection.** The reconstructed *after* image and native high-resolution *before* image are encoded using shared weights. Bidirectional cross-temporal attention fuses the two representations, followed by residual multi-scale decoding for binary change localization and semantic prediction.

from which we compute a semantic disagreement mask

$$D = \mathbf{1}[L_b \neq L_a]. \tag{9}$$

Connected components are extracted from the semantic predictions and filtered using class-dependent area, temporal-disagreement, changed-pixel, and overlap criteria. Each surviving component c_i is converted into a binary candidate mask

$$m_i(p) = \mathbf{1}[p \in c_i]. \tag{10}$$

An image-text encoder then removes semantically implausible candidates using similarity, area, and overlap thresholds. The resulting set $\mathcal{M}_{\text{cand}}$ provides the initial states for policy optimization. Detailed connected-component and pre-filtering algorithms are provided in the appendix.

3.5 Retrieval-Augmented Sequential Adaptation

Rather than treating pseudo-label selection as a one-shot thresholding problem, we formulate candidate refinement as a finite-horizon decision process. Each candidate $m_i^{(0)} \in \mathcal{M}_{\text{cand}}$ initializes one episode.

Retrieval and state construction. For the current candidate $m_i^{(t)}$, we retrieve support examples from the seen-class training database,

$$\mathcal{R}(m_i^{(t)}) = \left\{ x_j \in \mathcal{D}_{\text{ret}} \mid \text{sim}(x_j, m_i^{(t)}) \geq \tau_r \right\}. \tag{11}$$

No masks from the held-out target adaptation images are included in $\mathcal{D}_{\text{ret}}$.

The policy state is

$$s_i^{(t)} = \left(I_b, \hat{I}_a, m_i^{(t)}, \mathcal{R}(m_i^{(t)}), h_i^{(t)} \right), \tag{12}$$

where $h_i^{(t)}$ records the previous actions.

The action space is

$$\mathcal{A} = \{\text{expand}, \text{contract}, \text{shift}_x, \text{shift}_y, \text{noop}, \text{accept}, \text{reject}\}. \tag{13}$$

The first five actions modify or retain the current mask, whereas `accept` and `reject` terminate the episode.

Retrieval-augmented quality score. We define

$$Q(m_t) = \gamma \, \text{IoU}(m_t, D) + \alpha \, f_{\text{img}}(m_t) + \beta \, f_{\text{text}}\left(m_t \mid \mathcal{R}(m_t)\right), \tag{14}$$

where f_{img} measures visual-semantic consistency and f_{text} is the frozen retrieval-augmented LLM judgment.

Improvement-based reward. Using the absolute quality score as a reward at every step could favor unnecessarily long trajectories. We instead reward improvements and explicitly supervise terminal decisions:

$$r_t = \begin{cases} Q(m_{t+1}) - Q(m_t) - \lambda_{\text{step}}, & a_t \in \mathcal{A}_{\text{refine}}, \\ Q(m_t) - \tau_{\text{acc}}, & a_t = \text{accept}, \\ \tau_{\text{acc}} - Q(m_t), & a_t = \text{reject}, \end{cases} \tag{15}$$

where

$$\mathcal{A}_{\text{refine}} = \{\text{expand}, \text{contract}, \text{shift}_x, \text{shift}_y, \text{noop}\}. \tag{16}$$

The step cost $\lambda_{\text{step}} > 0$ discourages unnecessarily long episodes. In particular, a no-op yields $-\lambda_{\text{step}}$, preventing repeated accumulation of positive reward without improving the mask. The threshold τ_{acc} explicitly determines whether terminal acceptance or rejection is favored.

Policy optimization and rollout. The policy $\pi_\theta(a \mid s)$ and value function $V_\theta(s)$ are optimized with PPO over complete trajectories. At inference, the policy repeatedly retrieves support examples, constructs the state, and either refines the mask or issues a terminal decision. The episode terminates upon `accept`, `reject`, or reaching the maximum horizon T. Candidates that reach the horizon without acceptance are not added to the pseudo-label set.

Accepted terminal masks form

$$\mathcal{M}_{\text{good}} = \{m_i : a_i^{(T_i)} = \text{accept}\}. \tag{17}$$

3.6 Semantic Assignment and Detector Adaptation

Only after terminal mask selection do we assign semantic categories. For every $m_i \in \mathcal{M}_{\text{good}}$, a frozen semantic labeler G_{label}, trained over the unified taxonomy, predicts $\hat{y}_i$, producing the pseudo-label pair

$$(m_i, \hat{y}_i). \tag{18}$$

The semantic labeler is external to the task-specific seen/held-out split and may therefore contain prior semantic knowledge of $\mathcal{C}_{\text{unseen}}$; it does not provide target-scene ground-truth masks.

The base semantic head C_{base} predicts classes in $\mathcal{C}_{\text{seen}}$, while a new head ΔC introduces logits for the held-out categories:

$$C_{\text{adapt}}(\mathbf{z}_{256}) = [C_{\text{base}}(\mathbf{z}_{256}); \Delta C(\mathbf{z}_{256})]. \tag{19}$$

During adaptation, the shared encoder and C_{base} remain frozen, and only ΔC is optimized using the accepted pseudo-labels. Freezing C_{base} preserves its parameters and isolates learning to the new-class head; it does not imply that the final joint argmax is mathematically guaranteed to remain unchanged after the new logits are introduced.

The complete execution path is therefore $I_a^{\downarrow} \to \hat{I}_a \to$ base CD $\to \mathcal{M}_{\text{cand}} \to$ retrieval $\to$ policy refinement $\to \mathcal{M}_{\text{good}} \to G_{\text{label}} \to C_{\text{base}} + \Delta C$.

This ordering separates resolution alignment, change localization, candidate selection, semantic assignment, and detector adaptation, making explicit where each source of supervision enters the framework.

4 Experimental Evaluation and Results

Evaluation protocol. We evaluate on three cross-resolution change-detection benchmarks: **LEVIR-CD** ($4\times$), where the post-event image is downsampled by $4\times$; **SV-CD** ($8\times$), which contains strong seasonal appearance variation and uses $8\times$ post-event downsampling; and **DE-CD** ($3.3\times$), a real cross-sensor setting using Planet ($3\,\text{m}$) and Sentinel-2 ($10\,\text{m}$) imagery, corresponding to an effective resolution ratio of approximately $3.3\times$. Dataset construction, preprocessing, and computational details are provided in the appendix.

The detector produces a binary change mask for localization and semantic predictions for change-category assignment. For held-out-class adaptation, classes in $\mathcal{C}_{\text{unseen}}$ are excluded from task-specific detector and policy training, and no ground-truth target masks for these classes are used during adaptation. A frozen external semantic model may provide prior knowledge over the unified taxonomy, as described in the methodology.

Metrics. For change detection, we report Precision, Recall, F1, IoU, and Overall Accuracy (OA). For super-resolution and semantic alignment, we report PSNR, SSIM, CLIP, SigLIP, and MAMMUT. PSNR is reported in dB, while SSIM, CLIP, SigLIP, and MAMMUT are reported on a 0-100 scale by multiplying their native scores by 100.

Table 1 reports results at input resolutions 32×32, 64×64, and 128×128. The upper block presents change-detection performance, while the lower block evaluates reconstruction fidelity and semantic alignment. Higher input resolution generally improves reconstruction and semantic-alignment quality while preserving strong change-detection performance across all three benchmarks.

Table 1: The top block reports CD performance for each dataset at three input resolutions. The bottom block reports the corresponding reconstruction and semantic-alignment quality. PSNR is reported in dB, while SSIM, CLIP, SigLIP, and MAMMUT scores are multiplied by 100 for readability. CLIP is used off-the-shelf, while SigLIP and MAMMUT are fine-tuned.

	32×32					**64×64**					**128×128**				
Dataset	Pre ↑	Rec ↑	F1 ↑	IoU ↑	OA ↑	Pre ↑	Rec ↑	F1 ↑	IoU ↑	OA ↑	Pre ↑	Rec ↑	F1 ↑	IoU ↑	OA ↑
LEVIR-CD	89.57	90.00	89.78	81.46	95.43	90.21	93.42	91.79	84.83	98.89	90.58	93.46	92.00	85.19	99.01
SV-CD	98.01	98.03	98.02	96.12	98.67	98.07	98.54	98.30	96.66	99.02	98.45	98.78	98.61	97.26	99.40
DE-CD	91.56	72.98	81.22	68.38	91.85	93.52	75.52	83.56	71.76	99.61	93.58	76.58	84.23	72.76	95.73
Super-resolution fidelity and semantic alignment															
	32×32					**64×64**					**128×128**				
Dataset	PSNR ↑	SSIM ↑	CLIP ↑	SigLIP ↑	MAMMUT ↑	PSNR ↑	SSIM ↑	CLIP ↑	SigLIP ↑	MAMMUT ↑	PSNR ↑	SSIM ↑	CLIP ↑	SigLIP ↑	MAMMUT ↑
LEVIR-CD	34.62	90.47	78.85	88.39	84.72	38.11	93.12	81.64	91.67	88.23	41.58	95.36	84.92	94.28	91.57
SV-CD	35.24	91.18	79.53	89.27	85.36	39.03	93.89	82.41	92.34	88.94	42.55	95.78	85.77	94.91	92.41
DE-CD	33.98	89.82	77.62	87.04	83.21	37.45	92.55	80.39	90.26	86.87	40.92	94.93	83.68	93.72	90.44

Table 2: Comparison on the three cross-resolution change-detection benchmarks. Baseline results are reproduced from the corresponding cited publications; therefore, evaluation and metric-aggregation protocols may differ across methods. "–" denotes metrics not reported by the original source, and "*" denotes methods specifically designed for cross-resolution change detection. All reported scores are percentages (%).

	LEVIR-CD (4×)					**SV-CD (8×)**					**DE-CD (3.3×)**				
Method	Pre ↑	Rec ↑	F1 ↑	IoU ↑	OA ↑	Pre ↑	Rec ↑	F1 ↑	IoU ↑	OA ↑	Pre ↑	Rec ↑	F1 ↑	IoU ↑	OA ↑
FC-EF Daudt et al. (2018)	79.57	71.48	75.31	60.40	97.64	74.25	45.03	56.06	38.95	91.67	–	–	–	–	–
FC-Siam-Conc Daudt et al. (2018)	84.23	69.90	76.40	61.81	97.82	73.11	50.87	59.99	42.85	92.00	–	–	–	–	–
FC-Siam-Diff Daudt et al. (2018)	86.12	60.15	70.83	54.83	97.50	76.10	56.68	64.97	48.12	92.79	–	–	–	–	–
SNUNet Fang et al. (2021)	89.67	81.00	85.11	74.09	98.57	92.61	83.80	87.98	78.55	97.30	38.21	37.16	37.68	23.21	99.27
BIT Chen et al. (2021)	89.57	82.11	85.68	74.94	98.61	97.09	84.80	90.53	82.69	97.91	62.05	33.38	43.41	27.72	99.48
ICIFNet Feng et al. (2022)	87.84	84.62	86.20	75.75	98.63	95.68	90.56	93.05	87.00	98.40	63.50	25.04	35.92	21.89	99.47
DMINet Feng et al. (2023)	89.66	84.28	86.89	76.82	98.72	97.77	89.76	93.60	87.96	98.55	71.47	33.84	45.93	29.81	99.53
SUNet Shao et al. (2021)	64.12	**93.54**	76.08	61.40	97.03	63.55	97.98	77.10	62.73	93.13	32.60	71.00	44.68	28.77	98.96
SRCDNet Liu et al. (2022b)	66.29	84.18	74.17	58.94	97.04	91.30	91.89	91.59	84.49	98.01	39.62	33.22	36.13	22.05	99.30
STANet Chen & Shi (2020)	57.87	45.47	50.93	34.16	95.58	83.06	70.73	76.41	61.82	94.85	11.80	46.89	18.85	10.41	97.61
IFNet Zhang et al. (2020)	86.81	80.85	83.73	72.01	98.41	94.94	79.62	86.61	76.38	97.09	26.20	52.64	34.98	21.20	98.84
CGNet Wu et al. (2020)	–	91.54	89.17	80.46	98.75	–	–	–	–	–	–	63.31	39.93	24.94	99.53
USSFC-Net Lei et al. (2023)	–	90.81	88.26	78.99	**98.97**	–	–	–	–	–	–	72.31	35.72	21.74	**99.69**
CDMamba Zhang et al. (2025)	–	89.87	88.30	79.05	98.66	–	–	–	–	–	–	68.61	47.91	23.39	99.53
C2FNet Han et al. (2024)	–	91.80	89.28	80.64	98.85	–	–	–	–	–	–	68.42	44.16	30.60	99.41
MM-Trans* Xiang et al. (2023)	–	88.17	85.71	75.00	98.33	–	–	–	–	–	–	67.61	53.64	31.65	99.09
SILI* Chen et al. (2023)	**90.55**	86.30	88.38	79.18	98.86	95.29	93.36	94.32	89.24	**98.67**	61.35	42.32	50.10	33.42	99.50
HWA-Net* Yang et al. (2025)	–	92.08	90.64	81.00	98.82	–	–	–	–	–	–	73.76	54.39	35.49	99.55
Ours	90.21	93.42	**91.79**	**84.83**	98.89	**98.01**	**98.03**	**98.02**	**96.12**	**98.67**	**93.52**	**75.52**	**83.56**	**71.76**	99.61

Table 2 presents a matched cross-resolution comparison. Although several baselines were originally designed for same-resolution *CD*, all methods are evaluated using the same HR *before* and LR *after* inputs, dataset splits, degradation factors, and training conditions. On LEVIR-CD ($4\times$), our method achieves the best F1 and IoU. On SV-CD ($8\times$), it provides the strongest overall balance across precision, recall, F1, and IoU. On DE-CD ($3.3\times$), it substantially improves F1 and IoU while remaining competitive in OA. These results demonstrate the effectiveness of the proposed resolution-alignment framework under a consistent evaluation protocol. Additional qualitative and implementation analyses are provided in the appendix. We compare the full method with threshold filtering, retrieval without RL, and RL without the LLM reward. Table 3 shows that the full method achieves the best overall performance, with 95.9% acceptance precision, 85.0% Independent-Evaluator Coverage, 95.0% Independent-Evaluator Class Accuracy, and 77.0% Independent-Evaluator mask IoU. It

also improves the base detector. The evaluation metrics are computed using an independent frozen evaluator G_{eval} that is never used for pseudo-label generation, policy training, reward computation, or detector adaptation. The pseudo-label generator G_{label} and evaluator G_{eval} are therefore strictly separated. The complete eight-variant ablation is provided in the appendix.

Table 3: **End-to-end evaluation of adaptation to classes held out from base-detector training.** Acceptance Precision measures pseudo-label reliability. Coverage, Class Accuracy, and Mask IoU are computed using an independent frozen evaluator that is not used for pseudo-label generation, reward computation, policy training, or detector adaptation. F1 and mIoU report detector performance after adaptation.

Method	Acceptance Precision ↑	Independent-Evaluator Coverage ↑	Independent-Evaluator Class Accuracy ↑	Independent-Evaluator Mask IoU ↑	Adapted F1 ↑	Adapted mIoU ↑
Base detector	–	50.1	65.2	55.3	64.5	52.3
Threshold filtering	84.7	65.0	82.0	62.1	66.5	54.3
Retrieval without RL	89.1	72.2	87.1	67.0	68.5	55.3
RL without LLM reward	92.6	77.1	90.2	71.0	69.5	57.3
Full method	**95.9**	**85.0**	**95.0**	**77.0**	**72.5**	**59.3**

We isolate the contribution of the super-resolution stage through a matched comparison in which all variants use the same inputs, splits, resolution factors, change detector, optimization settings, and training budget; only the resolution-alignment method is changed. We compare direct CD on the LR input, bicubic upsampling, SRGAN, SwinIR (Liang et al., 2021), flow-matching SR, OSEDiff (Wu et al., 2024), and our diffusion-based alignment. SwinIR and OSEDiff outputs are matched to the benchmark-specific HR spatial size before being passed to the same detector.

Table 4 shows that learned resolution alignment consistently improves over direct CD. Our method achieves the best non-oracle results across all three benchmarks and approaches the HR-after oracle on LEVIR-CD and SV-CD. The oracle is not reported for DE-CD because no native HR post-event reference is available under our evaluation protocol.

Table 4: Ablation of the super-resolution stage across three cross-resolution change-detection benchmarks. We compare direct CD, interpolation-based, GAN-based, deterministic, flow-matching, and diffusion-based resolution alignment. An HR-after oracle is reported only when a native high-resolution post-event reference is available. All metrics are percentages (%) and are computed using the same evaluation protocol. Best non-oracle results are shown in bold.

	LEVIR-CD (4×)					SV-CD (8×)					DE-CD (3.3×)				
Method	Pre ↑	Rec ↑	F1 ↑	IoU ↑	OA ↑	Pre ↑	Rec ↑	F1 ↑	IoU ↑	OA ↑	Pre ↑	Rec ↑	F1 ↑	IoU ↑	OA ↑
CD-only	66.34	61.02	63.57	46.59	96.12	70.45	68.33	69.37	53.10	94.02	62.71	49.86	55.55	38.46	97.82
Bicubic+CD	74.28	71.03	72.62	57.01	97.01	82.17	79.45	80.78	67.76	96.35	78.66	60.48	68.38	51.95	98.74
SRGAN+CD	82.64	84.91	83.76	72.06	98.06	91.57	91.86	91.71	84.69	97.88	88.14	69.33	77.62	63.43	99.12
SwinIR+CD	83.92	85.80	84.85	73.69	98.18	92.85	93.45	93.15	87.18	98.05	89.41	71.16	79.25	65.63	99.18
Flow Matching SR+CD	84.39	86.57	85.47	74.63	98.24	93.46	93.74	93.60	87.97	98.11	89.97	71.84	79.89	66.51	99.21
OSEDiff+CD	89.28	92.06	90.64	82.88	98.76	97.18	97.30	97.24	94.63	98.54	92.68	74.19	82.41	70.08	99.54
Diffusion SR+CD (Ours)	**90.21**	**93.42**	**91.79**	**84.83**	**98.89**	**98.01**	**98.03**	**98.02**	**96.12**	**98.67**	**93.52**	**75.52**	**83.56**	**71.76**	**99.61**
HR-after Oracle+CD	91.45	93.88	92.63	86.27	99.04	98.38	98.52	98.45	96.95	98.85	–	–	–	–	–

5 Conclusion and Future Work

We presented a resolution-aligned cross-resolution *CD* framework combining diffusion-based *SR* with RL-guided adaptation to held-out semantic classes. The *SR* module aligns the *LR after* image with the *HR before* image, while the RL policy refines and selects pseudo-labels using geometric, visual, and language-guided rewards. Experiments show improved reconstruction, change detection, held-out-class coverage, mask quality, and post-adaptation performance. Future work will consider temporal sequences, multimodal imagery, and continual adaptation across regions and sensors.

AI USE STATEMENT

In this work, generative AI tools were used for text polishing and assistance with code writing. All AI-assisted text was reviewed and edited by the authors, and all generated or modified code was manually inspected and tested for correctness. The authors take full responsibility for the final content, implementation, results, and claims presented in this work.

ETHICS STATEMENT

This work focuses on remote-sensing change detection using publicly available or research-licensed imagery and benchmark datasets. The proposed framework is intended for scientific analysis of land-cover and structural change and does not rely on personally identifiable information or sensitive individual-level attributes. Human evaluation is used only to assess the semantic validity of retrieved examples and predicted change regions; no sensitive personal data are collected as part of these evaluations.

As with other automated remote-sensing systems, incorrect predictions may lead to misleading interpretations of geographic change. In particular, super-resolution artifacts, pseudo-labeling errors, and domain shifts across sensors or regions may introduce false or missed detections. We therefore report these limitations explicitly and recommend human verification before using model outputs in high-stakes operational settings. The method is intended for research and decision-support applications rather than fully autonomous deployment.

REPRODUCIBILITY STATEMENT

We provide the implementation and evaluation details required to reproduce our results throughout the main paper and appendix. The main paper specifies the full resolution-alignment, change-detection, and policy-guided adaptation pipeline, including the training objectives and evaluation protocol. The appendix provides the diffusion and change-detection hyperparameters, PPO architecture and optimization settings, reward formulation, retrieval and LLM prompting configuration, candidate-generation procedure, and semantic labeling protocol. It also documents dataset preprocessing, resolution factors, held-out-class construction, leakage-control measures, human-evaluation procedures, ablation settings, and theoretical assumptions. Algorithms for policy-guided mask refinement and additional quantitative and qualitative analyses are included to clarify the complete experimental workflow.

REFERENCES

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. IEEE, 2022.

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.

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

Hao Chen, Haotian Zhang, Keyan Chen, Chenyao Zhou, Song Chen, Zhengxia Zou, and Zhenwei Shi. Continuous cross-resolution remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–20, 2023. doi: 10.1109/TGRS.2023.3325829.

Yongqi Chen, Shou Feng, Chunhui Zhao, Yingjie Tang, Junjie Tang, Di Dong, and Nan Su. Deep learning for high-resolution remote sensing change detection: Task taxonomy, datasets, and perspectives. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

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

Runmin Dong, Shuai Yuan, Bin Luo, Mengxuan Chen, Jinxiao Zhang, Lixian Zhang, Weijia Li, Juepeng Zheng, and Haohuan Fu. Building bridges across spatial and temporal resolutions: Reference-based super-resolution via change priors and conditional diffusion model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27674–27684. IEEE, 2024.

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

Yuchao Feng, Honghui Xu, Jiawei Jiang, Hao Liu, and Jianwei Zheng. Icif-net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

Yuchao Feng, Jiawei Jiang, Honghui Xu, and Jianwei Zheng. Change detection on remote sensing images using dual-branch multi-level intertemporal network. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. doi: 10.1109/TGRS.2023.3241257. URL https://doi.org/10.1109/TGRS.2023.3241257.

Chengxi Han, Chen Wu, Meiqi Hu, Jiepan Li, and Hongruixuan Chen. C2f-semicd: A coarse-to-fine semi-supervised change detection method based on consistency regularization in high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–21, 2024.

Paul Hilt, Maedeh Zarvandi, Edgar Kaziakhmedov, Sourabh Bhide, Maria Leptin, Constantin Pape, and Anna Kreshuk. Reinforcement learning for instance segmentation with high-level priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3913–3922, 2023.

Duojun Huang, Xinyu Xiong, Jie Ma, Jichang Li, Zequn Jie, Lin Ma, and Guanbin Li. Alignsam: Aligning segment anything model to open context via reinforcement learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3205–3215. IEEE, 2024.

Weicheng Kuo, AJ Piergiovanni, Dahun Kim, Benjamin Caine, Wei Li, Abhijit Ogale, Luowei Zhou, Andrew M Dai, Zhifeng Chen, Claire Cui, et al. Mammut: A simple architecture for joint learning for multimodal tasks. *Transactions on Machine Learning Research*, 2023.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11286–11315, 2024.

Tao Lei, Xinzhe Geng, Hailong Ning, Zhiyong Lv, Maoguo Gong, Yaochu Jin, and Asoke K Nandi. Ultralightweight spatial–spectral feature cooperation network for change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

Xiang Li, Congcong Wen, Yuan Hu, and Nan Zhou. Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision. *International Journal of Applied Earth Observation and Geoinformation*, 124:103497, 2023.

Ying Li, Haokui Zhang, Xizhe Xue, Yenan Jiang, and Qiang Shen. Deep learning for remote sensing image classification: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(6):e1264, 2018.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF international conference on computer vision workshops (ICCVW)*, pp. 1833–1844. IEEE, 2021.

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

Chenyang Liu, 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.

Mengxi Liu, Qian Shi, Andrea Marinoni, Da He, Xiaoping Liu, and Liangpei Zhang. Super-resolution-based change detection network with stacked attention module for images with different resolutions. *IEEE Transactions on Geoscience and Remote Sensing*, 2022b. doi: 10.1109/TGRS.2021.3091758. URL https://doi.org/10.1109/TGRS.2021.3091758.

Yi Liu, Chao Pang, Zongqian Zhan, Xiaomeng Zhang, and Xue Yang. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geoscience and Remote Sensing Letters*, 18(5):811–815, 2020.

Yunxin Liu, Di Yuan, Zhenghua Xu, Yuefu Zhan, Hongwei Zhang, Jun Lu, and Thomas Lukasiewicz. Pixel level deep reinforcement learning for accurate and robust medical image segmentation. *Scientific Reports*, 15(1):8213, 2025.

D Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Seyedehanita Madani and Vishal M Patel. Morphing through time: Diffusion-based bridging of temporal gaps for robust alignment in change detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 431–439, 2026.

Seyedehanita Madani, Rama Chellappa, and Vishal M Patel. Diffregcd: Integrated registration and change detection with diffusion features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7503–7512, 2026.

Hyeoncheol Noh, Jingi Ju, Minseok Seo, Jongchan Park, and Dong-Geol Choi. Unsupervised change detection based on image reconstruction loss. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1352–1361, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Shaheer U Saeed, Yunguan Fu, Zachary MC Baum, Qianye Yang, Mirabela Rusu, Richard E Fan, Geoffrey A Sonn, Dean C Barratt, and Yipeng Hu. Learning image quality assessment by reinforcing task amenable data selection. In *International conference on information processing in medical imaging*, pp. 755–766. Springer, 2021.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

Ruizhe Shao, Chun Du, Hao Chen, and Jun Li. Sunet: Change detection for heterogeneous remote sensing images from satellite and uav using a dual-channel fully convolution network. *Remote Sensing*, 13(18):3750, 2021.

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.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Juan Tian, Daifeng Peng, Haiyan Guan, and Haiyong Ding. Racdnet: Resolution-and alignment-aware change detection network for optical remote sensing imagery. *Remote Sensing*, 14(18): 4527, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The multi-temporal urban development spacenet dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6398–6407, 2021.

Guoqing Wang, He Chen, Tingting Qiao, Jue Wang, and Wenchao Liu. Resolution-difference embedded network for cross-resolution remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

Chen Wu, Bo Du, and Liangpei Zhang. Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9774–9788, 2023.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37: 92529–92553, 2024.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *Advances in Neural Information Processing Systems*, 38:88167–88190, 2026.

Tianyi Wu, Sheng Tang, Rui Zhang, Juan Cao, and Yongdong Zhang. Cgnet: A light-weight context guided network for semantic segmentation. *IEEE Transactions on Image Processing*, 30:1169–1179, 2020.

Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6254–6264, 2023.

Yunfan Xiang, Xiangyu Tian, Yue Xu, Xiaokun Guan, and Zhengchao Chen. Egmt-cd: Edge-guided multimodal transformers change detection from satellite and aerial images. *Remote Sensing*, 16 (1):86, 2023.

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

Qingsong Xu, Yilei Shi, Jianhua Guo, Chaojun Ouyang, and Xiao Xiang Zhu. Ucdformer: Unsupervised change detection using a transformer-driven image translation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023.

Hualin Yang, Boran Ren, Zhijun Yang, Jing Xiong, Xiying Li, and Yu-Chian Calvin. Hwa-net: Hierarchical window aggregate network for cross-resolution remote sensing change detection. *Expert Systems with Applications*, pp. 129829, 2025.

Yanan You, Jingyi Cao, and Wenli Zhou. A survey of change detection methods based on remote sensing images for multi-source and multi-objective scenarios. *Remote Sensing*, 12(15):2460, 2020.

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

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

Xiangtao Zheng, Xiumei Chen, Xiaoqiang Lu, and Bangyong Sun. Unsupervised change detection by cross-resolution difference learning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021. doi: 10.1109/TGRS.2021.3079907.

Zhuo Zheng, Yanfei Zhong, Liangpei Zhang, and Stefano Ermon. Segment any change. *Advances in Neural Information Processing Systems*, 37:81204–81224, 2024.

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017.

A Inference-Time Policy Rollout

At inference time, our framework executes the learned policy $\pi_\theta(a \mid s)$ to refine candidate masks through sequential interaction rather than one-shot filtering. Starting from an initial candidate mask $m_i^{(0)} \in \mathcal{M}_{\text{cand}}$, the agent repeatedly observes the current state s_t, selects an action $a_t \sim \pi_\theta(\cdot \mid s_t)$, and transitions to a new state s_{t+1} until a terminal `accept`/`reject` action is emitted or a maximum horizon T is reached. This allows the policy to refine candidates through reward-guided rollout without updating the policy parameters at inference time.

A.1 Inference-Time RL Formulation

Each inference episode is defined over a finite-horizon Markov decision process. The state s_t contains the image pair, the current candidate mask $m_i^{(t)}$, the retrieved support set $\mathcal{R}(m_i^{(t)})$, and the action history. The action space includes local mask-refinement operations together with terminal `accept` and `reject` actions.

We first define the retrieval-augmented mask-quality score

$$Q(m_i^{(t)}) = \gamma\, \mathrm{IoU}(m_i^{(t)}, D) + \alpha\, f_{\text{img}}(m_i^{(t)}) + \beta\, f_{\text{text}}\left(m_i^{(t)} \mid \mathcal{R}(m_i^{(t)})\right), \tag{20}$$

where γ, α, and β weight geometric consistency, image-text alignment, and language-guided semantic feedback, respectively.

Rather than using the absolute quality score as a per-step reward, we reward improvements produced by refinement actions and explicitly define rewards for terminal decisions:

$$r_t = \begin{cases} Q(m_i^{(t+1)}) - Q(m_i^{(t)}) - \lambda_{\text{step}}, & a_t \in \mathcal{A}_{\text{refine}}, \\ Q(m_i^{(t)}) - \tau_{\text{acc}}, & a_t = \text{accept}, \\ \tau_{\text{acc}} - Q(m_i^{(t)}), & a_t = \text{reject}, \end{cases} \tag{21}$$

where

$$\mathcal{A}_{\text{refine}} = \{\text{expand}, \text{contract}, \text{shift}_x, \text{shift}_y, \text{noop}\}. \tag{22}$$

The step penalty $\lambda_{\text{step}} > 0$ discourages unnecessarily long trajectories, while τ_{acc} defines the quality threshold that governs terminal acceptance and rejection. In particular, a no-op action receives the negative step reward $-\lambda_{\text{step}}$.

The learned policy selects actions that maximize the expected discounted return

$$J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1} \gamma_{\text{disc}}^t r_t\right], \tag{23}$$

where $\tau = (s_0, a_0, r_0, \ldots, s_T)$ denotes a refinement trajectory and γ_{disc} is the discount factor. An episode terminates when the policy selects `accept` or `reject`, or when the maximum horizon is reached. Candidates that reach the horizon without acceptance are not added to the pseudo-label set.

A.2 Policy Rollout at Inference

During inference, the learned policy is executed autoregressively over candidate masks. Given state s_t, the agent selects an action

$$a_t \sim \pi_\theta(\cdot \mid s_t), \tag{24}$$

applies the corresponding mask transformation or terminal decision, and transitions to the next state according to the environment dynamics. If the chosen action is `accept`, the final mask is added to the accepted set; if it is `reject`, the candidate is discarded. Formally, the accepted set is

$$\mathcal{M}_{\text{good}} = \left\{ m_i^{(T_i)} \;\middle|\; \text{episode } i \text{ terminates with } \texttt{accept} \right\}, \tag{25}$$

where T_i denotes the timestep at which the terminal action is selected for episode i.

A.3 Practical Approximation

To reduce inference cost, we restrict the horizon to a small number of refinement steps and initialize episodes only from candidates that have already passed geometric and image-text pre-filtering. This preserves the sequential decision-making structure while keeping rollout tractable in practice. Compared with one-shot thresholding, inference under the learned policy enables multi-step correction of candidate masks before terminal decisions are made, leading to more reliable acceptance of valid unseen semantic changes.

A.4 Accepted Trajectories and Adaptation Set

After policy rollout, each candidate episode terminates either with an `accept` action, a `reject` action, or at the maximum horizon. The accepted terminal masks define the adaptation set

$$\mathcal{D}_{\text{new}} = \{(s^{(i)}_{0:T_i}, a^{(i)}_{0:T_i}, r^{(i)}_{0:T_i}, m^{(T_i)}_i) \mid a^{(i)}_{T_i} = \texttt{accept}\}, \tag{26}$$

where T_i denotes the timestep at which the terminal `accept` action is selected. This dataset contains the full refinement trajectory together with the accepted terminal mask, and therefore reflects the behavior of the learned policy under the retrieval-augmented reward model.

In practice, the terminal masks extracted from $\mathcal{D}_{\text{new}}$ define the high-confidence supervision set used for downstream adaptation of the change detector:

$$\mathcal{M}_{\text{good}} = \{m^{(T_i)}_i \mid (s^{(i)}_{0:T_i}, a^{(i)}_{0:T_i}, r^{(i)}_{0:T_i}, m^{(T_i)}_i) \in \mathcal{D}_{\text{new}}\}. \tag{27}$$

Unlike one-step reward thresholding, this construction preserves the sequential decision process that produced each accepted mask and ties the adaptation set directly to policy rollouts rather than to a static filtering operator.

The full policy-guided refinement pipeline is illustrated in Fig. 5: the learned policy interacts with candidate masks over multiple steps, the retrieval-augmented judge provides reward feedback at each step, and terminal accepted masks are collected into $\mathcal{D}_{\text{new}}$ for detector adaptation.

B Ablation of Policy Learning, Retrieval, and Reward Components

A central question is whether the adaptation stage truly requires sequential decision making, or whether similar accepted-mask quality can be obtained with a simpler static filter. Our method is formulated as a finite-horizon policy that iteratively refines each candidate mask and then emits a terminal `accept` or `reject` action under a retrieval-augmented reward model. This design is motivated by the observation that many candidate regions are not purely correct or incorrect at initialization: some require local refinement before a reliable terminal decision can be made. Accordingly, the full method combines three ingredients: *(i)* multi-step policy rollout, *(ii)* retrieval of semantically similar support examples, and *(iii)* a composite reward integrating geometric consistency, image-text alignment, and frozen LLM-based semantic feedback.

Ablation design. To isolate the contribution of each component, we compare the full method against the following alternatives:

1. **Threshold filtering only (No RL):** candidates are accepted or rejected directly using a fixed score threshold without sequential refinement.
2. **One-shot scorer:** a learned single-step scorer predicts `accept`/`reject` from the initial candidate state.
3. **Retrieval without RL:** retrieved examples are used for scoring, but no sequential refinement is performed.
4. **LLM judge without RL:** the frozen LLM-based judge is applied once to the initial candidate without policy rollout.
5. **RL without retrieval:** the policy refines masks sequentially, but retrieval context is removed from the state and reward.

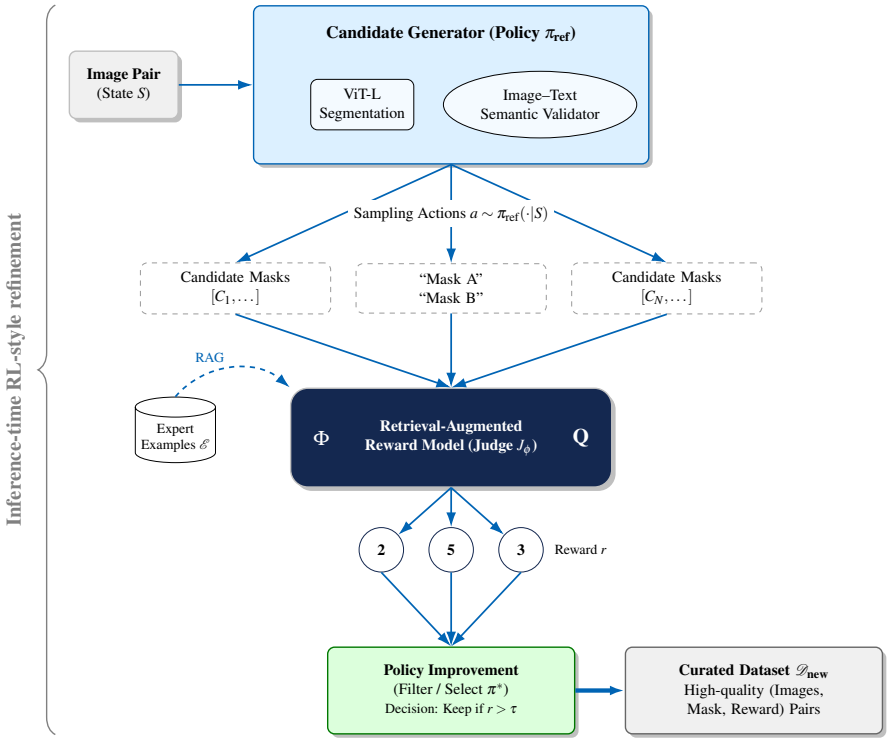

Figure 5: Illustration of our policy-guided refinement process. Starting from an initial candidate mask for a given image pair (state s), the learned policy interacts with the environment over multiple steps by proposing mask refinements or terminal `accept`/`reject` actions. At each step, a retrieval-augmented reward model (judge $\mathcal{J}_\Phi$), assisted by retrieved support examples $\mathcal{E}$, assigns reward based on semantic-visual alignment and geometric consistency. Episodes that terminate with an `accept` action produce terminal masks that are added to the adaptation set $\mathcal{D}_{\text{new}}$. This procedure implements sequential policy execution under reward feedback and yields high-confidence masks for downstream detector adaptation.

6. **RL without LLM reward:** the policy uses only geometric and image-text reward terms.
7. **RL without image-text reward:** the policy uses only geometric and frozen LLM-based reward terms.
8. **Full method:** sequential policy rollout with retrieval-augmented reward, including geometric, image-text, and frozen LLM-based components.

Evaluation protocol and metrics. Because the purpose of this stage is to identify *valid discovered instances* rather than optimize dense pixel overlap, we evaluate accepted terminal masks using the same human-verified instance-level protocol as in the novel-class discovery analysis. Specifically, from the set of accepted masks, human annotators determine whether each accepted region corresponds to a valid unseen semantic change or to a false acceptance such as a noisy artifact, boundary failure, or semantically incorrect region.

Let TP_{inst} denote accepted masks judged valid by human annotators, and let FP_{inst} denote accepted masks judged invalid. We report acceptance precision, defined as Acceptance Precision $= \frac{TP_{\text{inst}}}{TP_{\text{inst}}+FP_{\text{inst}}}$, and false discovery rate, defined as FDR $= \frac{FP_{\text{inst}}}{TP_{\text{inst}}+FP_{\text{inst}}} = 1 -$ Acceptance Precision.

This evaluation is better aligned with the goal of the adaptation stage, since the accepted masks are exactly the pseudo-labels used to construct the adaptation set for downstream detector update. A high acceptance precision indicates that the selected pseudo-labels are semantically reliable, while a low FDR indicates that only limited noise is introduced into adaptation.

Results and interpretation. Table 5 shows that the full method achieves the best accepted-mask quality, reaching an acceptance precision of 95.9% and the lowest FDR of 4.1%. Static alternatives such as threshold filtering and one-shot scoring are clearly weaker, indicating that sequential refinement before terminal acceptance is important. This supports the claim that many candidates benefit from iterative correction and should not be decided from the initial state alone.

Retrieval also provides a clear benefit: removing it reduces acceptance precision, showing that contextual support examples contribute useful semantic evidence beyond local geometry alone. Likewise, removing either the LLM-based reward or the image-text reward degrades performance, indicating that these two signals are complementary. The image-text term helps maintain visual-semantic consistency, whereas the frozen LLM-based judge contributes higher-level semantic validation. Overall, the results show that the method is not merely a complicated thresholding scheme. Rather, the combination of sequential policy rollout, retrieval context, and complementary reward signals yields the cleanest adaptation set for downstream detector update.

Table 5: Ablation of the adaptation stage under the human-verified instance-level evaluation protocol. Acceptance Precision measures the fraction of accepted masks judged to be valid unseen semantic changes, while False Discovery Rate (FDR) measures the fraction of accepted masks judged invalid. Higher Acceptance Precision and lower FDR are better. The full method performs best, showing that sequential policy rollout, retrieval context, and both semantic reward terms are all important.

Variant	Acceptance Precision ↑	FDR ↓
Threshold filtering only (No RL)	84.7	15.3
One-shot scorer	87.3	12.7
Retrieval without RL	89.1	10.9
LLM judge without RL	88.4	11.6
RL without retrieval	91.2	8.8
RL without LLM reward	92.6	7.4
RL without image-text reward	93.1	6.9
Full method	**95.9**	**4.1**

B.1 Seen-Class Retention After Adaptation

Introducing logits for held-out classes can alter the final prediction even when the parameters of C_{base} remain frozen. We therefore explicitly evaluate retention of the previously learned semantic classes before and after adaptation. For this analysis, evaluation is restricted to the seen-class subset, and the same evaluation protocol is applied to the base and adapted detectors.

Table 6: Seen-class retention before and after held-out-class adaptation. Δ denotes the performance change after adaptation (After $-$ Before), where values closer to zero indicate lower seen-class forgetting. All metrics are reported as percentages (%).

Dataset	F1 Before	F1 After	ΔF1	mIoU Before	mIoU After	ΔmIoU
LEVIR-CD	91.42	91.08	-0.34	84.21	83.74	-0.47
SV-CD	97.86	97.71	-0.15	95.91	95.63	-0.28
DE-CD	83.18	82.63	-0.55	71.24	70.47	-0.77

As shown in Table 6, adaptation introduces only a small degradation on previously learned classes. Across the three benchmarks, the average decrease is 0.35 F1 points and 0.51 mIoU points. SV-CD shows the smallest change, with reductions of only 0.15 F1 and 0.28 mIoU, while the more challenging cross-sensor DE-CD setting exhibits the largest drop of 0.55 F1 and 0.77 mIoU. These results indicate that optimizing only the newly introduced adaptation head produces limited interference with the previously learned semantic classes, although freezing C_{base} does not mathematically guarantee unchanged joint predictions.

C Theoretical Justification for Stage Composition

We provide a formal analysis explaining why composing super-resolution (SR) and change detection (CD) is theoretically sound, even though the two stages are not jointly optimized.

Problem Setup. Let $\mathcal{X}_{\mathrm{HR}}$ and $\mathcal{X}_{\mathrm{LR}}$ denote the high- and low-resolution image domains. We define the downsampling operator

$$D : \mathcal{X}_{\mathrm{HR}} \to \mathcal{X}_{\mathrm{LR}}, \tag{28}$$

and the super-resolution operator

$$S : \mathcal{X}_{\mathrm{LR}} \to \mathcal{X}_{\mathrm{HR}}. \tag{29}$$

The change detector is defined as

$$C : \mathcal{X}_{\mathrm{HR}} \times \mathcal{X}_{\mathrm{HR}} \to \mathcal{Y}, \tag{30}$$

where $\mathcal{Y}$ denotes the space of change masks.

In our setting, the before image $x_b \in \mathcal{X}_{\mathrm{HR}}$ is observed at high resolution, while the after image $x_a \in \mathcal{X}_{\mathrm{LR}}$ is low resolution. We assume the existence of a latent high-resolution after image $x_a^* \in \mathcal{X}_{\mathrm{HR}}$ such that

$$x_a = D(x_a^*). \tag{31}$$

Our inference pipeline is therefore

$$\hat{y} = C(x_b, S(x_a)). \tag{32}$$

—

Equivariance Assumption. Following our main analysis, we assume that the change detector is equivariant to resolution-preserving operators L, i.e.,

$$C(L(x_b), L(x_a^*)) = C(x_b, x_a^*). \tag{33}$$

This implies that applying a consistent transformation to both inputs does not alter the prediction, formalizing the intuition that CD depends on relative semantic differences rather than absolute resolution.

—

Bounded SR Reconstruction Error Assume the SR model achieves bounded reconstruction error on the test distribution:

$$\|S(x_a) - x_a^*\| \le \delta, \tag{34}$$

for some $\delta > 0$.

This assumption reflects empirical SR fidelity.

—

Local Stability of Change Detection Assume C is locally Lipschitz in its second argument in a neighborhood of x_a^*, i.e.,

$$\|C(x_b, u) - C(x_b, v)\| \le L\|u - v\|, \tag{35}$$

for all $u, v \in \mathcal{B}(x_a^*, r)$.

This models the smooth behavior of neural detectors under small perturbations.

—

theorem Composition Error Bound Under Lemmas 1–2, the composed prediction satisfies

$$\|C(x_b, S(x_a)) - C(x_b, x_a^*)\| \le L\delta. \tag{36}$$

Proof. By Lemma 2,

$$\|C(x_b, S(x_a)) - C(x_b, x_a^*)\| \le L\|S(x_a) - x_a^*\|. \tag{37}$$

Applying Lemma 1 yields

$$\|C(x_b, S(x_a)) - C(x_b, x_a^*)\| \le L\delta. \tag{38}$$

□

—

Interpretation. The theorem shows that the error induced by inserting SR before CD is bounded by the SR reconstruction error. Consequently, if S produces faithful reconstructions, the composed pipeline approximates the ideal HR-to-HR detector.

—

Connection to Equivariance. If SR were perfect ($\delta = 0$), then

$$C(x_b, S(x_a)) = C(x_b, x_a^*), \tag{39}$$

which, by equivariance, is equivalent to applying SR to both inputs. This formally justifies our design choice of enhancing only the after image.

—

Conclusion. Although SR and CD are trained independently, their composition is theoretically justified under mild reconstruction and stability assumptions. This explains the consistent empirical gains observed in Table 4.

D Policy Network and PPO-Based Mask Refinement

The adaptation module uses a lightweight policy to sequentially refine candidate masks before making a terminal `accept`/`reject` decision. For candidate $m_i^{(t)}$, the state is

$$s_i^{(t)} = \left(I_b, \hat{I}_a, m_i^{(t)}, \mathcal{R}(m_i^{(t)}), h_i^{(t)} \right), \tag{40}$$

where $\mathcal{R}(m_i^{(t)})$ contains retrieved support examples and $h_i^{(t)}$ records the action history.

The image pair, current mask, retrieval context, and history are encoded and fused into

$$z_i^{(t)} = \mathrm{Fuse}\Big(E_{\mathrm{vis}}([I_b, \hat{I}_a, m_i^{(t)}]), E_{\mathrm{ret}}(\mathcal{R}(m_i^{(t)})), E_{\mathrm{hist}}(h_i^{(t)}) \Big), \tag{41}$$

from which the policy predicts

$$\pi_\theta(a_i^{(t)} \mid s_i^{(t)}) = \mathrm{Softmax}\Big(W_\pi z_i^{(t)} + b_\pi \Big). \tag{42}$$

The action space is

$$\mathcal{A} = \{\mathrm{expand}, \mathrm{contract}, \mathrm{shift}_x, \mathrm{shift}_y, \mathrm{noop}, \mathrm{accept}, \mathrm{reject}\}, \tag{43}$$

where refinement actions update the mask through

$$m_i^{(t+1)} = \mathcal{T}(m_i^{(t)}, a_i^{(t)}), \tag{44}$$

while `accept` and `reject` terminate the episode.

Reward formulation. We first define the retrieval-augmented mask-quality score

$$Q(m_i^{(t)}) = \gamma\, \mathrm{IoU}(m_i^{(t)}, D) + \alpha\, f_{\mathrm{img}}(m_i^{(t)}) + \beta\, f_{\mathrm{text}}\left(m_i^{(t)} \mid \mathcal{R}(m_i^{(t)}) \right), \tag{45}$$

where D is the semantic difference mask, f_{img} measures image-text consistency, and f_{text} is the frozen retrieval-augmented LLM judge score.

The reward is defined by improvement in candidate quality together with explicit terminal rewards:

$$r_i^{(t)} = \begin{cases} Q(m_i^{(t+1)}) - Q(m_i^{(t)}) - \lambda_{\mathrm{step}}, & a_i^{(t)} \in \mathcal{A}_{\mathrm{refine}}, \\ Q(m_i^{(t)}) - \tau_{\mathrm{acc}}, & a_i^{(t)} = \mathrm{accept}, \\ \tau_{\mathrm{acc}} - Q(m_i^{(t)}), & a_i^{(t)} = \mathrm{reject}, \end{cases} \tag{46}$$

with

$$\mathcal{A}_{\mathrm{refine}} = \{\mathrm{expand}, \mathrm{contract}, \mathrm{shift}_x, \mathrm{shift}_y, \mathrm{noop}\}. \tag{47}$$

The step penalty $\lambda_{\mathrm{step}} > 0$ discourages unnecessarily long trajectories, while τ_{acc} controls terminal acceptance. The LLM judge remains frozen throughout policy optimization.

PPO optimization. The policy is trained using the clipped PPO objective

$$\mathcal{L}_{\mathrm{PPO}} = -\mathbb{E}_t \left[\min \left(\rho_t(\theta) \hat{A}_t, \mathrm{clip} \left(\rho_t(\theta), 1-\epsilon, 1+\epsilon \right) \hat{A}_t \right) \right], \tag{48}$$

where

$$\rho_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t \mid s_t)}. \tag{49}$$

We additionally optimize the value function and use entropy regularization:

$$\mathcal{L} = \mathcal{L}_{\mathrm{PPO}} + \lambda_v \left(V_\theta(s_t) - \hat{R}_t \right)^2 - \lambda_H \mathcal{H} \left(\pi_\theta(\cdot \mid s_t) \right). \tag{50}$$

Algorithm 1 summarizes the complete refinement and terminal-selection procedure.

Algorithm 1 PPO-Guided Mask Refinement for Held-Out Semantic Changes

Input: I_b, $\hat{I}_a$, $\mathcal{M}_{\mathrm{cand}}$, D, $\mathcal{D}_{\mathrm{ret}}$, horizon T
Output: Accepted masks $\mathcal{M}_{\mathrm{good}}$
Initialize $\mathcal{M}_{\mathrm{good}} \leftarrow \emptyset$
for $m_i^{(0)} \in \mathcal{M}_{\mathrm{cand}}$ **do**
  $h_i^{(0)} \leftarrow \emptyset$
  for $t = 0$ to $T-1$ **do**
    Retrieve $\mathcal{R}_i^{(t)}$ and construct $s_i^{(t)} = (I_b, \hat{I}_a, m_i^{(t)}, \mathcal{R}_i^{(t)}, h_i^{(t)})$
    Sample $a_i^{(t)} \sim \pi_\theta(\cdot \mid s_i^{(t)})$
    Compute

$$Q(m_i^{(t)}) = \gamma \, \mathrm{IoU}(m_i^{(t)}, D) + \alpha f_{\mathrm{img}}(m_i^{(t)}) + \beta f_{\mathrm{text}}(m_i^{(t)} \mid \mathcal{R}_i^{(t)})$$

    if $a_i^{(t)} = \mathrm{accept}$ **then**
      $r_i^{(t)} \leftarrow Q(m_i^{(t)}) - \tau_{\mathrm{acc}}$
      $\mathcal{M}_{\mathrm{good}} \leftarrow \mathcal{M}_{\mathrm{good}} \cup \{m_i^{(t)}\}$
      break
    else if $a_i^{(t)} = \mathrm{reject}$ **then**
      $r_i^{(t)} \leftarrow \tau_{\mathrm{acc}} - Q(m_i^{(t)})$
      break
    else
      $m_i^{(t+1)} \leftarrow \mathcal{T}(m_i^{(t)}, a_i^{(t)})$
      Compute $Q(m_i^{(t+1)})$
      $r_i^{(t)} \leftarrow Q(m_i^{(t+1)}) - Q(m_i^{(t)}) - \lambda_{\mathrm{step}}$
      $h_i^{(t+1)} \leftarrow h_i^{(t)} \cup \{a_i^{(t)}\}$
    end if
  end for
end for
Estimate advantages $\hat{A}_t$ from the collected trajectories
Update π_θ and V_θ using PPO
return $\mathcal{M}_{\mathrm{good}}$

E Model Efficiency

To assess the computational footprint of our approach, we compare the number of trainable parameters, floating-point operations (FLOPs), and training time per epoch with representative change detection baselines. All models are evaluated on the LEVIR-CD training set, where the input image resolution is fixed to $256 \times 256 \times 3$ and the batch size is set to 8. The results are summarized in Table 7.

Our model is highly compact, containing only 4.26M parameters and requiring 14.83G FLOPs, which is substantially lower than most CNN-based and Transformer-based competitors. Despite its

lightweight design, the training time per epoch remains competitive compared to larger architectures, while still delivering strong detection accuracy across benchmarks. In contrast, models such as IFNet, ICIFNet, and DMINet introduce significantly higher computational cost due to deeper backbones and attention-heavy feature fusion modules.

These results indicate that the proposed framework strikes a favorable balance between performance and efficiency. The compact architecture enables faster training and reduced memory consumption, making the model suitable for practical deployment scenarios and large-scale cross-resolution change detection tasks.

Table 7: Comparison of model efficiency. We report the number of model parameters (Params.), floating-point operations per second (FLOPs), and training time for one epoch on the LEVIR-CD training set. The input image to the model has a size of $256 \times 256 \times 3$. The batch size is set to 8.

Model	Params.(M)	FLOPs (G)	Training Time (s)
FC-EF Daudt et al. (2018)	1.35	7.14	52.33
FC-Siam-conc Daudt et al. (2018)	1.55	10.64	64.99
FC-Siam-diff Daudt et al. (2018)	1.35	9.44	63.21
STANet Chen & Shi (2020)	16.93	26.32	270.74
IFNet Zhang et al. (2020)	50.71	164.70	372.66
SNUNet Fang et al. (2021)	12.03	109.76	305.82
BIT Chen et al. (2021)	3.55	17.42	140.46
ICIFNet Feng et al. (2022)	25.83	50.50	555.37
DMINet Feng et al. (2023)	6.76	28.38	172.29
SUNet Shao et al. (2021)	15.56	79.78	321.31
SRCDNet Liu et al. (2022b)	12.77	30.98	81.86
ChangeFromer Bandara & Patel (2022)	13.06	17.5	103.81
Ours	4.26	14.83	258.6

F Implementation Details

Diffusion-Based Super-Resolution. The super-resolution module uses a conditional diffusion backbone with `UNetSuperRes` (`base_ch`= 64, `time_dim`= 256) and is trained on 16× NVIDIA A100 GPUs. At inference, we use 200 reverse-diffusion steps with batch size 32.

The training objective is

$$\mathcal{L}_{\mathrm{SR}} = \lambda_1 \mathcal{L}_{\mathrm{noise}} + \lambda_2 \mathcal{L}_{\mathrm{rec}} + \lambda_3 \mathcal{L}_{\mathrm{perc}}, \tag{51}$$

with $\lambda_1 = 1.0$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.01$, selected using the validation set. We use a linear noise schedule with β_t increasing from 10^{-4} to 0.02 over $T_{\mathrm{diff}} = 1000$ timesteps. Training is performed for 500 epochs using AdamW with learning rate 1×10^{-4}, weight decay 0.01, and coefficients $(0.9, 0.999)$.

For the synthetic $4\times$ setting, LR observations are generated by bicubic downsampling of the HR target followed by zero-mean Gaussian noise with $\sigma \sim \mathcal{U}(0, 0.05)$. Dataset-specific resolution factors and preprocessing are described in the dataset setup.

Resolution-Aligned Change Detection. The CD model is trained for 400 epochs using AdamW with $\eta = 5 \times 10^{-5}$, coefficients $(0.9, 0.99)$, and weight decay 10^{-6}. We use batch size 16, random horizontal and vertical flips, and cosine learning-rate scheduling with $T_{\max} = 50$.

The multi-scale prediction heads are optimized using

$$\mathcal{L}_{\mathrm{CD}} = 0.6\, \mathcal{L}_{\mathrm{focal}} + 0.4\, \mathcal{L}_{\mathrm{Dice}}, \tag{52}$$

where the loss weights are selected on the validation set.

Policy-Guided Discovery. Candidate masks are generated by a transformer-based segmentation model with a ViT/L encoder and ViT-Lite decoder. Retrieval is region-adaptive: large connected components use $k = 3$, while smaller candidates use $k = 5$. Retrieved examples provide semantic context to the policy and reward model; final pseudo-label acceptance is determined by the policy's terminal `accept`/`reject` action.

Semantic Labeler. The frozen semantic labeler G_{label} consists of a ViT encoder pretrained with masked autoencoding and a ViT-based decoder. During semantic fine-tuning, the encoder remains frozen and the decoder is trained for 120 epochs with learning rate 1×10^{-5} using a focal–Dice loss.

Training data are drawn from DLRSD, SpaceNet, Inria Aerial Image Labeling, Massachusetts Buildings and Roads, and AIRS, with their annotations mapped to the unified taxonomy $\mathcal{C}_{\text{all}}$. After fine-tuning, G_{label} is frozen and assigns a semantic pseudo-label $\hat{y}_i$ to each accepted mask $m_i \in \mathcal{M}_{\text{good}}$. Only the new-class head ΔC is then optimized using $(m_i, \hat{y}_i)$, while the shared encoder and C_{base} remain frozen.

Independent evaluation model. Pseudo-label generation and evaluation use separate frozen models. G_{label} constructs the adaptation labels, whereas G_{eval} is trained independently and used only to compute Coverage, Class Accuracy, and Mask IoU. G_{eval} is never used for candidate generation, retrieval, reward computation, pseudo-label assignment, policy training, or detector adaptation, and no target adaptation masks are used in its training.

F.1 LLM-Based Reward Model

We use frozen **Gemma 3 4B** Team et al. (2025) as the language-based semantic component of the reward model. The model operates only at inference and is never fine-tuned.

At each policy step, the LLM receives a description of the before/after pair, the current candidate region, and the region-adaptive retrieved support examples. The prompt is:

```
 You are given a pair of satellite images (before and after)
and a highlighted region indicating a potential change.
Determine whether this region corresponds to a meaningful
semantic change (e.g., construction, new object, removal) or
a spurious artifact (e.g., noise, shadow, misalignment).
You are also given similar examples of valid and invalid
changes for reference.
Answer with a single word:  VALID or INVALID.
```

The response is mapped to $f_{\text{text}}(m^{(t)}) \in \{0, 1\}$, with `VALID`$= 1$ and `INVALID`$= 0$. We use greedy decoding with temperature 0, random seed 42, and the region-adaptive retrieval setting $k \in \{3, 5\}$ described above.

All LLM inference is performed locally on an A100 GPU. Because decoding is greedy and the model is frozen, the semantic reward is deterministic under the fixed inputs and retrieval set.

To quantify the contribution of the language signal, we additionally evaluate a variant without f_{text}, leaving only geometric and image-text reward components. Removing the LLM reward reduces acceptance precision, indicating that language-guided validation provides complementary semantic information.

G Equivalence of Super Resolution Placement

Let $\mathcal{X}_{\text{HR}}$ and $\mathcal{X}_{\text{LR}}$ denote the high-resolution and low-resolution image domains, respectively. We denote by

$$D : \mathcal{X}_{\text{HR}} \to \mathcal{X}_{\text{LR}}, \qquad S : \mathcal{X}_{\text{LR}} \to \mathcal{X}_{\text{HR}} \tag{53}$$

a downsampling operator D and a super-resolution operator S. In our setting, the before image $x_b \in \mathcal{X}_{\text{HR}}$ is high resolution, while the after image $x_a \in \mathcal{X}_{\text{LR}}$ is low resolution. A change detection model is a mapping

$$C : \mathcal{X}_{\text{HR}} \times \mathcal{X}_{\text{HR}} \to \mathcal{Y}, \tag{54}$$

where $\mathcal{Y}$ is the space of change predictions (for example, pixel-wise change masks).

We assume that S approximately inverts D on the support of the data, that is,

$$S \circ D(x_b) = x_b + \varepsilon(x_b), \qquad \|\varepsilon(x_b)\| \ll \|x_b\|, \tag{55}$$

and that C is equivariant with respect to a shared post-processing operator $L : \mathcal{X}_{\text{HR}} \to \mathcal{X}_{\text{HR}}$ in the sense that

$$C(L(x_b), L(x_a)) = C(x_b, x_a), \qquad \forall x_b, x_a \in \mathcal{X}_{\text{HR}}. \tag{56}$$

Under assumptions equation 55 and equation 56, performing super resolution on the low-resolution after image x_a only is equivalent, up to the reconstruction error $\varepsilon(x_b)$, to a formulation where

both before and after images are obtained by applying the same super-resolution operator to low-resolution inputs. Consequently, without loss of generality we can model super resolution as acting solely on the after image.

Consider the following two formulations.

(i) Super resolution on the after image only. We compare the original high-resolution before image with the super-resolved after image:

$$y_{\text{after-SR}} = C\big(x_b,\, S(x_a)\big). \tag{57}$$

(ii) Symmetric super resolution on both images. We downsample the high-resolution before image, then super-resolve both images:

$$\tilde{x}_b = S\big(D(x_b)\big), \tilde{x}_a = S(x_a), y_{\text{both-SR}} = C\big(\tilde{x}_b,\, \tilde{x}_a\big). \tag{58}$$

Using equation 55, we write $\tilde{x}_b = x_b + \varepsilon(x_b)$. Applying equation 56 with $L = \mathrm{Id} + D_\varepsilon$ gives

$$\begin{aligned} C(\tilde{x}_b, \tilde{x}_a) &= C\big(x_b + \varepsilon(x_b), S(x_a)\big) \\ &\approx C\big(x_b, S(x_a)\big) \\ &= y_{\text{after-SR}}. \end{aligned} \tag{59}$$

Thus, the predictions produced when super resolution is applied symmetrically to both images coincide, up to a small reconstruction error, with the predictions obtained when super resolution is applied only to the low-resolution after image. This justifies treating the latter formulation as representative without loss of generality.

H Why Stage-wise Optimization Can Outperform Joint Training

Let the input be $x \in \mathcal{X}$ with ground-truth change label $y \in \mathcal{Y}$, and optionally a high-resolution target $x^\star \in \mathcal{X}_{\text{HR}}$. Our pipeline consists of a super-resolution module $g_\theta : \mathcal{X} \to \mathcal{X}_{\text{HR}}$ and a change detector $f_\phi : \mathcal{X}_{\text{HR}} \times \mathcal{X}_{\text{HR}} \to \mathcal{Y}$.

Objectives. A natural *joint* objective is

$$\min_{\theta,\phi}\ \mathbb{E}\big[\mathcal{L}_{\text{cd}}(f_\phi(g_\theta(x_1), g_\theta(x_2)), y)\big]\ +\ \lambda\, \mathbb{E}\big[\mathcal{L}_{\text{sr}}(g_\theta(x), x^\star)\big], \tag{60}$$

where $\mathcal{L}_{\text{cd}}$ is the task loss and $\mathcal{L}_{\text{sr}}$ is a reconstruction/perceptual SR loss (when $x^\star$ exists). In contrast, a *stage-wise* approach solves

$$\begin{aligned} \theta^\star &\in \arg\min_\theta\ \mathbb{E}\big[\mathcal{L}_{\text{sr}}\big(g_\theta(x), x^\star\big)\big]\,, \\ \phi^\star &\in \arg\min_\phi\ \mathbb{E}\big[\mathcal{L}_{\text{cd}}\big(f_\phi(g_{\theta^\star}(x_1), g_{\theta^\star}(x_2)), y\big)\big]\,. \end{aligned} \tag{61}$$

Key difficulty: non-identifiability of g_θ under $\mathcal{L}_{\text{cd}}$. If supervision is only available for y (common in change detection), then the term $\mathbb{E}[\mathcal{L}_{\text{cd}}(f_\phi(g_\theta(\cdot), g_\theta(\cdot)), y)]$ does not uniquely constrain g_θ to be a faithful SR mapping. In particular, g_θ can drift toward *task-hack* representations that help f_ϕ while distorting photometric fidelity or introducing hallucinated textures.

Task-equivalent SR mappings Assume f_ϕ has sufficient capacity. For any mapping g_θ, there exists ϕ such that

$$\begin{aligned} &\mathbb{E}\big[\mathcal{L}_{\text{cd}}\big(f_\phi(g_\theta(x_1), g_\theta(x_2)), y\big)\big] \\ &\quad\approx \inf_{\tilde{\phi}}\ \mathbb{E}\Big[\mathcal{L}_{\text{cd}}\big(f_{\tilde{\phi}}(x_1', x_2'), y\big)\Big]\,. \end{aligned} \tag{62}$$

where (x_1', x_2') denotes any intermediate representation of the pair.

This highlights that minimizing $\mathcal{L}_{\text{cd}}$ alone can leave θ underconstrained; different g_θ can yield similar task loss, including ones that degrade visual realism or generalization.

Gradient interference between $\mathcal{L}_{\text{sr}}$ and $\mathcal{L}_{\text{cd}}$. Even when $\mathcal{L}_{\text{sr}}$ is present, joint training can suffer from competing gradients. Let

$$\nabla_\theta \mathcal{J}(\theta, \phi) = \nabla_\theta \, \mathbb{E}\big[\mathcal{L}_{\text{cd}}\big] + \lambda \, \nabla_\theta \, \mathbb{E}\big[\mathcal{L}_{\text{sr}}\big]. \tag{63}$$

If the two terms point in opposing directions, updates that improve SR fidelity may degrade the CD objective (and vice versa).

Destructive gradient interference If

$$\langle \nabla_\theta \mathbb{E}[\mathcal{L}_{\text{cd}}], \; \nabla_\theta \mathbb{E}[\mathcal{L}_{\text{sr}}] \rangle < 0, \tag{64}$$

then for sufficiently small learning rate, a step that decreases $\mathbb{E}[\mathcal{L}_{\text{sr}}]$ can increase $\mathbb{E}[\mathcal{L}_{\text{cd}}]$ (or the converse), yielding slower convergence or worse stationary points than optimizing each stage separately.

Practical implication. Stage-wise training fixes g_θ to satisfy a well-posed fidelity criterion (Eq. equation 61, left), and then trains f_ϕ on a stable input distribution (Eq. equation 61, right). This avoids (i) task-hacking in g_θ under weak supervision, and (ii) destructive gradient interference between $\mathcal{L}_{\text{sr}}$ and $\mathcal{L}_{\text{cd}}$, which can otherwise produce visually implausible SR outputs or brittle CD performance under domain shift.

I Cross-Resolution LEVIR-CD vs Regular-Resolution Baselines

Despite being evaluated under a significantly more challenging cross-resolution setting (4×), our method surpasses prior approaches that operate under the standard same-resolution scenario. This highlights the robustness and generalization ability of our model, showing that it not only handles resolution gaps effectively but also achieves higher overall accuracy than methods specifically designed for regular-resolution change detection.

Method	Setting	OA↑
Madani et al. (2026)	Regular (1×)	98.70
Madani & Patel (2026)	Regular (1×)	98.00
Ours	Cross-resolution (4×)	**98.89**

Table 8: Overall accuracy comparison on LEVIR-CD. DiffRegCD and Morphing-T evaluate under standard same-resolution conditions (1×), while our method operates in a harder cross-resolution (4×) setting.

J Qualitative Results

We further provide qualitative visualizations to illustrate the behavior of the proposed framework under both change and no-change scenarios. Representative examples are provided in the supplementary materials, where each sample contains the before image, low-resolution after image, reconstructed after image produced by our diffusion-based super-resolution module, the high-resolution after reference, the ground-truth mask, and the final prediction. In the change cases, the reconstructed after images recover fine structural patterns such as roof edges, road intersections, and building footprints, which in turn enable more accurate localization of change regions. The predicted masks closely follow object boundaries and preserve semantic consistency with the ground-truth annotations, demonstrating that aligning the LR after image to the HR domain mitigates the resolution mismatch that typically degrades detection quality. In the no-change cases, the model produces stable and noise-free predictions, avoiding spurious detections even in textured natural scenes and residential regions with subtle illumination or appearance variations. These examples show that the framework can suppress many reconstruction-induced false positives in unchanged regions, although the diffusion model does not explicitly model reconstruction uncertainty. Overall, the qualitative observations are consistent with the quantitative results and highlight the ability of our method to (i) enhance structural fidelity in reconstructed LR images, (ii) improve boundary alignment in detected change regions, and (iii) reduce false-positive responses in no-change areas.

J.1 Detection of Small-Scale Changes

Figure 6 highlights our method's ability to adapt to small and complex change categories such as *swimming pools*. These regions are challenging because they occupy only a few pixels relative to the full image resolution, often leading to misdetections or omissions in existing benchmarks. Our approach accurately segments these small changes, confirming that the adaptive filtering and hierarchical validation stages allow the model to generalize effectively even in fine-grained, low-area scenarios.

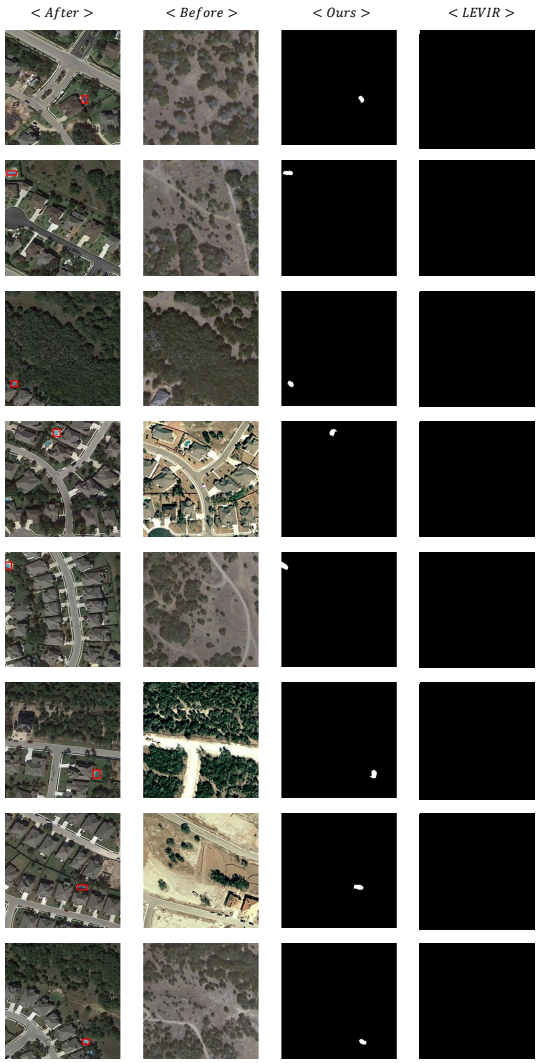

Figure 6: Visual comparison on challenging small-object cases. Our method demonstrates superior adaptability in detecting fine-grained changes such as *swimming pools*, which often occupy very small regions within high-resolution images. Compared to *LEVIR*, our approach successfully identifies subtle localized changes while avoiding false positives in surrounding areas.

To assess the robustness of our method in identifying fine-grained semantic changes, we evaluate its ability to detect subtle variations in vegetation, building structures, and newly constructed roads.

J.2 Qualitative Comparison: Missed Changes

Figure 7 illustrates qualitative examples highlighting cases where our method successfully detects meaningful semantic changes that are labeled as "no change" in the public Chen & Shi (2020)

dataset. While others provide strong building-level annotations, it occasionally omits fine-grained structural or surface-level changes (e.g., new roads, terrain modifications, or partial constructions). Our segmentation pipeline demonstrates greater sensitivity to these overlooked regions, producing accurate masks that capture finer structural or environmental variations. Although we do not report a direct metric comparison against the ground truth, we qualitatively observe that our method yields more semantically consistent results, especially in cases where LEVIR annotations fail to reflect real-world changes.

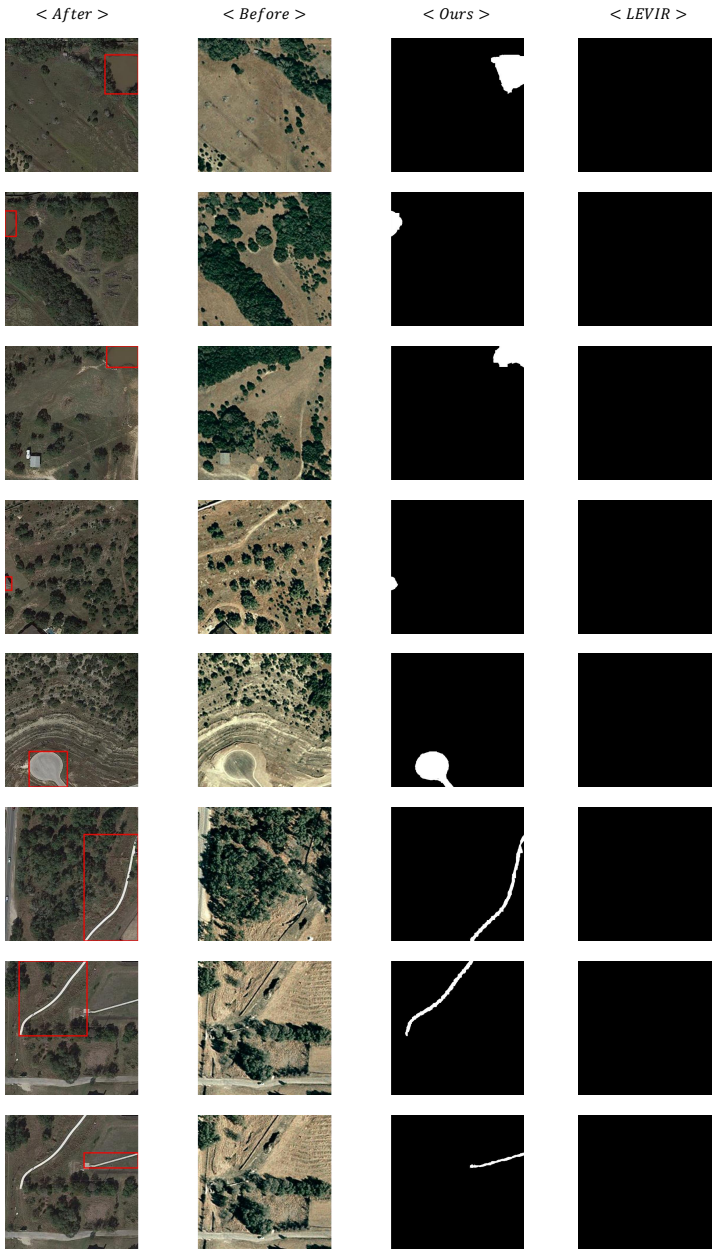

Figure 7: Qualitative comparison between our method and the LEVIR ground truth. Our model detects subtle structural or environmental changes that are annotated as “no change” in the public dataset.

As shown in Figure 8, vegetation and tree-level changes that our method successfully detects but are missing in the *LEVIR* annotations, while example 3 demonstrates a minor building modification accurately captured by our pipeline. These examples underscore the advantage of our segmentation-

and vision-language-guided framework, which remains sensitive to localized and small-scale variations often misrepresented in existing ground truth datasets.

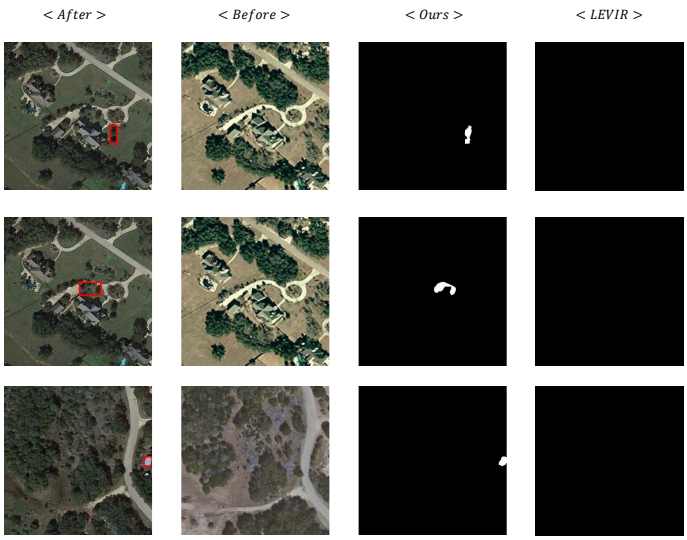

Figure 8: Comparison of small-scale change detection in vegetation and buildings. Examples 1 and 2 show successful detection of vegetation changes, while example 3 captures a subtle building modification. Our method detects these fine-grained changes that are commonly overlooked in *LEVIR* due to incomplete ground truth annotations.

K IoU Threshold Calibration

To determine an appropriate intersection-over-union (IoU) threshold for identifying true object-level changes, we conducted a controlled calibration experiment using human evaluation over a randomly selected subset of 100 image pairs. Each pair was manually annotated to indicate whether the observed differences corresponded to genuine structural change (e.g., new construction, building removal) or false change (e.g., illumination shifts, vegetation growth, or shadow variations). We then compared these binary human judgments against a continuous range of IoU values computed between masks at times t and $t + \Delta t$. Figure 9 illustrates the receiver operating characteristic (ROC) analysis used to optimize the IoU threshold τ_{iou}. The ROC curve evaluates the trade-off between true positive rate (TPR) and false positive rate (FPR) when varying τ_{iou} over the interval $[0, 1]$. The analysis achieved an area under the curve (AUC) of 0.79, confirming that IoU is a strong discriminator of true versus false changes. The optimal threshold $\tau_{\text{iou}} = 0.18$ corresponds to the maximal Youden's J index, representing the most balanced separation between changed and unchanged regions as perceived by human evaluators.

To further illustrate the decision boundary, Figure 10 shows the ROC curve corresponding to the optimal threshold $\tau_{\text{iou}} = 0.18$. The near-perfect separation (AUC $= 0.79$) demonstrates that human judgments align well with this empirical threshold, providing an interpretable and reproducible boundary for localizing true changes while suppressing noisy detections.

L Human-Agreement Evaluation of Reward-Filtered Pseudo-Labels

L.1 On Convergence under Noisy Pseudo-Labels

We formalize the intuition that, after RL-style filtering, training on many pseudo-labeled examples allows the model to learn the correct pixel-level features even if individual masks are sometimes imperfect.

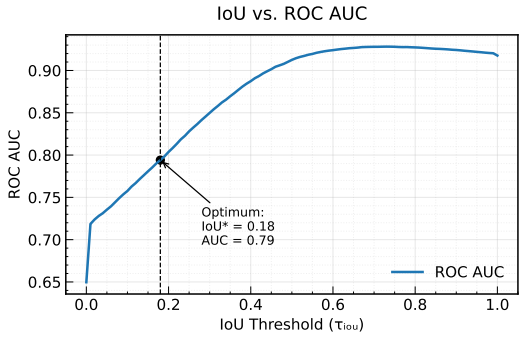

Figure 9: **IoU vs. ROC AUC.** ROC-AUC analysis over the IoU range $[0, 1]$ using human-annotated subtest images. The peak AUC is achieved near $\tau_{\text{iou}} = 0.18$, indicating this value yields the best trade-off between sensitivity and specificity in detecting true structural changes.

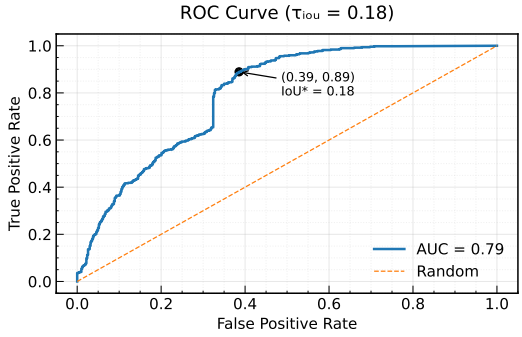

Figure 10: **ROC Curve for IoU Threshold Optimization.** The optimal intersection-over-union threshold $\tau_{\text{iou}} = 0.18$ achieves the best trade-off between true and false detections, as determined from 100 human-annotated image pairs. High TPR and low FPR values confirm that this threshold reliably distinguishes meaningful structural changes from spurious differences.

Convergence under High-Confidence Pseudo-Labels Let (X, Y) be random variables where $X \in \mathcal{X}$ denotes an input pixel (or patch) and $Y \in \{0, 1\}$ the true change label. Let $\tilde{Y} \in \{0, 1\}$ denote the pseudo-label produced by the segmentation and discovery pipeline, and let $D \in \{0, 1\}$ indicate whether a sample is retained after RL-style filtering ($D = 1$ means the mask is kept).

Assume the following:

$$\mathbb{P}(\tilde{Y} = Y \mid X, D = 1) \geq 1 - \varepsilon \quad \text{for all } X \in \mathcal{X}, \tag{65}$$

for some $\varepsilon < \frac{1}{2}$, and that retained samples $\{(X_i, \tilde{Y}_i) : D_i = 1\}_{i=1}^{n}$ are drawn i.i.d. from the conditional distribution $(X, \tilde{Y} \mid D = 1)$. Let $\mathcal{H}$ be a hypothesis class with finite capacity (e.g., finite VC dimension), and let $h_n \in \mathcal{H}$ be the empirical risk minimizer on the retained pseudo-labels with respect to a classification-calibrated loss ℓ:

$$h_n \in \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell\big(h(X_i), \tilde{Y}_i\big). \tag{66}$$

Then, under standard statistical learning assumptions, the sequence $\{h_n\}$ converges in probability to a classifier $h^\star$ whose decision boundary coincides almost everywhere with the Bayes-optimal classifier trained on the true labels Y. In particular, as $n \to \infty$, the probability that $h_n(X)$ predicts the correct class for a random pixel converges to the optimal value achievable by $\mathcal{H}$, even though individual pseudo-labels may occasionally be incorrect.

Informal interpretation. The RL-style filtering stage ensures that, among the retained pseudo-labels, the probability of being correct is strictly biased toward the true class (i.e., correct more often than

not). When the model is trained on a large number of such high-confidence but noisy examples, the aggregate supervision remains statistically aligned with the true decision boundary. As a result, the network learns the dominant visual features of the new change classes at the pixel level, rather than overfitting to occasional mis-labeled masks. Thus, the goal of filtering is not to obtain perfect pseudo-labels, but to guarantee that their noise is controlled and biased toward the correct class, which is sufficient for convergence in the limit.

To assess the quality of the pseudo-labels selected by our RL-style filtering pipeline, we construct a small held-out evaluation subset. For each candidate region mask m_i produced by the segmentation and discovery pipeline, a human reviewer provides a binary judgment indicating whether the region corresponds to a meaningful and visually coherent change. Since no manual pixel-level annotations are drawn, the evaluation is performed at the mask level through agreement rather than geometric overlap.

Binary human-agreement labels. Each evaluated mask is assigned a binary label

$$a_i \in \{0, 1\}, \tag{67}$$

where $a_i = 1$ denotes that the reviewer agrees that m_i represents a valid semantic change, and $a_i = 0$ otherwise. The RL-style filtering stage produces a binary decision variable

$$d_i \in \{0, 1\}, \tag{68}$$

where $d_i = 1$ if the mask is retained as a high-quality pseudo-label and $d_i = 0$ if it is discarded.

Reward-filtered human-agreement precision. We measure the fraction of retained pseudo-labels that humans judge as valid:

$$P_{\mathrm{HA}} = \frac{\sum_i d_i \, a_i}{\sum_i d_i}. \tag{69}$$

A higher P_{HA} indicates that the RL-style reward and filtering mechanism successfully favors region proposals that humans perceive as meaningful semantic changes, rather than noisy or trivial segments.

This evaluation subset is used *only for assessment* and is not used for training, filtering, or model adaptation. The pipeline itself remains fully unsupervised during discovery and fine-tuning; human judgment is employed solely to quantify the reliability of the pseudo-labels selected by our filtering process. The results in Table 9 show that 97% of the retained pseudo-labels were judged as valid changes by human reviewers, empirically confirming that the filtering stage produces a high-confidence supervision set suitable for reliable fine-tuning under our convergence assumptions.

Metric	Value (%)
Human-agreement precision P_{HA}	97.0

Table 9: Human-agreement evaluation of reward-filtered pseudo-labels. P_{HA} denotes the fraction of retained masks that human reviewers judged as valid semantic changes.

M Satellite Image Acquisition and Preprocessing

We constructed a custom dataset to exclude regions dominated by vegetation or trees in both temporal images (before and after change). To achieve this, we applied a patch filtering procedure based on statistical and colorimetric thresholds, as detailed in Algorithm 2. This filtering ensures that only urban or built-up regions with minimal vegetation are retained for further analysis.

For comparison and benchmarking, we additionally employed the LEVIR-CD dataset Chen & Shi (2020), a widely used benchmark for building change detection in very high-resolution (VHR) remote sensing imagery. LEVIR-CD consists of paired satellite images captured at different times over the same geographic regions, with pixel-level annotations of building changes. This dataset enables the study of semantic changes over time under realistic urban conditions.

Importantly, we present all LEVIR-CD samples *as is*, without applying any of our filtering criteria, to maintain consistency with prior work and preserve the original dataset distribution.

1. **Uniformity filter:** Patches with low pixel intensity variation are discarded. If
$$\sigma(X) < \tau_{\text{std}}, \quad \tau_{\text{std}} = 0.08, \tag{70}$$
the patch is excluded.
2. **Low saturation filter:** Patches with low mean saturation in HSV space are removed. If
$$\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} S_{ij} < \tau_{\text{sat}}, \quad \tau_{\text{sat}} = 0.15, \tag{71}$$
the patch is discarded.
3. **Vegetation filter:** Using the Excess Green Index,
$$\text{ExG} = 2G - R - B, \tag{72}$$
patches dominated by vegetation are removed if
$$\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} \text{ExG}_{ij} > \tau_{\text{ExG}}, \quad \tau_{\text{ExG}} = 0.35. \tag{73}$$

Algorithm 2 Patch Filtering Process

Input: Image patch $X \in \mathbb{R}^{H\times W\times 3}$, thresholds $\tau_{\text{std}}, \tau_{\text{sat}}, \tau_{\text{ExG}}$
Output: Keep or discard decision
Compute pixel intensity standard deviation $\sigma(X)$
if $\sigma(X) < \tau_{\text{std}}$ **then**
    Discard patch
end if
Convert X to HSV, compute mean saturation $\bar{S}$
if $\bar{S} < \tau_{\text{sat}}$ **then**
    Discard patch
end if
Compute Excess Green Index: $\text{ExG} = 2G - R - B$
Compute mean $\overline{\text{ExG}}$
if $\overline{\text{ExG}} > \tau_{\text{ExG}}$ **then**
    Discard patch
end if
Keep patch

N TOP-k SIMILARITY EVALUATION

To assess the reliability of the retrieval stage, we conducted a human evaluation experiment using a subset of 100 test images. For each query patch x, the SigLIP model retrieved the top-k candidates $\{x_1, x_2, \ldots, x_k\}$ ranked by cosine similarity:

$$s_i = \frac{f_\theta(x) \cdot f_\theta(x_i)}{\|f_\theta(x)\|\|f_\theta(x_i)\|}, \quad s_i \in [0, 1]. \tag{74}$$

A human annotator was then asked to decide whether the retrieved patch x_i corresponds to the same semantic class as the query image. For each value of $k \in \{1, 2, \ldots, 10\}$, we computed the average percentage of retrieved samples judged as correct:

$$A(k) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\big[\text{HumanAgree}(x_n, x_{n,i\le k})\big], \tag{75}$$

where $\mathbf{1}[\cdot]$ is an indicator function and $N = 100$ is the number of evaluated images.

As shown in Figure 11, retrieval accuracy increases sharply from Top-1 to Top-5, then reaches a stable plateau. This indicates that for all 100 test samples, the correct semantic class was consistently present within the top five retrieved neighbors, with no improvement for larger k. Consequently, we fixed $k = 5$ as the retrieval depth for all subsequent experiments, balancing recall, semantic diversity, and computational efficiency.

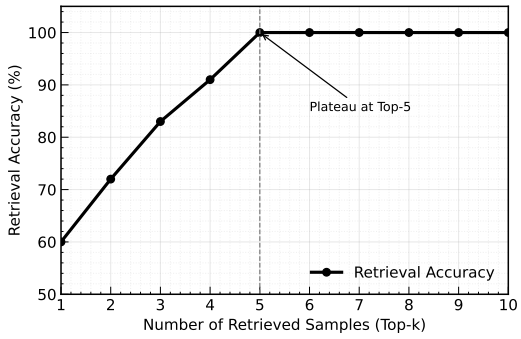

Figure 11: **Top-k Retrieval Consistency.** Human-annotated retrieval accuracy for SigLIP embeddings over 100 test images. Accuracy increases rapidly up to $k = 5$, after which a plateau at 100% agreement is observed, confirming that Top-5 retrieval captures all semantically correct matches.

N.1 Candidate Screening and LLM Reward Statistics

The image–text encoder first removes semantically implausible candidate regions. Candidates that pass this stage are evaluated by the frozen LLM judge as part of the PPO reward model. Importantly, the LLM output does not directly determine pseudo-label acceptance; final acceptance is produced by the policy's terminal `accept`/`reject` decision.

Table 10 summarizes these stages. Of $15{,}253$ segmentation candidates, $8{,}325$ (54.6%) are removed by image–text screening, leaving $6{,}928$ candidates. The LLM judge assigns a positive semantic judgment to $1{,}613$ (23.3%) and a negative judgment to $5{,}315$ (76.7%) of these candidates.

Table 10: Candidate-screening and LLM-judgment statistics. Image–text rates are computed over all candidates, while LLM rates are computed over the $6{,}928$ candidates surviving image–text screening. LLM judgments contribute to the PPO reward and do not directly determine pseudo-label acceptance.

Stage	Count	Rate	Role
Total candidates	$15{,}253$	—	Segmentation output
Rejected by image–text encoder	$8{,}325$	54.6%	Discarded
Forwarded to LLM judge	$6{,}928$	45.4%	Reward evaluation
Positive LLM judgment	$1{,}613$	23.3%	Semantic reward signal
Negative LLM judgment	$5{,}315$	76.7%	Semantic reward signal

N.2 Human Agreement Across Dataset Creation Pipelines

As summarized in the results section and illustrated in Figure 12, our full pipeline (**Segmentation + Encoder + RLtR**) achieves the highest human agreement of 98%, substantially outperforming all other baselines. Encoder-only and zero-shot models show limited semantic consistency, while few-shot adaptation-based on cosine-similarity retrieval-improves alignment across all LLMs. These results highlight the critical role of both visual filtering and reinforcement-based refinement in achieving human-level interpretability and reliability.

N.3 Evaluation of Distance Metrics for Human-Annotated Similarity

To assess the effectiveness of different distance metrics for similarity-based retrieval, we conducted a human evaluation on a subset of $N = 100$ test images. For each image x_i, annotators were asked to confirm whether the true class appeared within the Top-5 retrieved results under each metric. The final accuracy for each metric m was computed as:

$$A(m) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\text{HumanAgree}(x_i, m)], \tag{76}$$

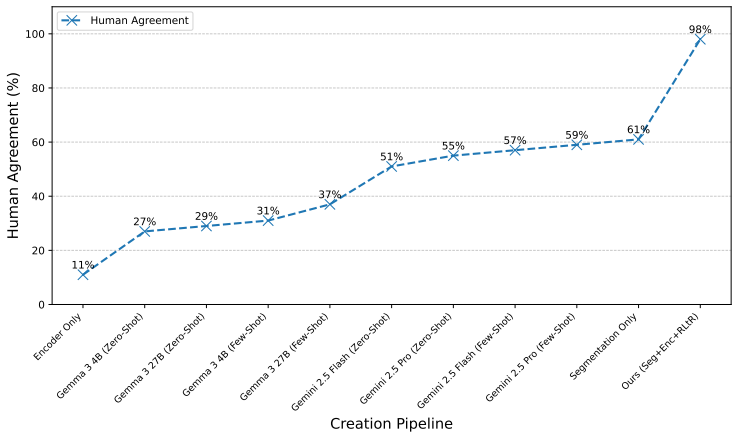

Figure 12: Human agreement across dataset creation pipelines. Models are sorted in ascending order by agreement rate. Few-shot variants consistently outperform their zero-shot counterparts, while our full method reaches near-perfect human alignment.

where $\mathbf{1}[\cdot]$ is an indicator function equal to 1 when the annotator confirmed the retrieval as correct.

Distance Metrics. We compared four commonly used distance metrics for feature-space similarity evaluation:

- **Cosine Similarity:**
$$S_{\cos}(E_1, E_2) = \frac{E_1 \cdot E_2}{\|E_1\| \, \|E_2\|}, \tag{77}$$
which measures angular similarity between two embedding vectors.
- **L1 Distance:**
$$d_{L1}(E_1, E_2) = \sum_{j=1}^{D} |E_{1j} - E_{2j}|, \tag{78}$$
providing a robust measure of absolute deviation across dimensions.
- **L2 Distance:**
$$d_{L2}(E_1, E_2) = \sqrt{\sum_{j=1}^{D} (E_{1j} - E_{2j})^2}, \tag{79}$$
which emphasizes larger differences by squaring deviations.
- **Wasserstein Distance:**
$$W(E_1, E_2) = \inf_{\gamma \in \Gamma(E_1, E_2)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|], \tag{80}$$
representing the minimal transport cost required to align the distributions of two embedding sets.

Results. Figure 13 illustrates the agreement rate between human annotators and each metric. Cosine similarity achieved perfect alignment (100% agreement), while L_2, L_1, and Wasserstein distances achieved 94%, 93%, and 84% respectively. The results suggest that cosine similarity best captures the semantic relationships in the embedding space, while distance-based metrics are less effective in representing perceptual alignment with human judgment.

N.4 Embedding Model Comparison for Top-5 Retrieval

To evaluate the retrieval efficacy of different embedding models on our Top-k similarity (Cosine Similarity) task, we compared three representative vision-language encoders: **CLIP** Radford et al. (2021), **MaMMUT** Kuo et al. (2023), and **SigLIP 2** Tschannen et al. (2025).

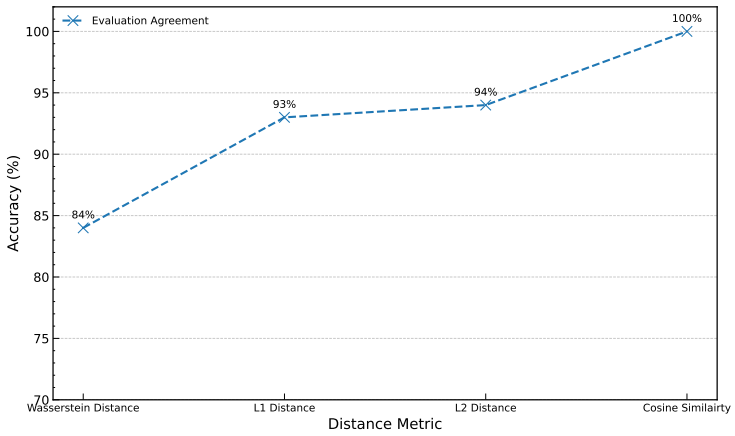

Figure 13: Human-annotated retrieval agreement across different distance metrics. Accuracy values are computed for $N = 100$ test images with Top-5 similarity retrieval. Cosine similarity achieved the highest correspondence with human evaluation.

CLIP jointly optimizes an image encoder f_θ and a text encoder g_ϕ using a contrastive loss:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(f_\theta(I_i), g_\phi(T_i))/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(f_\theta(I_i), g_\phi(T_j))/\tau)}, \tag{81}$$

where sim denotes cosine similarity and τ is a temperature parameter. SigLIP replaces the softmax loss with a sigmoid-based regression objective:

$$\mathcal{L}_{\text{SigLIP}} = -[y \cdot \log\sigma(s) + (1-y)\cdot\log(1-\sigma(s))], \tag{82}$$

which improves training efficiency and scalability. Finally, MaMMUT employs a multi-task loss combining masked language modeling and image-text matching:

$$\mathcal{L}_{\text{MaMMUT}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{MIM}} + \mathcal{L}_{\text{ITM}}, \tag{83}$$

enhancing fine-grained alignment for subtle semantic reasoning.

For each query patch, a human annotator determined whether the correct semantic class appeared within the top-5 retrieved candidates generated by each model. Figure 14 presents the resulting annotation agreement rates. As shown, CLIP achieved 61%, MaMMUT reached 89%, and SigLIP 2 achieved 100%. These results motivated our use of SigLIP 2 as the retrieval backbone for downstream ranking.

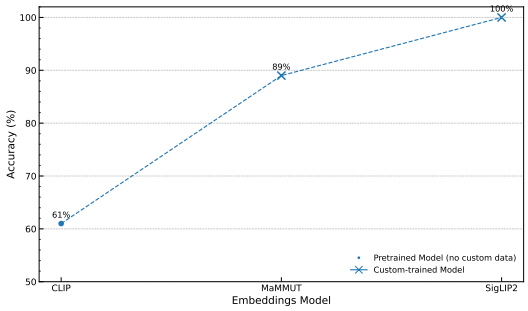

Figure 14: **Top-5 Retrieval Accuracy across Embedding Models.** Human annotator agreement rates for Top-5 candidate sets generated by each model. Marker types differentiate: "`.`" denotes a purely pretrained model, and "`x`" denotes a model fine-tuned on our domain.

O SEMANTIC VERIFICATION OF FALSE-POSITIVE REGIONS

To address noisy annotations in existing benchmarks, we introduce a filtering stage that detects and removes false-positive change predictions in unchanged regions. Figure 15 highlights representative "no change" cases often misclassified as semantic changes in prior benchmarks.

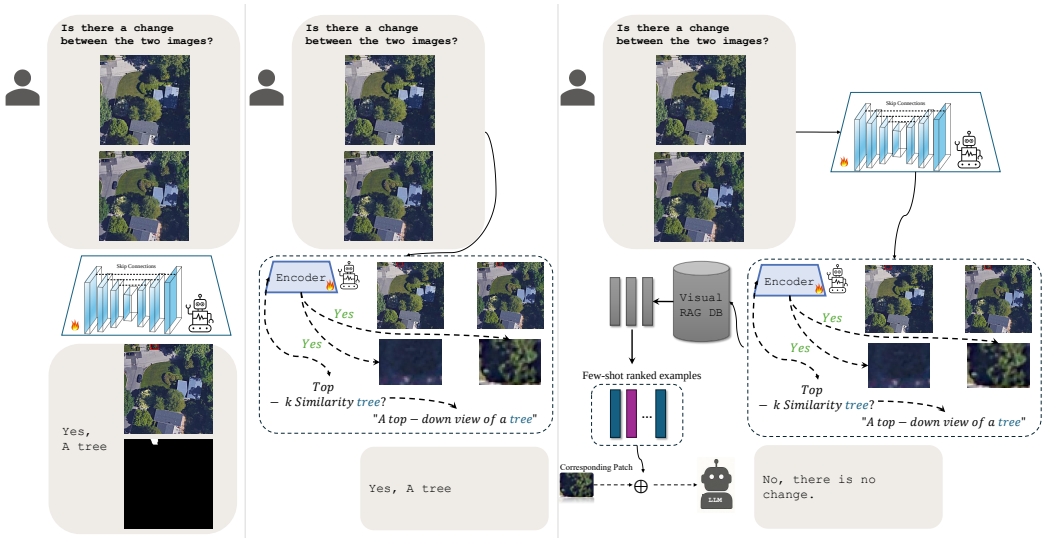

Figure 15: **Filtering false positives of "no change."** State-of-the-art segmentation models often misclassify unchanged regions as changes. Given image patches x_{before} and x_{after}, we use an encoder $f(\cdot)$ to compute similarity. If the similarity $s \geq \tau_{\text{sim}}$, the region is classified as *no change*. To ensure robustness, we retrieve class-based examples $\mathcal{E} = \{(x_k, y_k)\}_{k=1}^{K}$ and validate with a LLM J_ϕ. Only if J_ϕ confirms the same class across time, the patch is discarded as false change. This cleans the dataset and prevents noisy ground truth labels.

P UNIFIED SEMANTIC TAXONOMY AND EVALUATION OF NOVEL CLASS DISCOVERY

Standard pixel-level metrics do not fully capture the goal of novel class discovery, which is to determine whether the adaptation stage accepts semantically valid unseen change instances rather than artifacts. In addition, existing cross-resolution change-detection benchmarks do not provide a shared semantic ontology for evaluating unseen-class discovery across datasets. To address this, we introduce a unified semantic taxonomy and evaluate novel class discovery at the instance level.

Unified semantic taxonomy. We define a common semantic label space

$$\begin{aligned}\mathcal{C}_{\text{all}} = \{&\texttt{driveway}, \texttt{ground_natural},\\ &\texttt{parking_lot}, \texttt{road_unpaved}, \texttt{sidewalk},\\ &\texttt{trail}, \texttt{tree}, \texttt{other_human_made},\\ &\texttt{track}, \texttt{bike_lane}, \texttt{crosswalk},\\ &\texttt{painted_median}, \texttt{railway_tracks},\\ &\texttt{shipping_container}, \texttt{stairs},\\ &\texttt{swimming_pool}, \texttt{athletic_field}\}.\end{aligned} \tag{84}$$

This taxonomy serves as a benchmark-level ontology for candidate generation, retrieval, reward evaluation, and unseen-class analysis. It is introduced for consistent semantic evaluation and does not replace the original annotation space of the source datasets.

Dataset mapping. After spatial alignment and candidate extraction, each candidate region is assigned to one of the classes in $\mathcal{C}_{\text{all}}$ using the semantic relabeling procedure described in the methodology. This converts binary or dataset-specific change annotations into a shared semantic representation and enables unified evaluation across datasets. Table 11 summarizes both the original annotation space of each dataset and the fine-grained mapped classes used in our unified benchmark.

Held-out-class benchmark setup. To evaluate adaptation to semantic classes excluded from task-specific detector training, we partition the unified label space into disjoint seen and held-out subsets,

$$\mathcal{C}_{\text{seen}} \cap \mathcal{C}_{\text{unseen}} = \emptyset, \qquad \mathcal{C}_{\text{seen}} \cup \mathcal{C}_{\text{unseen}} = \mathcal{C}_{\text{all}}. \tag{85}$$

The base detector, candidate scoring modules, and adaptation policy are trained using only classes in $\mathcal{C}_{\text{seen}}$. Evaluation is then performed on unlabeled image pairs containing changes from $\mathcal{C}_{\text{unseen}}$. This protocol measures whether the learned policy can recover semantically valid changes from classes held out from task-specific training without target-scene mask supervision for those classes.

For each dataset d, let $\mathcal{C}^{(d)} \subseteq \mathcal{C}_{\text{all}}$ denote the subset of unified classes present after relabeling. We define dataset-specific seen and unseen partitions

$$\mathcal{C}^{(d)}_{\text{seen}} \subset \mathcal{C}^{(d)}, \qquad \mathcal{C}^{(d)}_{\text{unseen}} = \mathcal{C}^{(d)} \setminus \mathcal{C}^{(d)}_{\text{seen}}. \tag{86}$$

Only instances in $\mathcal{C}^{(d)}_{\text{seen}}$ are used during supervised training and policy optimization, while unseen-class evaluation is restricted to instances whose semantic label belongs to $\mathcal{C}^{(d)}_{\text{unseen}}$.

Evaluation protocol. We evaluate novel class discovery at the instance level rather than with pixel-level precision and recall. Specifically, we assess whether an accepted terminal mask corresponds to a valid unseen semantic change or to a false acceptance such as a segmentation artifact, localization error, boundary error, or semantically incorrect region. This directly tests the quality of the discovered novel-class instances rather than dense mask overlap.

Human verification. To validate semantic correctness, three independent annotators jointly reviewed the same set of 1000 accepted terminal masks. Each example was labeled according to whether the accepted mask corresponded to a valid unseen semantic change. Final labels were determined by annotator agreement. Among the 1000 reviewed examples, 959 were judged to be valid discoveries, yielding a human-verified accuracy of 95.9%. This result indicates that the proposed adaptation stage is able to recover semantically meaningful unseen change instances with high reliability.

Qualitative analysis. We further provide representative examples of correct novel-class discoveries, rejected artifacts, and failure cases involving visually plausible but semantically invalid regions. These examples illustrate both the strengths of the policy and the main sources of residual error.

Leave-one-class-out evaluation. For each dataset d with unified class set $\mathcal{C}^{(d)} = \{c_1, \ldots, c_{K_d}\}$, we perform K_d independent evaluation folds. In fold j, class c_j is treated as unseen:

$$\mathcal{C}^{(d,j)}_{\text{unseen}} = \{c_j\}, \qquad \mathcal{C}^{(d,j)}_{\text{seen}} = \mathcal{C}^{(d)} \setminus \{c_j\}.$$

The base detector, candidate-scoring modules, policy, and retrieval support set are constructed using only $\mathcal{C}^{(d,j)}_{\text{seen}}$. The held-out class is used only for unseen-class evaluation and unlabeled adaptation. The reported results are macro-averaged across all leave-one-class-out folds:

$$\overline{\mathcal{M}}^{(d)} = \frac{1}{K_d} \sum_{j=1}^{K_d} \mathcal{M}^{(d,j)},$$

where $\mathcal{M}^{(d,j)}$ denotes the evaluation metric when c_j is held out.

Q Class Distribution Analysis

Table 13 and Figure 16 present the distribution of detected changes across semantic categories in our generated dataset (test only). The results indicate a clear dominance of the *building* and *tree* classes, which together account for over 75% of all detected changes. This imbalance aligns with the urban-suburban nature of the underlying LEVIR-CD imagery, where most structural and vegetation modifications occur.

Table 11: Unified benchmark construction across datasets. The second column lists the original annotation space of each source dataset. The third column lists the fine-grained semantic subclasses assigned by our relabeling pipeline for unified novel-class evaluation. These mapped classes are not native annotation labels of the source datasets.

Dataset	**Original labels**	**Mapped classes from $\mathcal{C}_{\text{all}}$**
LEVIR-CD	Binary change/no-change; changed instances restricted to buildings.	`driveway`, `parking_lot`, `sidewalk`, `tree`, `other_human_made`, `crosswalk`, `painted_median`, `swimming_pool`, and `stairs`.
SV-CD	Binary change/no-change under strong seasonal variation.	`ground_natural`, `road_unpaved`, `trail`, `tree`, `track`, `railway_tracks`, `parking_lot`, `other_human_made`, `athletic_field`, and `driveway`.
DE-CD	`impervious`, `agriculture`, `forest_vegetation`, `wetlands`, `soil`, `water`, and `snow_ice`.	`ground_natural`, `parking_lot`, `road_unpaved`, `sidewalk`, `trail`, `tree`, `other_human_made`, `track`, `bike_lane`, `crosswalk`, `painted_median`, `railway_tracks`, `shipping_container`, `stairs`, `swimming_pool`, `athletic_field`, and `driveway`.

Table 12: Human-verified accuracy of novel class discovery on accepted terminal masks.

Evaluation Set	Reviewed Samples	Accuracy ↑
Accepted terminal masks	1000	95.9%

While smaller categories such as *roads*, *sidewalks*, and *parking lots* appear less frequently, they provide valuable diversity for model generalization. The long-tail behavior visible in Figure 16 emphasizes the need for adaptive sampling and weighting during training to ensure balanced learning across object types. Overall, this distribution validates the coverage of both man-made and natural change patterns in our dataset.

Table 13: Number of Detected Changes per Class in the Test Set. The proportion represents the ratio of each class relative to the total number of detected changes.

Changed Object Category	**Number of Changes**	**Proportion (%)**
Building	6209	54.9
Tree	2570	22.7
Sidewalk	779	6.9
Road paved	524	4.6
Swimming pool	213	1.9
Parking lot	188	1.7
Water	157	1.4
Trail	116	1.0
Shipping container	79	0.7
Road unpaved	58	0.5
Athletic field	50	0.4
Driveway	39	0.3
Railway tracks	25	0.2
Painted median	24	0.2
Crosswalk	20	0.2
Bike lane	16	0.1
Track	5	0.04
Stairs	4	0.04
Total	**11376**	**100.0**

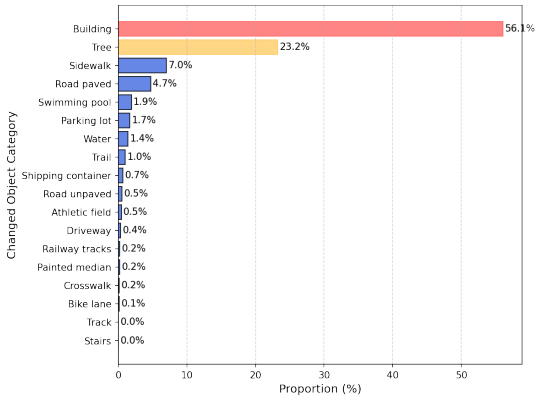

Figure 16: Class distribution of detected changes in the test set. The majority of changes correspond to *building* and *tree* categories, followed by smaller contributions from man-made and natural structures. This distribution highlights the long-tail nature of the dataset.

Table 14: Summary of Processed Image Pairs

Description	**Count**
Rows with at least one change	1362
Rows with no change	567
Total individual changes	11076
Total rows processed	1929

Table 14 provides a quantitative summary of our dataset after the automated filtering and validation stages. Out of the total 1,929 processed image pairs, 1,362 contain at least one verified change, while 567 are classified as *no-change* samples. In total, the dataset includes 11,076 individual annotated changes, demonstrating both large-scale coverage and balanced representation across change and non-change cases. These statistics confirm the robustness of our automated pipeline in producing reliable, diverse examples suitable for change detection and captioning tasks.

R Prompt Engineering

Prompt-engineering ablation. The following experiment evaluates alternative prompting strategies for the Gemma semantic judge. The 1–5 scoring formulation is used only for this prompt-ablation study and is not the reward representation used by the PPO policy. In the final method, the judge outputs `VALID` or `INVALID`, which is mapped to $f_{\text{text}} \in \{0, 1\}$.

R.1 System Prompt for Gemma Filtering

1. **Instruction:** The model is prompted as an expert in satellite image interpretation to score a query patch based on the presence of a specific target class $\{selected_class\}$. It receives reference examples and must assign a score from 1 to 5 according to visibility and clarity.

   ```
   ``You are an expert in recognizing objects from
   satellite images.  Your task is to score a query
   image patch from 1 to 5.  You need to specify if
   {selected_class} appears in the image.  All images are
   satellite images.  Return only the numerical score
   (1, 2, 3, 4, or 5).''
   ```

2. **Scoring Guide:** The scoring criteria used for filtering are defined as follows:

   ```
   ``5:  There is definitely a {selected_class} in the
   last image.  The object's shape, shadow, and features
   are clearly visible from above.
   ```

```
4: Very likely that the image contains a
{selected_class}. Features are mostly clear.
3: Probably the image contains a {selected_class},
but visibility or details are ambiguous.
2: Unlikely that a {selected_class} appears in the
image.
1: Definitely does not contain a {selected_class}.''
```

3. **Example Format:** The model is shown several examples of reference and query images structured as:

```
``Example (1): {start_of_image} Score = 5
Example (2): {start_of_image} Score = 3
...
Example (5): {query_image} Score = ?''
```

This systematic prompt design enforces consistent, interpretable visual filtering behavior, ensuring the model evaluates satellite imagery with quantitative reliability across all semantic classes.

All experiments in this section were conducted to analyze the effect of different prompt formulations on the Gemma-based satellite image filtering task. The goal is to evaluate how textual guidance impacts the model's ability to accurately score the presence of a target semantic class within a given query image.

For each image sample I, the model receives a textual instruction prompt T and returns a classification score prediction:

$$R = \mathcal{V}(T, I), \tag{87}$$

where $\mathcal{V}$ denotes the Gemma inference function parameterized by the given prompt.

To systematically evaluate the effect of context and structure, we experiment with progressively more informative prompt variants as described below.

1. **Instruction-only:** A minimal instruction without additional guidance.

```
``You are an expert in recognizing objects from
satellite images. Your task is to score whether a
given class appears in the last image. Return only
the numerical score (1-5).''
```

2. **Score Guide:** Adds explicit definitions of scoring criteria to improve rating consistency.

```
``Score the image from 1 to 5, where 5 means the
object is clearly visible, 3 indicates uncertainty,
and 1 means it is definitely absent.''
```

3. **Image Examples:** Extends the prompt with several labeled examples of reference and query image pairs.

```
``Example (1): {image_1} Score = 5. Example (2):
{image_2} Score = 3. Example (3): {query_image} Score
= ?''
```

4. **Combined (Final):** Combines both the scoring guide and reference examples for maximum clarity and contextual learning.

```
``You are an expert in satellite image
interpretation. Rate whether the object class
appears in the last image, using a score from 1 to 5.
Follow the scoring guide: 5 = Definitely visible; 4
= Very likely visible; 3 = Unclear; 2 = Unlikely; 1 =
Definitely not visible. Example (1): {image_1} Score
= 5. Example (2): {image_2} Score = 3. Example (3):
{query_image} Score = ?''
```

To evaluate the effectiveness of each Gemma prompt variant, we employed a human evaluator to assess the model's predictions on a set of 100 satellite image pairs. All pre-processing stages-including semantic segmentation, connected component extraction, and image-text encoding-were held constant to isolate the effect of the prompt itself.

The results show that the simplest instruction-only prompt achieved a moderate agreement of 61%. Introducing explicit scoring guidance improved interpretability and consistency, reaching 74%. When example-based context was added, the accuracy further increased to 87%, while the combined prompt, which integrates both scoring definitions and visual exemplars, achieved the highest human-aligned performance of **98%**. These findings highlight that structured contextual prompts substantially enhance Gemma's reasoning ability and scoring precision across visual categories.

S Robustness to the Choice of Segmentation Backbone

To evaluate the sensitivity of the proposed filtering mechanism to the choice of segmentation backbone, we compare our custom segmentation model to two established baselines: SegFormer (Xie et al., 2021), a transformer-based semantic segmentation model, and Mask2Former (Cheng et al., 2021), which jointly models pixel and mask representations. Both baselines were fine-tuned on the OpenEarthMap (Xia et al., 2023). For each model, we generate segmentation masks on a representative subset of the dataset and compute the pairwise IoU between the resulting masks. Table 15 reports the average IoU scores between the three segmentation models. These results suggest that the proposed Best-of-N filtering mechanism is largely insensitive to the specific segmentation backbone employed.

Table 15: Pairwise IoU between segmentation models.

Model	**Ours**	**SegFormer**	**Mask2Former**
Ours	1.00	0.94	0.96
SegFormer	0.94	1.00	0.98
Mask2Former	0.96	0.98	1.00

T Implementation Details and Leakage Control

To make the adaptation stage fully reproducible, we specify the main implementation choices used in the candidate-generation, retrieval, reward-modeling, and policy-optimization stages. Candidate masks are generated using three segmentation backbones: a ViT-based encoder-decoder model with a lightweight segmentation head, SegFormer, and Mask2Former. For each before/after pair, all segmentation models are applied independently, and their semantic difference maps are converted into connected components. Components that do not satisfy the minimum area, overlap, and image-text consistency thresholds are discarded before reinforcement learning begins.

For image-text scoring, we use SigLIP as the remote-sensing image-text encoder. For language-based semantic feedback, we use Gemma-3-4B as a frozen judge. The judge receives the cropped before/after regions, the candidate mask description, and retrieved support examples, but it is not given the held-out labels. This prevents the reward model from directly exploiting the target class names during evaluation. Retrieval is restricted to the training split only. No images, masks, captions, or candidate regions from the validation or test regions are included in the retrieval pool.

For each candidate, we retrieve the top-5 most similar support examples according to the SigLIP embedding similarity. The retrieval set contains both accepted and rejected examples from the training data, allowing the judge to compare a candidate against positive and negative reference cases. Thresholds for candidate filtering are selected on the validation split and then kept fixed for all test experiments. We use the same thresholds across datasets unless stated otherwise, avoiding per-test-set tuning.

The RL policy is optimized with PPO for a small fixed budget of 20,000 environment steps, using a maximum episode horizon of $T = 5$ refinement actions before a terminal accept/reject decision is forced. At each step, the agent may expand, contract, shift, keep, accept, or reject the current mask.

To keep computation bounded and avoid over-sampling easy regions, we retain at most 20 candidate masks per image after pre-filtering, ranked by image-text consistency and geometric validity. These candidates are then refined independently by the policy.

We parameterize the PPO objective with a clipping threshold $\epsilon = 0.2$, a value function coefficient $\lambda_v = 0.5$, and an entropy bonus $\lambda_H = 0.01$ to encourage exploration. Advantages are computed using Generalized Advantage Estimation (GAE) with a discount factor $\gamma = 0.99$ and a trace decay parameter $\lambda = 0.95$. The network is trained using the Adam optimizer with a learning rate of 3×10^{-4}, a batch size of 128, and 4 update epochs per rollout. The reward shaping terms are weighted by $\alpha = 1.0$, $\beta = 0.1$, and $\gamma_{\text{disc}} = 0.99$. For the discrete spatial actions, the expand, contract, and shift operations modify the predicted mask bounding box by a fixed uniform step size of 10%.

We also enforce strict leakage control. Retrieval examples are drawn only from the training partition, threshold selection is performed only on validation data, and final evaluation is performed on spatially disjoint test regions. The LLM judge is frozen, receives no test labels, and is not provided with the names of the held-out classes. Therefore, accepted masks are selected only from visual evidence, retrieved training examples, geometric consistency, and image-text alignment, rather than from direct access to test annotations or held-out-class labels.

U Qualitative Results

We further provide qualitative visualizations to illustrate the behavior of the proposed framework under both change and no-change scenarios. Representative examples are shown in Figure 20, where each sample contains the before image, low-resolution after image, reconstructed after image produced by our diffusion-based super-resolution module, the high-resolution after reference, the ground-truth mask, and the final prediction. In the change cases, the reconstructed after images recover fine structural patterns such as roof edges, road intersections, and building footprints, which in turn enable more accurate localization of change regions. The predicted masks closely follow object boundaries and preserve semantic consistency with the ground-truth annotations, demonstrating that aligning the LR after image to the HR domain mitigates the resolution mismatch that typically degrades detection quality. In the no-change cases, the model produces stable and noise-free predictions, avoiding spurious detections even in textured natural scenes and residential regions with subtle illumination or appearance variations. These examples show that the framework can suppress many reconstruction-induced false positives in unchanged regions, although the diffusion model does not explicitly model reconstruction uncertainty. Overall, the qualitative observations are consistent with the quantitative results and highlight the ability of our method to (i) enhance structural fidelity in reconstructed LR images, (ii) improve boundary alignment in detected change regions, and (iii) reduce false-positive responses in no-change areas.

U.1 Unlabeled Discovery Examples

Figures 17-19 provide additional qualitative results from our cross-resolution change discovery framework. Each example shows the $\textit{HighRes}_{\textit{Before}}$ image, the $\textit{LowRes}_{\textit{After}}$ observation, the reconstructed $\textit{Recon}_{\textit{After}}$ output produced by our super-resolution module, and the corresponding change-prediction maps generated by C_{base} (labeled data only) and $(C_{\text{base}} + \Delta C)$ (ours with RL discovery). The qualitative results illustrate resolution alignment, improved semantic consistency, and unsupervised discovery of previously unseen change classes.

Furthermore, Figure 22 presents cases of newly built or extended roads that our approach identifies correctly, while previous methods classify them as *no-change* due to incomplete ground truth coverage. This emphasizes our model's capability to generalize to fine-grained, low-contrast structural updates, even in challenging rural or semi-urban regions.

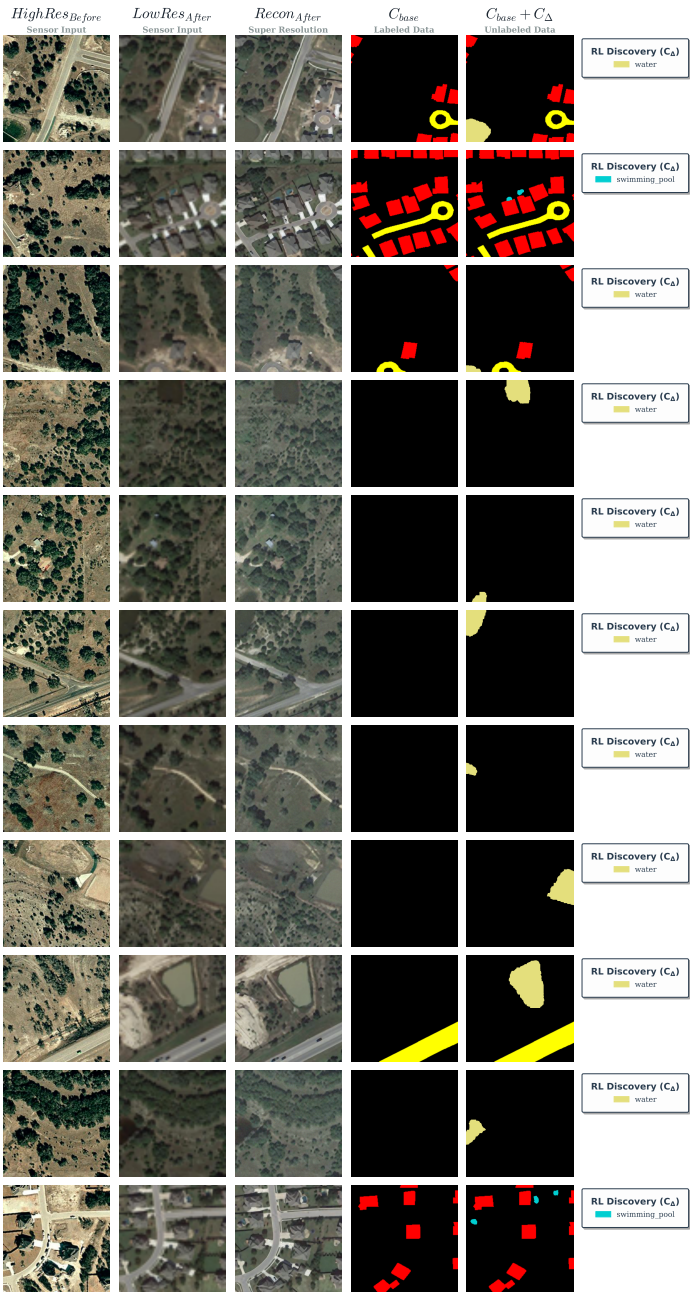

Figure 17: Cross-resolution change discovery qualitative examples 1.

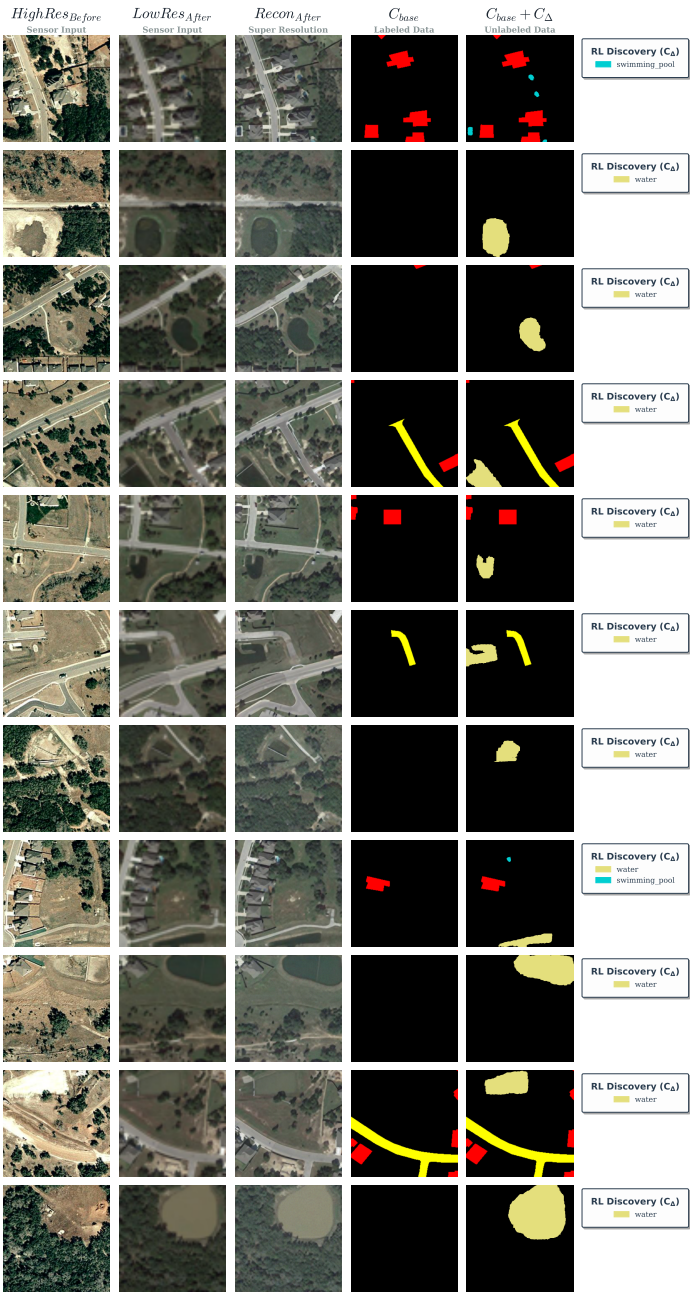

Figure 18: Cross-resolution change discovery qualitative examples 2.



Figure 19: Cross-resolution change discovery qualitative examples 3.

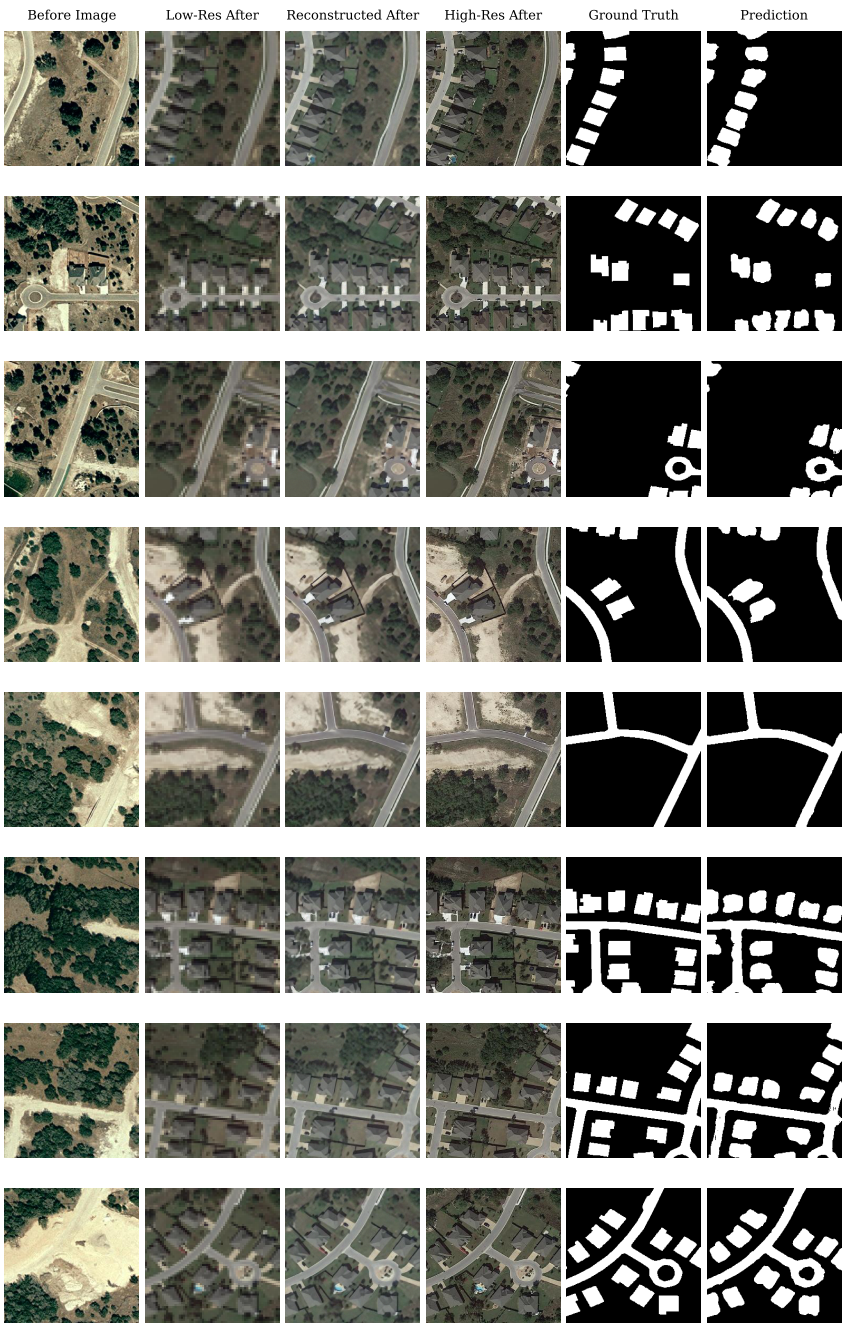

Figure 20: Qualitative examples of change cases. The model successfully reconstructs the low-resolution after image and produces accurate change maps that closely align with the ground-truth annotations across diverse urban layouts.



Figure 21: Qualitative examples of no-change cases. The model preserves structural consistency between bitemporal observations and avoids generating false alarms, demonstrating strong robustness in regions with minimal semantic variation.

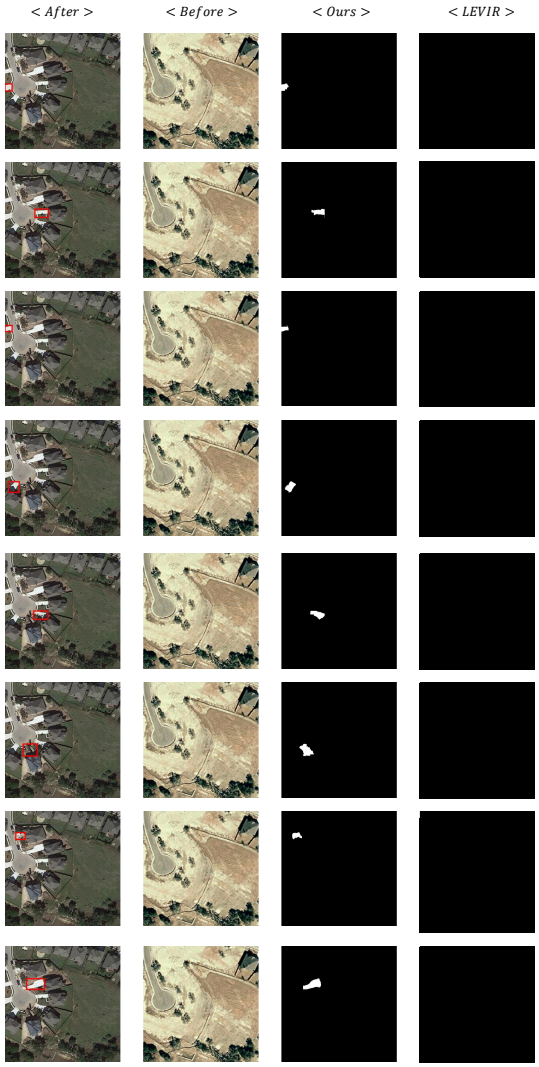

Figure 22: Detection of newly constructed or extended roads. Our approach correctly identifies subtle road-level changes that prior methods missed, primarily due to incomplete or misaligned ground truth annotations in *LEVIR*.